technische universität
dortmund

fakultät
statistik

Survival models with selection of genomic covariates in heterogeneous cancer studies

Dissertation

in Fulfillment of
the Requirements for the Degree of
Doktor der Naturwissenschaften

Submitted to
the Faculty of Statistics
of the TU Dortmund University

by
Katrin Madjar

July 16, 2018

Referees:
Prof. Dr. Jörg Rahnenführer
Prof. Dr. Katja Ickstadt

Date of Oral Examination:
September 7, 2018

Danksagung

Mein Dank gilt in erster Linie meinem Doktorvater Prof. Dr. Jörg Rahnenführer für die Ermöglichung und hervorragende Betreuung meiner Arbeit, sowie seine ständige Diskussions- und Hilfsbereitschaft. Er hat mir den Freiraum ermöglicht, mich mit den unterschiedlichen wissenschaftlichen Forschungsschwerpunkten auseinanderzusetzen und mich dabei stets ermutigt und motiviert. Ich danke meiner Zweitgutachterin Prof. Dr. Katja Ickstadt und Prof. Dr. Manuela Zucknick für die konstruktiven Gespräche und wertvollen Ideen, die einen wichtigen Beitrag zu meiner Arbeit geleistet haben. Bedanken möchte ich mich auch bei meinen Kollegen für die freundschaftliche Arbeitsatmosphäre, den bereichernden und konstruktiven Austausch und stete Hilfsbereitschaft. Bei meiner Familie und Tillmann möchte ich mich ganz besonders herzlich bedanken, dass sie immer für mich da waren und mich stets ermutigt und unterstützt haben.

Contents

Chapter 1

Introduction

Survival analysis is an important objective in various fields of biomedical research, particularly in cancer research. Main goals are the prediction of a patient's risk and the identification of new prognostic biomarkers to improve patients' prognosis. In recent years, molecular data such as microarray gene expression, next-generation sequencing, copy number variation (CNV) or single-nucleotide polymorphism (SNP) data have increasingly gained importance in diagnosis and prediction of disease outcome. It is well-known that genes do not act in isolation but are related to other genes in complex molecular networks and interact in regulatory or functional pathways. Exploring the underlying biological mechanisms and detecting relevant genes and pathways are an important task in biomedical research, as they help to better comprehend the cause of disease and enable the targeted development of new pharmaceuticals.

The use of high-throughput technologies allows simultaneous measurements of genome-wide data for patients and results in high-dimensional data, where the number of genomic predictors greatly exceeds the number of patients ($p >> n$). In this situation, the number of genes associated with outcome is typically small. Important objectives in modeling high-dimensional data are good prediction performance and finding a subset of predictors that are truly relevant to the outcome. A sparse model solution may reduce noise in estimation and increase interpretability of the results. Besides predictive accuracy and sparsity, another criterion for model evaluation in high-dimensional settings is stability in variable selection. This implies a similar set of selected covariates across different resampling data sets (or similar patient cohorts) (Lausser et al., 2013; Bommert, Rahnenführer, and Lang, 2017). Several studies have shown that published gene lists obtained for the same clinical type of patients differ greatly and show only small overlap. They are highly unstable and depend strongly on the subset of patients used for gene selection (Ein-Dor et al., 2005; Michiels, Koscielny, and Hill, 2005; Ein-Dor, Zuk, and Domany, 2006). This lack of reproducibility raises doubts about the reliability and robustness of the reported biomarkers. Main sources of instability in feature selection are the small number of patients used to generate the gene lists (limited amount of information in the data), correlations between genes, and genes with weak effects on outcome (Sauerbrei, Boulesteix, and Binder, 2011; He and Yu, 2010). Variable selection stability will also be considered in the course of this thesis, however, the main focus is on predictive accuracy.

One problem with high-dimensional data is that standard approaches for parameter estimation in regression models cannot handle such a large number of predictors. In survival analysis, the solution maximizing the partial likelihood of the Cox proportional hazards model (Cox, 1972) is not unique for $p >> n$. But even in cases where $p < n$, conventional regression techniques may result in an overfitted model that performs well

on fitted data but fails in external validation. Besides poor prediction performance, results may be hard to interpret if the number of included predictors is large. To tackle this problem, several adaptions have been proposed during the last years, such as prior dimension reduction through univariate selection (Bøvelstad et al., 2007; Witten and Tibshirani, 2010b) or supervised and unsupervised principal components (Bair and Tibshirani, 2004; Bair et al., 2006; Hastie, Tibshirani, and Friedman, 2009, chapter 3.5.1), traditional variable selection like forward stepwise selection (Bøvelstad et al., 2007; Witten and Tibshirani, 2010b), regularized regression introducing a penalty term into the (partial) likelihood such as ridge, lasso or the elastic net as combination of both (Verweij and Van Houwelingen, 1994; Tibshirani, 1997; Zou and Hastie, 2005), and boosting algorithms (Hothorn and Bühlmann, 2006; Tutz and Binder, 2006). Univariate feature selection has the disadvantage that prediction performance of the resulting multivariate model is not necessarily higher than in the univariate models, particularly when selected predictors are highly correlated with each other. In this case, the multivariate model does not provide substantial additional information (Witten and Tibshirani, 2010b). Bøvelstad et al. (2007) compare some of these methods with regard to their prediction performance and find that the ridge regression (Verweij and Van Houwelingen, 1994) performs best. However, a drawback of the ridge penalty is that it does not result in sparse models, because it only shrinks regression coefficients towards zero. In comparison, the lasso penalty (Tibshirani, 1997) estimates some or most regression coefficients exactly zero, thus implying automatic variable selection. Sparse models are also provided by boosting algorithms.

In addition to genomic predictors, established clinical covariates such as age, sex or tumor stage are often available in practice. In this case, it is of interest to determine the additional predictive value of genomic predictors over clinical covariates. Due to the large number of genomic predictors compared to the typically small number of clinical covariates, the latter might be dominated in a combined model including both types of predictors. This problem requires the mandatory inclusion of clinical covariates, while only genomic covariates are subject to variable selection or penalization. Comparative studies show that combined models including clinical and genomic predictors often outperform those models based only on one type of predictors with respect to better predictive abilities (Boulesteix and Sauerbrei, 2011; De Bin, Sauerbrei, and Boulesteix, 2014; Binder and Schumacher, 2008; Bøvelstad, Nygård, and Borgan, 2009).

Many diseases exhibit considerable heterogeneity with regard to biological characteristics and clinical outcomes. A major subject of research has been the study of cancer genomes and identification of molecular subtypes of cancer based on differences in genomic measurements. Different subtypes may have different prognoses, responses to therapy and progression patterns and thus, can be a challenge for cancer diagnosis and treatment. Clinical oncology has focused on personalized therapies with the aim of identifying patient subgroups that benefit from specific targeted treatment, and preventing unnecessary harm to patients, who are unlikely to respond. Therefore, exploring and understanding molecular mechanisms and tumor heterogeneity provides deeper biological insight into tumor development and improves targeted therapies (Curtis et al., 2012; Almendro, Marusyk, and Polyak, 2013; Bedard et al., 2013; Junttila and Sauvage, 2013). Despite distinct differences, different types of cancer have a shared genetic basis (Hanahan and Weinberg, 2011). Several cross-cancer analyses have been conducted to detect common pathways, risk loci and genetic variants, such as variants of DNA repair

genes (Scarbrough et al., 2016; Fehringer et al., 2016; Kar et al., 2016).

Within the scope of this thesis, we expect the underlying data to be heterogeneous due to known patient subgroups with different prognoses. We assume that some subgroups are closer related to one another with regard to sharing genomic predictors with similar effects on survival outcome, while other subgroups may be un- or less related due to subgroup-specific or even opposite effects (the latter are very rare). In a combined model that pools patients from all subgroups the results may be biased and subgroup-specific effects may get lost because the effects are averaged. Standard subgroup analysis on the other hand, relies only on the patients belonging to the subgroup of interest and disregards information from the other subgroups. Due to reduced sample size, this may lead to a loss of power or unstable results with high variance, especially in small subgroups such as patients with rare diseases. Therefore a tradeoff between both standard approaches is needed that combines their advantages. What we are looking for is the prediction of each subgroup and the detection of subgroup effects, but to allow sharing of information between subgroups, when supported by data.

In this thesis, two approaches are proposed to address the problem of predicting survival outcome based on potentially high-dimensional covariates such as gene expression data in a heterogeneous cohort with prespecified subgroups. Both proposed methods perform variable selection in the Cox proportional hazards model and provide a separate model fit for each subgroup. They consider heterogeneity of data and allow sharing information between subgroups, when appropriate. Besides identifying relevant predictors, they also help to uncover similarities between subgroups. We consider multiple cancer studies as subgroups, however, our approaches can be applied to any other subgroups, for example, defined by clinical covariates.

Our first approach is a frequentist weighted Cox regression model with lasso penalty to induce variable selection. Individual weights are estimated for each subgroup such that they represent the probability of belonging to that subgroup. Patients who are similar to the subgroup of interest receive higher weights in the subgroup-specific model. Sample size is not reduced as in standard subgroup analysis since all patients are used for estimation and weights account for the heterogeneity in the data. This approach not only allows estimation of a separate model for each subgroup, but also provides information on the similarities between subgroups. The idea for the derivation of these weights originates from Bickel et al. (2008) who apply the weights in logistic regression for modeling the success of HIV-drug therapies. Netzer (2013) uses these weights to predict the survival outcome of subgroups defined by clinical covariates based on CNV data. We adopt this weighted approach but with improved estimation of the weights, by comparing different methods and incorporating cross-validation in order to prevent overfitting. Furthermore, we conduct an extensive simulation study to evaluate the potential of our method which has not been done before.

Our second approach is a novel hierarchical Bayesian Cox model with Bayesian variable selection that assumes a graph structure among the predictors within and between subgroups. This network is used in variable selection and favors the selection of predictors that are related in the graph. Prior knowledge of the network is not required. Instead, simultaneous inference of the relevant predictors for each subgroup and the network among them is performed. This approach provides an insight into the relationships among predictors within and between subgroups and allows the

identification of related predictors that jointly impact the response. The novelty of our model is the combination of existing Bayesian approaches and the extension to subgroup analysis.

1.1 Structure of this thesis

This thesis begins with an overview of existing approaches for subgroup analysis. Chapter 2 provides information on the biological background, including a description of techniques for the measurement and preprocessing of gene expression data as well as key facts on lung cancer. Furthermore, an application example of real lung cancer studies is described.

Chapter 3 introduces statistical methods for the analysis of high-dimensional data with time-to-event endpoint and extensions to subgroup analysis. First, all methods are explained in the classical, frequentist setting. Central terms and standard methods for survival analysis are described including the Cox proportional hazards model. Regularization techniques for dealing with high-dimensional predictors are proposed as well as two measures of prediction performance for model evaluation. The standard Cox model is extended to a weighted version in order to take subgroups into account. An approach to derive individual weights for subgroup analysis is presented along with classification methods for their estimation. Then, classical frequentist approaches are followed by an outline of Bayesian approaches for subgroup analysis in high-dimensional data with survival outcome. This comprises the Bayesian formulation of the Cox model, Bayesian variable selection and structure learning in graphical models. Existing approaches in these fields of research are combined and extended to the proposed new Bayesian subgroup model.

In chapter 4 both, the proposed frequentist and the Bayesian subgroup model, are applied to simulated and real lung cancer studies and compared with standard subgroup and global analysis of all patients. The Bayesian model is applied to genomic predictors only, but can be extended to include mandatory clinical covariates as in the case of the frequentist model. The latter additionally serves to investigate the question, whether prediction accuracy can be improved by combining clinical and genomic covariates. A model including both types of predictors is compared to a model based on only clinical or only genomic predictors. This thesis concludes with a summary of the main results and an outlook for further research in chapter 5.

1.2 Overview of subgroup analysis

Subgroup analysis is an important objective in various medical applications. In epidemiology, a major concern is the investigation of heterogeneity in the relationship between exposure and disease subtypes. Assuming that subtypes are known a priori, Wang et al. (2016) provide an overview of statistical methods for risk assessment of disease subtype heterogeneity that are appropriate for different study types, such as cohort, case-control and case-case studies. This comprises hypothesis tests to assess the association of exposure with a specific subtype or to test if the exposure effects are the same across subtypes.

An important goal in the development of personalized therapies is the identification and confirmation of subgroups of patients who will benefit from a specific treatment and have a positive benefit-risk balance. Traditional subgroup analysis in clinical trials includes statistical tests for interaction to assess differences in treatment effects between patient subgroups (Pocock et al., 2002; Rothwell, 2005). Ondra et al. (2016) perform a systematic literature search on novel statistical methods to analyze the heterogeneity of treatment effects across patient subgroups defined by biomarkers in clinical trials. This includes confirmatory approaches, where treatment effects are investigated to demonstrate the treatment's efficacy in one subgroup (or a small number of predefined subgroups) and/ or the entire population by controlling the familywise error rate, and exploratory approaches that may compare multiple subgroups without error rate control or identify unknown subgroups. Adaptive designs are also thoroughly reported. In these trials, patients are recruited in several stages and after each stage, an interim analysis is conducted. The results of an interim analysis allow a continuation of the trial as initially planned, an earlier stop either for futility or success, or a modification of the trial's design to, for example, restrict patient recruitment in subsequent stages to a predefined subgroup.

Matsui et al. (2017) aim at identifying various association profiles of genes across treatment subgroups. First, for each gene, a univariate regression model is fitted separately for the treatment and control group. The distribution of the effect estimates from both subgroups is modeled by a two-component hierarchical mixture with Gaussian distributions at the gene level and a non-parametric prior for mean effect sizes across genes. The two mixture distributions represent the null genes with no effect in both subgroups and the non-null genes of interest. Then gene ranking and selection is performed using an optimal discovery procedure based on the estimated model with false discovery rate control.

In this thesis, subgroups are considered to be different studies. Thus, meta-analysis methods appear to be a natural choice to summarize information across studies. Classical meta-analysis analyzes each study separately and then pools summary statistics of the individual study results. Summary statistics can be obtained by combining the p-values from the studies (e.g. Fisher's and Stouffer's methods) or by modeling the effect sizes from the studies (fixed-effect and random-effects model). Traditionally, the fixed-effect (FE) model has been used when all studies are assumed to have a common effect and differences between studies are due to sampling variation, while the random-effects (RE) model has been preferred under the assumption of heterogeneity. However, in the presence of strong between-study heterogeneity, the RE is less powerful than the FE approach. The test based on Cochran's Q also has low power in detecting true heterogeneity among studies when the number of studies is small (Higgins et al., 2003; Ioannidis, Patsopoulos, and Evangelou, 2007; Thompson, Attia, and Minelli, 2011).

With advances in high-throughput technologies in recent years, the amount of genomic data for genome-wide association studies (GWAS) has increased. A review on meta-analysis methods for GWAS can be found in Evangelou and Ioannidis (2013) and Thompson, Attia, and Minelli (2011), and with respect to microarray gene expression data in Tseng, Ghosh, and Feingold (2012). In GWAS associations of genes with diseases often differ across studies, which might be due to noise or chance (homogeneity) or due to genuine underlying differences such as genetic diversity (heterogeneity). The latter

is of great interest because it allows to further explore and understand heterogeneous associations and possibly identify novel genetic variants for complex diseases (Ioannidis, Patsopoulos, and Evangelou, 2007; Pei et al., 2016).

Thus, the detection of between-study heterogeneity is an important goal in meta-analysis and various new approaches have been proposed to deal with this problem. Han and Eskin (2011) and Neupane et al. (2012) propose new RE methods that assume no heterogeneity under the null hypothesis and have higher power to detect heterogeneous associations across studies. Han and Eskin (2012) introduce a RE-type model that is the adaptively weighted sum of z-scores method. Adapted weights include a new statistic termed m-value which corresponds to the posterior probability that the effect exists in each study. In this model a greater weight is assigned to studies predicted to have an effect and a smaller weight to studies predicted to have no effect. Bayesian extensions of the FE model with varying levels of heterogeneity are suggested by Wen and Stephens (2014).

Böhning, Dietz, and Schlattmann (1998) and Raim, Neerchal, and Morel (2014) consider model-based clustering in meta-analysis by fitting a finite mixture of densities to the studies to identify unknown clusters of similar studies. To detect outlier studies, Beath (2014) assumes a mixture of outlier and non-outlier studies with RE models that differ in their variances. After the identification of outlier studies, the overall treatment effect is estimated by including all studies but with outliers down-weighted. A subset-based approach is proposed by Bhattacharjee et al. (2012) that tests all combinations of studies (subsets) for the presence of true effects in either the same or opposite directions. A FE-type test statistic is computed for each subset and the best subset is selected as the one with the maximum test statistic (strongest overall effect).

Classical methods for combining p-values of studies aim at testing the alternative hypothesis that at least one of the studies has a non-zero effect size against the null that the effect sizes are zero in all studies. Adaptions to genomic data have been developed to test for at least a prespecified number of non-zero effects. In this framework, Li et al. (2014) and Song and Tseng (2014) consider ordered p-value approaches. Song and Tseng (2014) utilize a single r-th ordered p-value (p-value with rank r in ordered list of p-values) as test statistic for testing the alternative that an effect exists in at least one given percentage of studies. In contrast, Li et al. (2014) extend the traditional Fisher's and Stouffer's method to a weighted sum of ordered p-values, where the weights are based on the order of the p-values. They test for non-zero effect sizes in a majority of studies. P-values closer to the median are weighted highest since they better represent the majority of studies, while the smallest/largest p-values are down-weighted. Another adaption of Fisher's method is the adaptively weighted statistic by Li and Tseng (2011), where the weights are used to maximize the significance of the summary statistic. A comparison of this method with the r-th ordered p-value method and with other classical meta-analysis approaches is performed by Chang et al. (2013).

When data are available at patient-level, meta-analysis might lead to a loss of power and less accurate results in studies with small sample sizes. Instead of analyzing multiple studies separately and combining their results as in meta-analysis, integrative analysis analyzes the raw data from all studies simultaneously. Thus, integrative analysis can be more informative and outperform meta-analysis (Ma, Huang, and Moran, 2009). Liu et al. (2014) and Liu, Huang, and Ma (2014) conduct integrative analysis of multiple

cancer subtypes and multiple heterogeneous cancer studies based on high-dimensional genomic predictors. In this framework, regression coefficients have two dimensions, the gene and the study dimension. To accommodate both, composite penalties are used for estimation and two-level gene selection. For a specific gene, the first level of selection (outer penalty) is to determine whether it is associated with at least one study. For the second level of selection different inner penalties are considered under the homogeneity and heterogeneity model. The homogeneity model assumes the same set of prognostic covariates in all studies. Thus, within each study, the ridge penalty is appropriate as inner penalty since it encourages shrinkage but no selection. Under the heterogeneity model, a covariate can be associated with the response in some studies but not others. In this case, a lasso type penalty is appropriate as inner penalty to identify which study a selected gene is associated with.

In contrast to integrative analysis for the aggregation of multiple studies with the same type of (omics) data, Boulesteix et al. (2017) perform integrative analysis of multiple omics data types available for the same patient cohort. They use a lasso penalty with different penalty parameters for the different data types that can be determined either by cross-validation or prespecified by the user. In simulation studies, the authors show that their approach (called IPF-LASSO) performs better in terms of prediction performance and sparsity than the standard and sparse group lasso when the data types are different with respect to relevant variables. Bergersen, Glad, and Lyng (2011) integrate external information provided by another genomic data type into variable selection in the Cox model. They propose a weighted lasso that penalizes each covariate individually with weights inversely proportional to the external information. Information is based on the Spearman correlation between genes of both data types or obtained from the ridge regression coefficients of the Cox model fitted to the other data type. An additional tuning parameter controlling the relative strength of all weights is optimized by cross-validation. Stingo et al. (2011) incorporate structural information on gene pathways as prior into variable selection in a Bayesian framework. The joint distribution of binary variable selection indicators is modeled by a Markov Random Field that includes prior information on the relationship between genes in a pathway. A review of integrative Bayesian analysis of different types of molecular data for the same set of samples is performed by Ickstadt, Schäfer, and Zucknick (2018). They discuss multiple approaches for integrative analysis with a focus on gene prioritization (identification of genes that differ between biological or clinical conditions involving a multiple testing problem), model-based clustering for identification of subgroups, variable selection in regression modeling, and structure learning for graphical models.

Instead of sharing information between subgroups by integrating external information into variable selection, Huang et al. (2011) propose a weighted approach for combining positive predictive value (PPV) and negative predictive value (NPV) across populations when the assumption of common classification accuracy is justified. ROC curve estimation is used to evaluate the ability of a risk prediction marker in discriminating diseased from non-diseased. The estimates of PPV and NPV are based on a weighted average of the ROC curves from a target and an auxiliary population.

Weighted regression approaches, that assign different weights to the observations in the likelihood, have been proposed in order to take into account heterogeneity in a cohort due to known subgroups. Weyer and Binder (2015) use a weighted and stratified

Cox regression model based on componentwise boosting for automatic variable selection. They consider the case of two subgroups (strata) and focus on predicting only one specific stratum. Observations from this stratum receive a weight of 1 in the stratum-specific likelihood, while observations from the other stratum are down-weighted with a fixed weight in the interval (0,1). A stratified likelihood allows estimating a separate baseline hazard for each stratum, which is advantageous for heterogeneous subgroups. However, fixed weights are less flexible than the estimated weights used in our proposed frequentist subgroup model.

In a Bayesian setting, Bogojeska and Lengauer (2012) apply a weighted logistic regression model to predict the binary treatment response of an HIV combination therapy. Each drug belonging to the target therapy is considered as a separate subgroup with a specific weight. Subgroup-specific weights share a common Gaussian prior with a mean drawn from a Gaussian hyperprior to relate all subgroups and model their similarity. Weights are estimated from a hierarchical logistic regression model that models the combined effects of all subgroups based on the training data. The maximum a posteriori estimates of the model parameters are subsequently used for prediction on the test data. Simon (2002) proposes a method to estimate subgroup-specific treatment effects as an average of observed within-subgroup differences and overall differences. He considers an ordinary Cox model including a binary treatment effect, a binary covariate or linear combination of binary covariates as subgroup indicators and the corresponding treatment-by-subgroup interactions. The maximum partial likelihood estimates of the regression coefficients are assumed to be multivariate normally distributed with the true regression coefficients as mean vector. A Gaussian prior is assigned to the unknown regression coefficients so that the resulting posterior distribution is also normal. Simon shows that the components of the posterior mean are linear combinations of the estimated treatment effects in different subgroups. Extracting the respective scalars gives the subgroup-specific weights that can be applied to the partial likelihood for parameter estimation. However, both approaches are not designed for high-dimensional data since they do not perform variable selection. In addition, Simon (2002) makes the rather restrictive assumption that all covariates are binary.

Local regression is another technique that utilizes weighted regression models but without predefined groups. A separate model is fitted to each observation (query point) based on its neighboring observations. Local neighborhood is specified by a kernel function representing the distance from the query point. Kernel functions are introduced as weights into the likelihood of the local regression model and determine to which extent the single observations influence the estimation. A special case are the K-nearest neighbors of each query point that receive equal weight in the local regression. The parameter defining the width of the neighborhood, such as K, is considered a tuning parameter. All single local regression models together form the local weighted regression based on all observations (Hastie, Tibshirani, and Friedman, 2009, chapters 2.8.2 and 6). A drawback of localized regression is, that it does not provide global regression parameters to describe the relationship between covariates and response, making interpretation difficult. Furthermore, only a small number of observations is used for each local fit, which complicates estimation in high-dimensional settings.

To deal with this problem, Tutz and Binder (2005) develop a penalized localized classification approach with automatic choice of localization, variable selection and penalty

parameters based on cross-validation. Binder et al. (2012) propose a cluster-localized logistic regression with weighted componentwise likelihood-based boosting for automatic variable selection and a special clustering for SNP data. In our weighted regression approach, the predefined subgroups can be considered as query points, each with a separate regression model. However, in contrast to localized regression, our weighted likelihood is based on all observations rather than only on neighboring observations. Our weights correspond to the relation between covariates and subgroup membership instead of the distance in covariate space.

We assume that subgroups are known in advance, however, if the data are expected to have underlying subgroups and subgroup membership is unknown, subgroups can be identified by cluster analysis or tree-based methods. The goal of cluster analysis is to group observations into subsets (clusters) such that those within each cluster are more similar to each other compared to those in different clusters. This requires a measure of dissimilarity or distance between pairs of observations. Some of the most popular clustering algorithms work directly on the observed data without relying on a probability model, such as K-means or hierarchical clustering (Hastie, Tibshirani, and Friedman, 2009, chapter 14.3).

Clustering of gene expression profiles has been used to identify cancer subtypes (Golub et al., 1999; Perou et al., 2000). Mauguen et al. (2017) perform a two-stage approach to define cancer subtypes and identify risk factors with distinctive influence on these subtypes. In the first step, K-means clustering is used and maximization of a heterogeneity measure as optimization criterion to identify cancer subtypes with similar characteristics and mutational profiles. The second step involves correlation of these subtypes with known risk factors in logistic regression to determine the distinctive risk factors.

In high-dimensional settings when the number of features is large compared to the number of observations and when the true underlying clusters are expected to differ only with respect to some features, sparse clustering might provide more accurate and interpretable results. In this context, Witten and Tibshirani (2010a) propose a lasso-type penalty to adaptively select features in K-means and hierarchical clustering. Shen, Olshen, and Ladanyi (2009) develop a latent variable model for integrative clustering of multiple genomic data types. To identify tumor subtypes, different molecular data types measured for the same tumor samples are clustered simultaneously by a K-means procedure and a lasso-type penalty is used to derive a sparse solution.

A Bayesian model-based clustering approach for integrative clustering of multiple genomic data sets is introduced by Kirk et al. (2012). Model-based clustering assumes that the data are an i.i.d sample drawn from a mixture model with K components. Each component density function describes one cluster. The model is usually fit by maximum likelihood using the EM algorithm, or by Markov Chain Monte Carlo (MCMC) methods in Bayesian approaches. Often a mixture of Gaussian densities is used as in Lee, Chen, and Wu (2016) and Chen and Ye (2015). Their approaches involve two problems, separating the cohort into homogeneous components by determining the latent membership of each observation (clustering) and performing variable selection within each component. An overview of model-based clustering can be found in McLachlan and Peel (2000) and Fraley and Raftery (2002). Cluster analysis is generally based on dissimilarities in feature space rather than in the relation to the response as in

tree-based approaches.

Classification and regression trees (CART) partition the predictor space into regions by recursive binary splits. At each node different splitting variables and split points are considered for possible partition, the response is modeled and predicted for each partition and the split resulting in the best fit is selected (for more details see chapter 3.2.2). For continuous or censored response this is a regression problem; for categorical response a classification problem. The response within the two resulting partitions should be preferably homogeneous (pure) and between nodes, the response averages should differ as much as possible. The terminal nodes of a final tree can be considered as subgroups (Breiman et al., 1984; Gordon and Olshen, 1985; LeBlanc and Crowley, 1993). Schmoor, Ulm, and Schumacher (1993) propose a two-step procedure including CART to adjust global comparison of treatment groups for patients' heterogeneity with respect to prognosis. First, CART is used to separate patients based on their covariate values into subgroups with different prognoses. Second, these subgroups are used as strata for the estimation and testing of the treatment effect.

The patient rule induction method (PRIM) also partitions the predictor space but not based on binary decision rules. The aim is to find regions in which the response average is much larger (or smaller) than its average over the entire predictor space (Hastie, Tibshirani, and Friedman, 2009, chapter 9.3; Friedman and Fisher, 1999). Instead of modeling and predicting a response as in CART, a test can be employed to determine if the considered partition results in a significantly higher mean response than expected under the (permutated) null distribution (Dyson and Sing, 2014). PRIM has several medical applications, such as investigating genomic variations as predictors of risk of disease in heterogeneous patient subgroups (Dyson and Sing, 2014). Chen et al. (2015) use PRIM to search for predictive signatures that identify patient subgroups (signature-positive groups) with maximally beneficial treatment effect.

Other tree-based methods have been proposed in the context of randomized clinical trials where the aim is to identify patient subgroups with heterogeneity of treatment effects. Patients are recursively divided into subgroups by focusing on treatment-by-covariate interactions. The splitting criterion measures the heterogeneity of treatment effect between the two resulting partitions. The best split is the one that induces the largest difference in treatment effect. Subgroups are determined by the terminal nodes of the final tree. Su et al. (2011) propose such a tree-structured approach, termed interaction tree, for longitudinal data with continuous response. At each node, a linear regression model is considered including the following: a binary treatment indicator, an indicator associated with a specific split of a specific covariate, and their interaction. The Wald test statistic of the estimated interaction effect is used as splitting criterion. The best split among all permissible splits of the covariate space is selected such that the Wald test statistic is maximized. The overall treatment-by-covariate interaction of the final tree is assessed with a permutation test to confirm the existence of heterogeneous treatment effects. A similar approach is used in Su et al. (2009), but instead of the Wald test statistic the t-test statistic of the treatment-by-covariate interaction effect is used as splitting criterion. Su et al. (2008) and Negassa et al. (2005) develop interaction tree methods for censored survival data with different versions of the partial likelihood ratio statistic (PLRS) as splitting criterion. Negassa et al. (2005) utilize a stratified PLRS with different baseline hazards for the resulting subgroups to compare the model including the treatment and interaction effect with the model

containing only the treatment effect under the assumption of homogeneity. Su et al. (2008) use an unstratified PLRS that compares the model including the treatment and covariate main effects and their interaction with the model based on the two main effects only.

In contrast to interaction trees, where the aim is to divide all patients into subgroups with largest possible difference in treatment effect, the *Virtual Twins* method (Foster, Taylor, and Ruberg, 2011) considers the two treatments as reference and alternative treatment and aims at identifying a subgroup of patients who may have an enhanced treatment effect. For a binary response, Virtual Twins estimates the probability of the response under the two treatments for each patient. The difference between the estimated probabilities represents that patient's individual, differential treatment effect and is used as splitting criterion in the tree algorithm. Some of these methods are compared in Doove et al. (2014).

Alternative approaches for the identification of subgroups with differential treatment effects aim at finding an optimal individualized treatment rule (ITR) which assigns each individual to an appropriate treatment based on observed patient characteristics. An optimal ITR is a map from the predictor space to the treatment space that maximizes the expected clinical outcome (Zhao et al., 2012; Zhang et al., 2012). Addition of penalties allows to deal with a large number of covariates (Xu et al., 2015). A comprehensive review of further statistical methods for subgroup identification in clinical trials is provided by Lipkovich, Dmitrienko, and D'Agostino (2017).

Chapter 2

Data and biological background

Cancer develops by transformation of normal cells to malignant, cancerous cells caused by genetic alterations and structural changes in DNA. These genetic changes may be inherited or occur spontaneously as a result of repeated exposure to carcinogens (such as tobacco smoke, or ultraviolet radiation) and defective DNA repair pathways. Cancer is a major cause of morbidity and mortality worldwide, with about 8 million cancer-related deaths and an estimated 14 million new cases of cancer diagnosed in 2012. It is a very complex disease due to the large number and different types of possible mutations, genetic diversity across world populations, and various environmental influences. Therefore, an important objective in cancer research is the identification of genetic aberrations and changes in regulatory pathways related to tumor development and progression. Better understanding of the underlying molecular mechanisms is improving diagnosis and more targeted treatment. Over the past two decades, technologies for the measurement of gene expression have made rapid progress and are widely used, resulting in a vast amount of genomic data, including publicly available databases, and an increasing number of cancer genome-wide association studies. Findings have revealed changes in tumor-related genes and pathways that allow tumor classification, patient stratification and provide potential therapeutic targets. Despite great advances and rapid growth of information, the identification of genomic biomarkers for clinical use is still challenging. Many of the reported biomarkers have failed in subsequent validation studies and only a few findings have made it to clinical practice (World Health Organization, 2014).

This chapter provides a brief outline of basic epidemiologic and biological information on lung cancer, followed by an explanation of all data collected for the application of the proposed statistical models in section 2.2. In particular, the data collection and curation process is reported, and a descriptive analysis of the applied lung cancer studies is presented. This is followed by an explanation of the Affymetrix microarray technology for gene expression measurements, and a certain preprocessing algorithm for this type of data, in sections 2.3 and 2.4, respectively.

2.1 Lung cancer

Lung cancer is the leading cause of cancer-related death worldwide, making up 20% of the total cancer mortality in 2012. Lung cancer is the most common cancer in men and the third most common in women (after breast and colorectal cancer). Despite recent advances, long-term survival remains poor with a 5-year overall survival rate of 10–15%. This is mainly due to late detection with the majority of patients being diagnosed with locally advanced or metastatic disease (World Health Organization,

2014). Guidelines for treatment decisions and assessing prognosis are still largely based on tumor staging. The TNM staging system describes the extent or size of the primary tumor (T), absence or presence of regional lymph node involvement (N), and absence or presence of distant metastases (M). After determination of the T, N, and M categories, a stage of 0 (in situ), or I to IV (from early to most advanced disease) is assigned. A distinction is made between clinical stage (cTNM) before surgery, and pathological staging (pTNM) post-surgical (for more details see PDQ® Adult Treatment Editorial Board, 2018). For early-stage NSCLC (stages I and II) the standard treatment with the best chance of cure is still surgical resection. Patients with advanced stage or inoperable disease are usually treated with chemotherapy, radiotherapy, molecularly targeted therapy, or recently immunotherapy. However, even when diagnosed in early stages without detectable nodal or metastatic involvement, relapse rate after surgery is high and approximately 50% of patients subsequently develop metastases (Anandappa and Popat, 2016).

The most important risk factor for the development of lung cancer worldwide is tobacco smoking. Risk increases with both quantity and duration of smoking. Smoking produces cellular injury in the whole respiratory tract. Many genetic alterations reverse within months of smoking cessation, but some are irreversible and may explain why former smokers continue to have an increased risk of lung cancer. Other risk factors include increasing age and exposure to second-hand smoke, radon, asbestos, outdoor air pollution (specifically particulate matter), and radiation (World Health Organization, 2014). Historically, small cell lung cancer has been distinguished from non–small cell lung cancer (NSCLC) by morphology. NSCLC accounts for approximately 85% of all lung cancers and the most common histological types are adenocarcinoma (ADC), squamous cell carcinoma (SQC), and large cell carcinoma (LCC). ADC exhibits the highest prevalence in never-smokers and women, whereas SQC predominates in male smokers.

In recent years, lung cancer has increasingly been classified according to molecular differences. Many prognostic gene expression signatures for lung cancer have been developed aiming at patient stratification into subgroups with distinct clinical outcomes. However, most of the early studies have been limited by shortcomings including overfitting, lack of sufficient validation and lack of proper evidence of medical use beyond existing treatment decisions (Subramanian and Simon, 2010). Only a few of the reported prognostic genes for lung cancer have been translated into clinical application. The identification of mutations in certain histological subtypes has led to the development of molecular targeted therapy to improve the survival of patient subsets. Genetic alterations are best established for ADC with well-known driver mutations involving EGFR, KRAS, HER2 (ERBB2), BRAF, and PIK3CA, as well as ALK and ROS1 fusion, and MET amplification. Mutations in PIK3CA, FGFR1, DDR2, PTEN, TP53, SOX2, and CDKN2A have shown to be associated with tumor-genesis in SQC. Therapeutic agents for some of these mutations, such as DDR2, FGFR1, EGFR, ALK and ROS1, are available, others are in development (World Health Organization, 2014).

2.2 Data description

A large number of data sets from different cancer types including survival endpoint, Affymetrix microarray gene expression data, and - to some extent - clinical pathologic

information were downloaded from the Gene Expression Omnibus (GEO) data repository (Edgar, Domrachev, and Lash, 2002) and manually curated. Raw expression data (CEL-files), mainly measured on the Affymetrix HG-U133 Plus 2.0 and HGU-133A array, were normalized using frozen robust multiarray analysis (fRMA) explained in section 2.4. All cohorts were checked for duplicates by looking at correlations of the expression values. Duplicates and normal (non-tumorous) samples, as well as samples with missing survival endpoint, were removed. Overall, this resulted in a collection of ten non-small cell lung cancer (NSCLC) cohorts ($n = 1779$), four colon cancer cohorts ($n = 893$), eight ovarian cancer cohorts ($n = 922$), and 16 breast cancer cohorts ($n = 2193$) with available survival outcome. Breast cancer cohorts can be divided according to treatment (with two cohorts consisting of two different treatment subgroups, thus counted twice below): seven estrogen receptor (ER) positive tamoxifen-treated cohorts ($n = 923$), six node-negative untreated cohorts ($n = 824$), and five cohorts with mixed adjuvant treatment ($n = 446$). A summary of these data sets, including information on the data curation process and clinical pathologic characteristics, can be found in Heimes et al. (2017) (for node-negative untreated breast cancers), Hellwig et al. (2016) and Marchan et al. (2017) (for ER-positive tamoxifen-treated breast cancers and all other cancer types). In addition, 14 neoadjuvant treated breast cancer cohorts ($n = 1045$) with available binary endpoint (response to treatment) instead of survival endpoint, were collected.

In this thesis, only NSCLC cohorts are used as application example in sections 4.1.3 and 4.2.5. Once restricting data to the availability of clinical variables of age at time of diagnosis, sex, pTNM stage, histology and smoking status, only five out of ten cohorts remain (GSE4573, GSE29013, GSE31210, GSE37745, GSE50081). For GSE4573 only preprocessed gene expression data normalized with the MAS5 algorithm are available that differ from the fRMA-normalized expression data of the other cohorts. Therefore, GSE4573 is removed from subsequent analysis. Expression data of the remaining four NSCLC cohorts were measured on the Affymetrix HG-U133 Plus 2.0 array comprising 54 675 probe sets that represent genes (see section 2.3 for more information). The majority of these 54 675 probe sets represent noise and do not contain relevant information regarding survival outcome. This makes identification of the prognostic genes challenging and slows down computation time. Therefore, two additional preselected gene sets are considered for analysis. One subset is defined by the 1000 probe sets with the highest variability in gene expression values across all cohorts, referred to as top-1000-variance genes. This selection is based on the assumption that relevant prognostic genes imply systematic changes in their expression values and thus, a larger variance in contrast to irrelevant noise genes. Alternatively, a literature-based selection of prognostic genes from the following two publications is considered. Both publications use training and validation data independent of the four NSCLC cohorts used in this thesis.

Kratz et al. (2012) develop and validate a 14-gene expression assay based on quantitative PCR in patients with non-squamous NSCLC. The assay improves prognostic accuracy for patients with early-stage non-squamous NSCLC at high risk for mortality after surgical resection. The assay comprises eleven cancer-related target genes (BAG1, BRCA1, CDC6, CDK2AP1, ERBB3, FUT3, IL11, LCK, RND3, SH3BGR, WNT3A) and three reference genes (ESD, TBP, YAP1). Genes were selected from literature review and previously published microarray and PCR-based studies described in a prior

study. Many of the cancer-related genes are involved in classical oncogenic pathways and all of them are intricately related to molecular lung cancer pathways. Tang et al. (2017) perform a systematic literature review and meta-analysis-based evaluation of published prognostic signatures for NSCLC. The review includes 42 gene signatures derived from genome-wide mRNA gene expression microarray studies. Expression levels of all single genes belonging to the original signature are combined using supervised principal component analysis based on a training data set. Prediction performance of each signature is assessed by a meta-analysis of all test data sets. The performance of the published signatures compared to random signatures is evaluated by a linear mixed-effect model, and the prognostic power independent of clinical variables is assessed by multivariate Cox models. A separate analysis of the histological types adenocarcinoma (ADC) and squamous cell carcinoma (SQC) results in 17 and 8 prognostic signatures for ADC and SQC, respectively (a total of 20 different signatures as 5 signatures are prognostic in both ADC and SQC). These signatures significantly outperform random signatures and remain prognostic after adjusting for clinical risk factors.

In this thesis, the 14 genes from Kratz et al. (2012) and all genes belonging to the 20 prognostic signatures from Tang et al. (2017) are combined to one prognostic gene list. This list includes only single genes and ignores how the genes were combined numerically in the original signatures (using statistical models). Gene symbols are translated into corresponding probe set IDs of the Affymetrix HG-U133 Plus 2.0 array using the R/Bioconductor annotation package `hgu133plus2.db` (version 3.2.3). Not all genes have a match on this array. Thus, a reduced gene list comprising 3429 different probe sets that are related to 1323 different genes is used for analysis and referred to as prognostic gene list.

Overall survival and gene expression data of the four NSCLC cohorts (GSE29013, GSE31210, GSE37745, GSE50081) analyzed in this thesis are illustrated in Figure 2.1 by a Kaplan-Meier plot of the estimated survival functions in each cohort, as well as by a PCA (Principal Component Analysis) plot based on the expression data of the selected prognostic genes in all cohorts. The Kaplan-Meier plot shows that patients in cohort GSE31210 have the best prognosis with a 10-year overall survival probability of about 75%, while GSE37745 exhibits the worst prognosis with a 10-year overall survival of 25%. GSE29013 has the shortest maximum follow-up time with about 7 years, and GSE37745 the longest maximum follow-up time with more than 15 years. A PCA plot helps to identify patterns in large multivariate data sets. The aim is to reduce the dimensionality of the data without losing information in the data. The original variables are transformed into a smaller number of uncorrelated principal components that are sorted so that they explain a decreasing proportion of the total variance. In other words, the first principal component explains most of the variance in the data and thus, contains the largest amount of information, whereas the last principal component is the least important. The first two principal components in Figure 2.1 explain 26.7% and 15.4% of the total variability in the data (axis labels). Each patient is represented by a point in the PCA plot and colors refer to the different cohorts. The arrangement of the data points in the PCA plot indicates that patients within each cohort cluster together, and cohorts GSE37745 and GSE50081 are almost inseparable. In contrast, GSE31210 and GSE29013 can be well distinguished from all other cohorts in the direction of both principal components, with GSE29013 being further away from all other cohorts. However, the proportion of explained variance by the first two principal components is

only moderate (less than 50% of the total variance). Similar PCA plots are obtained when using gene expression data of all available probe sets or of the top-1000-variance probe sets.

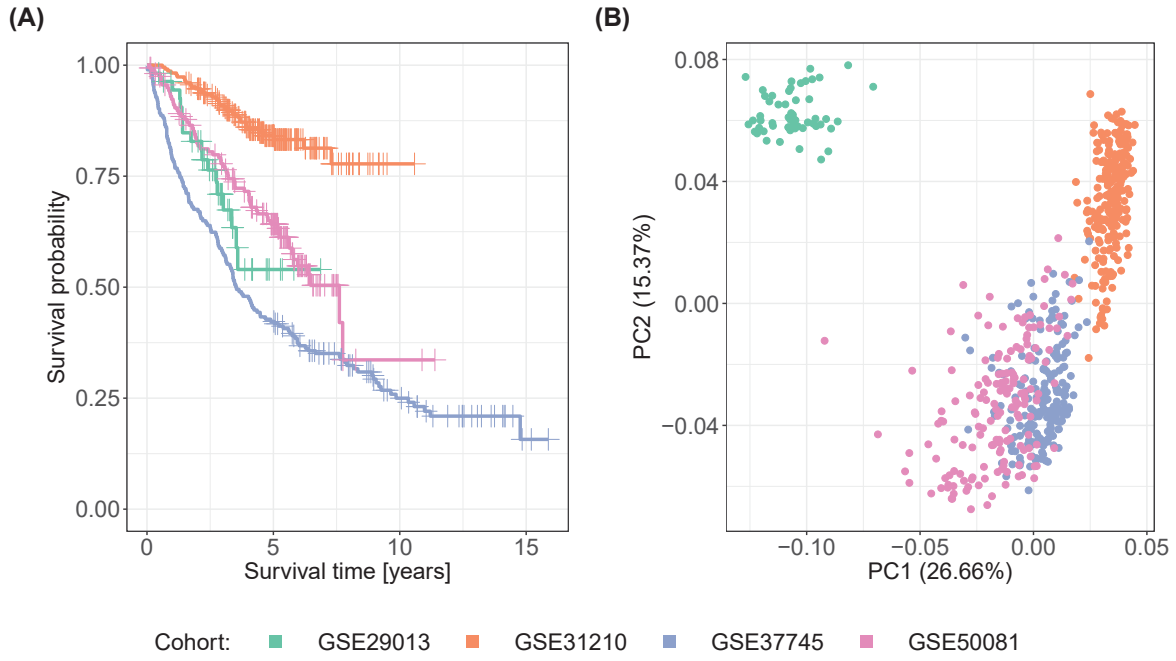


FIGURE 2.1: *(A) Kaplan-Meier plot of estimated survival functions for all lung cancer cohorts. (B) PCA (Principal Component Analysis) plot based on expression data of prognostic genes in all lung cancer cohorts.*

A summary of clinical pathologic variables of all cohorts is presented in Table C.1 (see Appendix). Mean age at diagnosis ranges between 60 and 68 years. The proportion of male patients in cohort GSE29013 is 69%; in all other cohorts it is approximately 50%. About two thirds of patients have stage I NSCLC, except for GSE29013 where stage I tumors make up less than half of all tumors. GSE31210 includes only ADC in contrast to the other cohorts where ADC form the largest histological type, followed by SQC, and other NSCLC. Apart from GSE31210, the vast majority of patients are former or current smokers. Censoring rates vary substantially between 26% and 85% across all cohorts.

2.3 Affymetrix gene expression microarrays

The genetic information of most organisms is encoded in deoxyribonucleic acid (DNA) consisting of a long sequence of nucleotides. Each nucleotide comprises a phosphate, a sugar (deoxyribose) and one of four different bases: adenine, cytosine, guanine and thymine. DNA is made of two complementary strands of nucleotides that are connected via hydrogen bonds between pairs of bases and form a double helix. The complementary bases cytosine and guanine bind together, as do adenine and thymine. DNA consists of sequences that code for proteins (referred to as genes), as well as non-coding sequences. Ribonucleid acid (RNA) differs from DNA in its sugar molecule (ribose), the base thymine is replaced by uracil, and RNA is single-stranded rather than double-stranded.

There are non-coding types of RNA and messenger RNA (mRNA) that code for proteins. Gene expression describes the transfer of genetic information from DNA for protein synthesis. This process takes place in two stages: the transcription of DNA into complementary mRNA, and the translation of mRNA into proteins (Göhlmann and Talloen, 2009, chapter 1). An illustration of this process is provided in Figure 2.2.

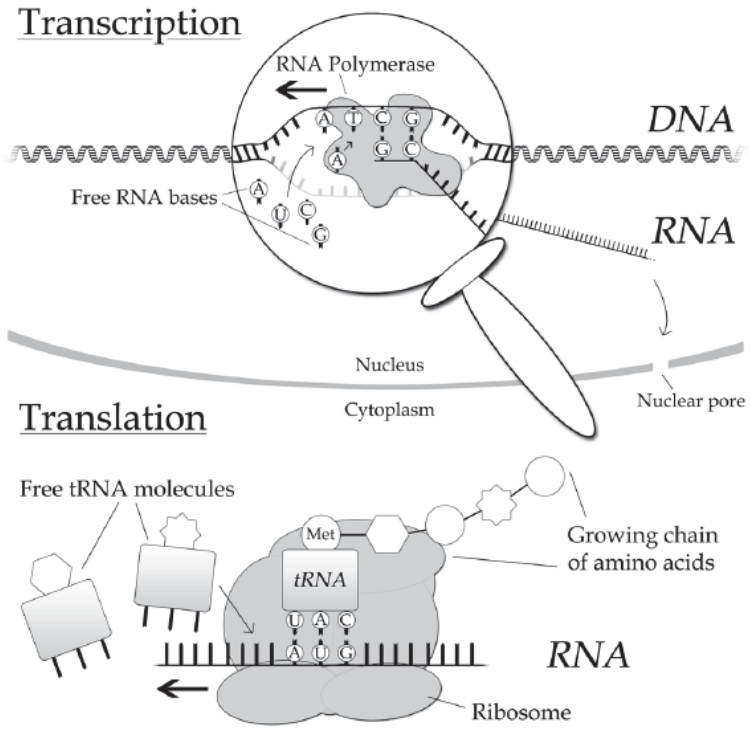


FIGURE 2.2: *Transcription of DNA to mRNA, and translation of mRNA into proteins (chains of amino acids) (Melzer et al., 2013).*

The development of microarray technology for gene expression measurement began in the late 1980s. Rapid progress in the 1990s allowed the simultaneous measurement of expression levels of many thousands of genes in a biological sample. The technology became increasingly popular and widely used in many fields of biomedical research (Bumgarner, 2013). This thesis focuses on the Affymetrix GeneChip® system based on high-density oligonucleotide arrays produced by Affymetrix. Affymetrix gene expression microarrays are composed of oligonucleotides (sequences of 25 nucleotides), referred to as probes, that are artificially synthesized onto the array surface. Each probe is present in millions of copies, all located at the same position on the array called probe cell. There are two types of probes: perfect match (PM) and mismatch (MM) probes. The PM probe is exactly complementary to the sequence of interest and measures the expression of the target gene transcript. The MM probe differs from the PM probe only by a substituted middle base, with the intention of quantifying non-specific hybridization and background signal detected by the corresponding PM probe. A PM and its corresponding MM probe constitute a probe pair and are adjacent on the array. Groups typically comprising 11 PM probes are referred to as probe sets and represent genes. Each array contains tens of thousands of probe sets (Bolstad, 2004; Göhlmann and Talloen, 2009, chapter 2). Affymetrix HG-U133 Plus 2.0 gene expression array data

are used in this thesis. This array comprises all probe sets from the previous generation HG-U133A and HG-U133B arrays, and nearly 10 000 additional probe sets, making up 54 675 probe sets in total. It allows to measure transcription over the entire human genome in a single hybridization (Affymetrix, 2003b).

The measurement of gene expression begins with the isolation and preparation of target mRNA from a biological sample. Target mRNA is reversed transcribed into complementary double-stranded DNA (cDNA) which, in turn, is used to produce cRNA. Next, cRNA is amplified to have sufficient material for the array, biotin-labeled, and fragmented. The labeled cRNA fragments bind to complementary probes on the array. After hybridization, unbound fragments are washed out and the array is stained with fluorescence dye that binds to the biotin label on the cRNA. The amount of fluorescence is measured by a laser scanner and the pixels of the resulting image give the probe intensities for each probe cell (Bolstad, 2004; Affymetrix, 2003a). To define a measure of gene expression that represents the amount of hybridized RNA, the probe intensities matching a probe set have to be combined in an appropriate way, which is explained in the following section.

2.4 Preprocessing of gene expression data

Preprocessing is important to transfer information of probe intensities into gene expression values. The most popular and widely used preprocessing algorithm for Affymetrix gene expression microarrays is robust multiarray analysis (RMA) proposed by Irizarry et al. (2003). It performs three steps: background correction, normalization, and summarization of log-transformed PM values. Background correction is carried out for each array individually, with the intention of removing background signals caused for example by non-specific binding, incomplete washing or optical noise. Simple approaches are based on subtracting the MM probes from the PM probes. However, the use of MM probes has some drawback: MM probes may also detect target signals and can have higher signal intensities than PM probes. Apart from that, Irizarry et al. (2003) show that the resulting measures of gene expression fail to adjust for the probe effect (within-array variability between probes within a probe set). Instead, they propose an alternative measure that only makes use of the PM probes and ignores the MM probes. For background correction they assume that PM intensities can be modeled as a mixture of a normally distributed background signal and an exponentially distributed signal. The background corrected PM value corresponds to the conditional expectation of the signal given the PM value. A detailed explanation of the procedure can be found in Bolstad (2004).

The second step of the RMA algorithm is normalization that is required to remove variation between arrays in order to make them comparable. Sources of variation can be sample preparation and array processing such as labeling, hybridization, and scanning. Quantile normalization has demonstrated the best performance in a comparison of normalization methods for high-density oligonucleotide arrays (Bolstad et al., 2003), and therefore is used in the RMA algorithm. Quantile normalization forces the probe intensity distribution to be the same for all arrays by normalizing the probe intensities to a common set of quantiles. This requires computation of an average distribution as reference distribution (see Bolstad et al., 2003 for details).

The third step of the RMA procedure is the summarization of probe intensities in each probe set to one gene expression value per probe set. Irizarry et al. (2003) propose a parametric linear model that accounts for probe effects

$$Y_{ijn} = \theta_{in} + \phi_{jn} + \epsilon_{ijn}, \tag{2.1}$$

where Y_{ijn} is the background-adjusted, normalized and $\log_2$-transformed PM intensity of probe $j \in \{1, ..., J_n\}$ in probe set $n \in \{1, ..., N\}$ on array $i \in \{1, ..., I\}$, θ_{in} represents the $\log_2$ scale expression level of probe set n on array i, ϕ_{jn} represents the probe effect for the j-th probe of probe set n, and ϵ_{ijn} is the independent identically distributed measurement error with mean 0. For identifiability, the sum of probe effects within a probe set is constrained to zero. Median polish is used to robustly estimate the parameters. Of interest is the estimate of θ that gives the gene expression measure for the corresponding probe set and is referred to as robust multi-array average (RMA) (Irizarry et al., 2003).

The last two steps require the simultaneous analysis of multiple arrays. This has the disadvantage that data sets preprocessed separately are not comparable and batch effects (e.g. differences due to gene expression measurements at different time periods or in different laboratories) are not accounted for. Therefore, McCall, Bolstad, and Irizarry (2010) propose an adaption of the RMA algorithm, termed frozen RMA (fRMA), that allows the separate analysis of microarrays and later combination of data for analysis. In contrast to RMA, fRMA uses not only the information in the present data but also information from large publicly available microarray databases. The latter serves as training data to generate a reference distribution for quantile normalization and to precompute parameter estimates for the summarization step. The model in equation (2.1) is extended by adding the index k for batch effects and a random effect term γ_{jkn} for the variability in probe effects across batches

$$Y_{ijkn} = \theta_{in} + \phi_{jn} + \gamma_{jkn} + \epsilon_{ijkn}.$$

The variances of the random effect γ (between-batch variance) and the measurement error ϵ (within-batch variance) are probe specific and estimated from the fixed reference data along with ϕ. For summarization, first the global batch effect estimate is subtracted from each intensity. Then the log gene expression is estimated by a weighted average of the probes in each probe set, with weights being defined by an M-estimation method and divided by the sum of the two precomputed variance estimators (McCall, Bolstad, and Irizarry, 2010). The fRMA algorithm is implemented in the R/Bioconductor package `fRMA` and used for the preprocessing of gene expression data in this work.

Chapter 3

Statistical methods

This chapter provides an overview of statistical methods for the analysis of high-dimensional data with time-to-event endpoint and extensions to subgroup analysis. This includes the Cox proportional hazards model, the most popular regression model in survival analysis, regularization and variable selection techniques for dealing with a large number of covariates, as well as extensions of standard models to take subgroups into account. These methods are first described in the classical, frequentist setting along with a general introduction to survival analysis in section 3.1, and in section 3.3 adapted for the Bayesian context. The Bayesian model mainly differs from the frequentist model in an adapted version of the Cox partial likelihood with prior distributions for the parameters, and in that it relies on a variable selection prior instead of adding a penalty term to the likelihood as in classical regularization. Parameter estimation in the Bayesian model is performed by Markov Chain Monte Carlo sampling rather than by classical techniques such as maximum likelihood. Furthermore, both approaches differ in how they address the problem of considering heterogeneity in data due to known subgroups and borrowing information across subgroups. The frequentist model is based on a weighted version of the likelihood with individual weights estimated from the data. Derivation of weights follows a Bayesian idea but frequentist classification is used for their estimation as described in section 3.2. In contrast, the Bayesian model uses a graphical model in the variable selection prior to link all subgroups.

3.1 Survival Analysis

The aim of survival analysis is to model and predict the time until a specified event. In the applied field of medicine and biology this event may be the death of a patient, the remission or recurrence of a disease. A typical starting point for the time measurement is the date of diagnosis or primary treatment. A common feature of time-to-event data is the presence of censoring, where the exact length of time is unknown. Instead, it is only known that the event of interest occurred before the start of the study (left censoring) or has not occurred until a given time point (right censoring). The combination of both is called interval censoring. This thesis exclusively deals with right censoring, the most common type in clinical practice. Examples of this censoring type are a patient leaving the study before its end or being still alive at the end of the study.

This chapter begins with an introduction to central terms and basic functions in survival analysis. In section 3.1.2 the Cox proportional hazards model is explained that serves to predict a patient's survival function from a set of clinical and/or genomic covariates. This is followed by a weighted version of the likelihood to take subgroups

into account. In high-dimensional settings, where the number of covariates is typically much larger than the number of observations, conventional techniques for parameter estimation do not work and adaptions are required such as regularization methods. In section 3.1.3 two measures of prediction performance for model evaluation are described.

3.1.1 Basic quantities

Let T be the time until a specified event, in the following termed *survival* or *failure time*, and let C be the *censoring time*. T and C are non-negative random variables and assumed to be stochastically independent. A further assumption is that the distribution of C provides no information about the distribution of T, referred to as non-informative censoring. Let $\widetilde{T} = \min(T, C)$ denote the observed time until an event or censoring and let $\delta = \mathbb{1}(T \leq C)$ be a binary indicator that indicates whether a patient experienced an event ($\delta = 1$) or censoring ($\delta = 0$). Assume the data consist of n independent patients and for each patient, the tuple $(\tilde{t}_m, \delta_m)$ and the vector of covariates $\boldsymbol{x}_m = (x_{m1}, ..., x_{mp})' \in \mathbb{R}^p$ are observed, $m = 1, ..., n$. The covariates may consist of clinical variables such as age, sex, tumor grade or tumor size and a potentially large set of genomic variables.

3.1.1.1 Survival function

The distribution of the survival time T is characterized by two main functions, the survival function and the hazard rate. The survival function is the probability of an individual surviving beyond a time point t and defined as

$$S(t) = P(T > t) = 1 - F(t), \quad t \in [0, \infty)$$

with $F(t)$ the distribution function of T. $S(t)$ is continuous from the right and strictly decreasing with limits $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$ (Klein and Moeschberger, 2003, chapter 2.2).

The standard non-parametric estimator of the survival function is the Kaplan-Meier estimator (Kaplan and Meier, 1958). Let $t_{(1)} < t_{(2)} < ... < t_{(D)}$ be the ordered observed failure times and d_g the number of events at time $t_{(g)}$. The risk set $\mathcal{R}_g = \{k : \tilde{t}_k \geq t_{(g)}\}$ denotes the indices of patients who are at risk at time $t_{(g)}$ (who have not experienced an event yet immediately prior to $t_{(g)}$). $r_g = |\mathcal{R}_g|$ is the corresponding number of patients who are at risk. The Kaplan-Meier estimator is defined as

$$\hat{S}(t) = \begin{cases} 1, & \text{if } t < t_{(1)} \\ \prod_{t_{(g)} \leq t} \left(1 - \frac{d_g}{r_g}\right), & \text{if } t \geq t_{(1)} \,. \end{cases}$$

It is derived from the product of estimates of the conditional probabilities that a patient survives the next time point given that he has not had an event before. The Kaplan-Meier estimator is a step function with jumps at the observed failure times. If at least one patient is censored at the largest observation time t_{max} then $\hat{S}(t_{max}) > 0$. For t beyond the largest observation time $\hat{S}(t)$ is not well defined. The variance of the

Kaplan-Meier estimator can be estimated by Greenwood's formula

$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \cdot \sum_{t_{(g)} \leq t} \frac{d_g}{r_g(r_g - d_g)}.$$

An approximate pointwise $(1-\alpha)\cdot 100\%$ confidence interval for the survival function at time t is defined by

$$\hat{S}(t) \pm u_{(1-\alpha/2)} \cdot \left[\widehat{Var}(\hat{S}(t))\right]^{1/2},$$

where $u_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ percentile of the standard normal distribution (Klein and Moeschberger, 2003, chapters 4.2 and 4.3). In this thesis, the R package `survival` (version 2.41-3) is used to estimate the survival function and to plot Kaplan-Meier curves.

3.1.1.2 Hazard rate

The hazard rate or function $h(t)$ is the risk of a patient to experience the event at time t given that the event has not occurred before

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \quad h(t) \geq 0 \; \forall t \in [0, \infty).$$

If T is continuous, the following relationship between hazard rate and survival function holds true

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{\mathrm{d}}{\mathrm{d}t}F(t)}{S(t)} = \frac{-\frac{\mathrm{d}}{\mathrm{d}t}S(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t}\ln(S(t)).$$

The cumulative hazard rate for continuous survival times is defined as

$$H(t) = \int_0^t h(u)\mathrm{d}u = \int_0^t \frac{f(u)}{S(u)}\mathrm{d}u = \int_0^t \frac{\frac{\mathrm{d}}{\mathrm{d}u}F(u)}{S(u)}\mathrm{d}u = -\int_0^t \frac{\frac{\mathrm{d}}{\mathrm{d}u}S(u)}{S(u)}\mathrm{d}u = -\ln S(t)$$

and thus,

$$S(t) = \exp(-H(t)) = \exp\left(-\int_0^t h(u)\mathrm{d}u\right)$$

(Klein and Moeschberger, 2003, chapter 2.3). An estimator of the cumulative hazard rate is the Nelson-Aalen estimator proposed by Nelson (1972) and Aalen (1978)

$$\hat{H}(t) = \sum_{t_{(g)} \leq t} \hat{h}(t_{(g)}) = \sum_{t_{(g)} \leq t} \frac{d_g}{r_g}.$$

Due to the relationship between cumulative hazard rate and survival function, an alternative estimator of the survival function is given by $\hat{S}(t) = \exp(-\hat{H}(t))$ (Klein and Moeschberger, 2003, chapter 4.2).

3.1.2 Cox proportional hazards model

The Cox proportional hazards model developed by Cox (1972) is the most popular regression model used in survival analysis. It models the hazard rate $h(t|\boldsymbol{x}_m)$ of an individual at time t and consists of two terms, the non-parametric baseline hazard rate

$h_0(t)$ and a parametric form of the covariate effect

$$h(t|\boldsymbol{x}_m) = h_0(t) \cdot \exp(\boldsymbol{\beta}'\boldsymbol{x}_m) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p} \beta_i x_{mi}\right). \tag{3.1}$$

$\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is the unknown parameter vector that represents the strength of influence of the covariates on the hazard rate and $\boldsymbol{x}_m \in \mathbb{R}^p$ is the observed vector of covariates.

The Cox model assumes proportional hazard rates, meaning that the ratio of hazard rates for two patients with covariate vectors $\boldsymbol{x}_m$ and $\boldsymbol{x}_{m'}$ is constant over time

$$\frac{h(t|\boldsymbol{x}_m)}{h(t|\boldsymbol{x}_{m'})} = \exp\left(\sum_{i=1}^{p} \beta_i (x_{mi} - x_{m'i})\right).$$

This is equivalent to the assumption of time-independent covariates. Residual plots can be used to check if this assumption is valid (Klein and Moeschberger, 2003, chapters 8.1 and 11).

3.1.2.1 Likelihood

The regression coefficients β_i are traditionally estimated via maximum likelihood based on the partial likelihood. Let $t_{(1)} < t_{(2)} < ... < t_{(D)}$ be the ordered observed event times, $\boldsymbol{x}_{(g)}$ the observed vector of covariates for a patient with event time $t_{(g)}$ and $\mathcal{R}_g = \{k : \tilde{t}_k \geq t_{(g)}\}$ the set of patients who are at risk at time $t_{(g)}$. The partial likelihood is derived from the product of the conditional probabilities that a patient with covariates $\boldsymbol{x}_{(g)}$ experiences an event at time $t_{(g)}$ given one of the patients in $\mathcal{R}_g$ has an event at this time

$$L(\boldsymbol{\beta}) = \prod_{g=1}^{D} \frac{h(t_{(g)}|\boldsymbol{x}_{(g)})}{\sum_{k\in\mathcal{R}_g} h(t_{(g)}|\boldsymbol{x}_k)} = \prod_{g=1}^{D} \frac{\exp(\sum_{i=1}^{p} \beta_i x_{(g)i})}{\sum_{k\in\mathcal{R}_g} \exp(\sum_{i=1}^{p} \beta_i x_{ki})}. \tag{3.2}$$

The partial log-likelihood is defined as

$$l(\boldsymbol{\beta}) = \ln(L(\boldsymbol{\beta})) = \sum_{g=1}^{D}\sum_{i=1}^{p} \beta_i x_{(g)i} - \sum_{g=1}^{D} \ln\left[\sum_{k\in\mathcal{R}_g} \exp\left(\sum_{i=1}^{p} \beta_i x_{ki}\right)\right].$$

The problem of maximizing $l(\boldsymbol{\beta})$ can be solved numerically using a Newton-Raphson technique (Cox, 1972; Klein and Moeschberger, 2003, chapter 8.3).

In the presence of ties (more events occur at the same observed event time), a modification of the partial likelihood is necessary. Two very popular methods for handling tied events are the ones suggested by Efron (1977) and Breslow (1974). Let d_g be the number of events at time $t_{(g)}$ and $\mathcal{D}_g$ the set of patients with an event at this time. The partial likelihood according to Breslow is defined as

$$L_B(\boldsymbol{\beta}) = \prod_{g=1}^{D} \frac{\exp\left(\sum_{k\in\mathcal{D}_g} \boldsymbol{\beta}'\boldsymbol{x}_k\right)}{\left[\sum_{k\in\mathcal{R}_g} \exp(\boldsymbol{\beta}'\boldsymbol{x}_k)\right]^{d_g}}.$$

When the number of ties is large, the partial likelihood proposed by Efron is more precise

$$L_E(\boldsymbol{\beta}) = \prod_{g=1}^{D} \frac{\exp\left(\sum_{k\in\mathcal{D}_g} \boldsymbol{\beta}'\boldsymbol{x}_k\right)}{\prod_{l=1}^{d_g}\left[\sum_{k\in\mathcal{R}_g}\exp(\boldsymbol{\beta}'\boldsymbol{x}_k) - \frac{l-1}{d_g}\sum_{k\in\mathcal{D}_g}\exp(\boldsymbol{\beta}'\boldsymbol{x}_k)\right]}.$$

When no ties are present between the event times, then $L(\boldsymbol{\beta}) = L_E(\boldsymbol{\beta}) = L_B(\boldsymbol{\beta})$.

After the estimation of the regression coefficients based on the partial likelihood, the baseline hazard rate $h_0(t)$ is estimated from the complete censored-data likelihood. For each patient, the triple $(\tilde{t}_m, \delta_m, \boldsymbol{x}_m)$ is observed, $m = 1, ..., n$. The likelihood function for right-censored data is defined as

$$\begin{aligned} L(\boldsymbol{\beta}, h_0(\tilde{\boldsymbol{t}})) &= \prod_{m=1}^{n} f(\tilde{t}_m|\boldsymbol{x}_m)^{\delta_m} S(\tilde{t}_m|\boldsymbol{x}_m)^{1-\delta_m} = \prod_{m=1}^{n} h(\tilde{t}_m|\boldsymbol{x}_m)^{\delta_m}\exp\left(-H(\tilde{t}_m|\boldsymbol{x}_m)\right) \\ &= \prod_{m=1}^{n} h_0(\tilde{t}_m)^{\delta_m}\exp(\boldsymbol{\beta}'\boldsymbol{x}_m)^{\delta_m}\exp\left(-H_0(\tilde{t}_m)\exp(\boldsymbol{\beta}'\boldsymbol{x}_m)\right), \end{aligned}$$

where $H_0(t) = \sum_{t_{(g)}\le t} h_0(t_{(g)})$ is the discrete cumulative baseline hazard rate and $h_0(\tilde{\boldsymbol{t}}) = (h_0(\tilde{t}_1), ..., h_0(\tilde{t}_n))'$ with $h_0(t) = 0\ \forall t \notin \{t_{(1)}, ..., t_{(D)}\}$.

For fixed $\boldsymbol{\beta}$ the likelihood can be maximized as a function of $h_0(t)$ only, yielding an estimator of the baseline hazard rate. Inserting the partial maximum likelihood (ML) estimator $\hat{\boldsymbol{\beta}}$ into the likelihood results in

$$\begin{aligned} L(h_0(\tilde{\boldsymbol{t}})) &= \left(\prod_{g=1}^{D} h_0(t_{(g)})\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_{(g)})\right)\exp\left(-\sum_{m=1}^{n} H_0(\tilde{t}_m)\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_m)\right) \\ &= \left(\prod_{g=1}^{D} h_0(t_{(g)})\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_{(g)})\right)\exp\left(-\sum_{m=1}^{n}\sum_{t_{(g)}\le\tilde{t}_m} h_0(t_{(g)})\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_m)\right) \\ &= \left(\prod_{g=1}^{D} h_0(t_{(g)})\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_{(g)})\right)\exp\left(-\sum_{g=1}^{D}\sum_{k\in\mathcal{R}_g} h_0(t_{(g)})\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_k)\right) \\ &= \prod_{g=1}^{D}\left[h_0(t_{(g)})\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_{(g)})\exp\left(-\sum_{k\in\mathcal{R}_g} h_0(t_{(g)})\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_k)\right)\right] \\ &= \prod_{g=1}^{D}\left[h_0(t_{(g)})\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_{(g)})\exp\left(-h_0(t_{(g)})\sum_{k\in\mathcal{R}_g}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_k)\right)\right] \\ &\propto \prod_{g=1}^{D} h_0(t_{(g)})\exp\left(-h_0(t_{(g)})\sum_{k\in\mathcal{R}_g}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_k)\right). \end{aligned}$$

Differentiating the likelihood with respect to $h_0(t_{(g)})$ and setting the derivative equal to zero, results in the ML estimator for the baseline hazard rate at time $t_{(g)}$

$$\hat{h}_0(t_{(g)}) = \frac{1}{\sum_{k\in\mathcal{R}_g}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_k)}$$

(Klein and Moeschberger, 2003, chapters 3.5 and 8.3) .

Consequently, Breslow's estimator of the cumulative baseline hazard rate $H_0(t)$ is defined as

$$\hat{H}_0(t) = \sum_{t_{(g)} \leq t} \left(\frac{d_g}{\sum_{k \in \mathcal{R}_g} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_k)} \right) .$$

Following from the relationship between cumulative hazard rate and survival function in subsection 3.1.1.2, Breslow's estimator provides an estimator of the baseline survival function of a patient with covariate vector $\boldsymbol{x}_m = \mathbf{0}_p$: $\hat{S}_0(t) = \exp(-\hat{H}_0(t))$. For a patient with covariate vector $\boldsymbol{x}_m$ an estimate of the individual survival function is given by

$$\begin{aligned} \hat{S}(t|\boldsymbol{x}_m) &= \exp\left(-\hat{H}(t|\boldsymbol{x}_m)\right) \\ &= \exp\left(-\hat{H}_0(t) \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_m)\right) \\ &= \exp\left(\ln(\hat{S}_0(t)) \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_m)\right) \\ &= \hat{S}_0(t)^{\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_m)} \end{aligned}$$

(Klein and Moeschberger, 2003, chapter 8.8).

3.1.2.2 Weighted partial likelihood

In the case of an unweighted partial likelihood, all patients contribute to the same extent to the estimation of the regression coefficients. This might not be desirable when the cohort is heterogeneous due to known subgroups that are associated with different prognosis. In this context it is reasonable to fit a separate Cox model for each subgroup. This can be done by using only the data from the subgroup of interest or by including information from the other subgroups. For the latter, individual weights $w_m \in [0, 1]$, $m = 1, ..., n$ are introduced to the partial likelihood in order to vary the level of contribution of each patient. The weighted partial log-likelihood is given by

$$l(\boldsymbol{\beta}) = \sum_{g=1}^{D} w_g \boldsymbol{\beta}' \boldsymbol{x}_{(g)} - \sum_{g=1}^{D} w_g \ln \left[\sum_{k \in \mathcal{R}_g} w_k \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_k\right) \right]$$

(Weyer and Binder, 2015). An approach for estimating individual weights from the data is proposed in section 3.2. Alternatively, fixed weights can be used as suggested in Weyer and Binder (2015). The idea is to focus on a specific subgroup of patients and assign each of these patients a weight of 1. All other patients are down-weighted and receive a fixed weight $w \in (0, 1)$. Let $s_m \in \{1, ..., S\}$ be the observed subgroup indicator for patient m and s the subgroup of interest. The fixed weights are defined as

$$w_m = \begin{cases} 1, & \text{if } s_m = s \\ w, & \text{else.} \end{cases}$$

$w = 1$ results in the unweighted standard partial log-likelihood based on all patients (*combined analysis*), whereas $w = 0$ corresponds to a *subgroup analysis* with a standard partial log-likelihood based only on the patients in the subgroup of interest.

3.1.2.3 Regularization methods

In high-dimensional settings when the number of covariates p exceeds the number of observations n, the problem of maximizing the partial log-likelihood cannot be solved uniquely. A way to deal with the $p >> n$ situation is to introduce a penalty term into the partial log-likelihood $l(\boldsymbol{\beta})$, referred to as regularization. This approach is also reasonable in $p < n$ settings since it considers collinearity among the predictors and helps to prevent overfitting. The adapted maximization problem of the partial log-likelihood is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ l(\boldsymbol{\beta}) - \lambda_P \cdot \left(\alpha \|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1 \right) \right\} \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ l(\boldsymbol{\beta}) - \lambda_P \cdot \sum_{i=1}^{p} \left(\alpha \beta_i^2 + (1-\alpha)|\beta_i| \right) \right\}.\end{aligned}$$

$\alpha \in [0,1]$ specifies the type of penalization. $\alpha = 1$ corresponds to the ridge regression (Hoerl and Kennard, 1970; Verweij and Van Houwelingen, 1994), while $\alpha = 0$ results in the lasso regression (Tibshirani, 1996; Tibshirani, 1997). For $\alpha \in (0,1)$ the elastic net introduced by Zou and Hastie (2005) constitutes a compromise between both penalties. An advantage of the lasso regression is that it provides an automatic variable selection procedure by estimating many regression coefficients equal to zero. Ridge regression, on the other hand, does not give a sparse solution. It shrinks all regression coefficients towards zero, so that their estimates are close to but unequal to zero. However, ridge regression better handles correlated predictors by shrinking their coefficients towards each other and giving them equal weight. Lasso regression typically selects one of the correlated variables while ignoring the others (Simon et al., 2011). In comparative studies, ridge regression tends to yield better prediction performance than the lasso (Bøvelstad et al., 2007; Kammers et al., 2011). More theory and many extensions for the lasso can be found in Bühlmann and Geer (2011).

The complexity parameter $\lambda_P \geq 0$ controls the absolute size and thus the amount of penalization of the regression coefficients. The higher λ_P the stronger the level of shrinkage and the higher the number of estimated coefficients equal to zero in the case of lasso regression. λ_P is typically optimized via cross-validation. In this thesis, the regression coefficients are estimated via cyclical coordinate descent from a regularized partial log-likelihood. The corresponding algorithm is implemented in the R package `glmnet` (version 2.0-13) and described in the Appendix, section A.1. Figure 3.1 visualizes the estimation of regression coefficients in a Cox regression model with lasso and ridge penalty utilizing `glmnet`.

In recent years, boosting algorithms have been developed as alternative to regularization for dealing with high-dimensional data and providing sparse models. Boosting was originally designed for classification problems and later extended to regression. The general idea is to combine many weak learners to receive one strong, competitive learner. An overview of boosting techniques is given in Hastie, Tibshirani, and Friedman (2009, chapter 10). For survival analysis, two boosting types are established: gradient and likelihood-based boosting. The former is based on gradient descent by minimizing a specific loss function (Hothorn and Bühlmann, 2006; Bühlmann and Hothorn, 2007; Hothorn et al., 2006). The latter maximizes a penalized form of the partial log-likelihood (Tutz and Binder, 2006; Binder and Schumacher, 2008).

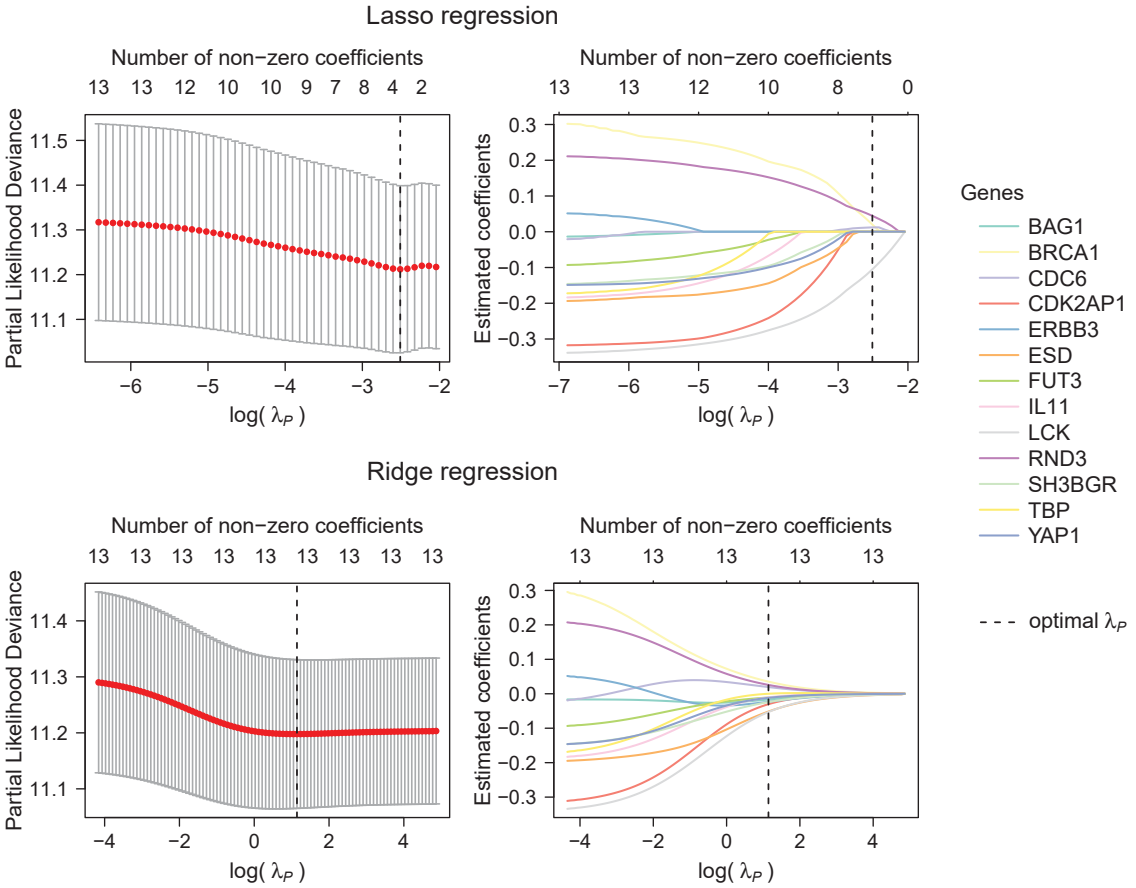


FIGURE 3.1: *Illustration of the lasso (upper row) and ridge (bottom row) regression as implemented in the R package* `glmnet`*. A penalized Cox model is fitted to 194 lung cancer patients with overall survival endpoint and 13 genomic predictors. The optimal* λ_P *is chosen by 10-fold cross-validation and gives the minimum mean cross-validated error.*

3.1.3 Model evaluation

In order to assess the prediction performance of a fitted model, it is important to validate it on independent test data. If the same (training) data are used for learning and evaluating a model, the estimated prediction error will generally be too optimistic and underestimate the true prediction error. This means that the model performs well on the training data but worse on independent test data. The more complex a model becomes, the more training data are used for learning, which makes the model more specific to the training data but less generalizable. This is called *overfitting* (Hastie, Tibshirani, and Friedman, 2009, chapter 7.2).

When no independent test data are available for validation, resampling can be applied. Resampling is the random (repeated) partitioning of the entire available data into training and test sets. The model is fitted on the training set and predictions are made on the test set. Two common resampling techniques that are used in this thesis are the K-fold cross-validation and subsampling. In K-fold cross-validation the data are randomly split into K disjoint subsets of approximately the same size. $K-1$ parts serve as training set and the remaining k-th part as test set. This is repeated

for $k = 1, ..., K$ and the K estimated prediction errors are combined to one mean prediction error. Typical choices of K are $K = 5, 10$ or $K = n$ termed leave-one-out cross-validation (Hastie, Tibshirani, and Friedman, 2009, chapter 7.10). Subsampling differs from bootstrap only in that training sets are drawn from the data without replacement. The data are randomly split into a training set $\mathcal{I}_k \subset \{1, ..., n\}$ and a test set $\{1, ..., n\} \backslash \mathcal{I}_k$. This is repeated K times. In this thesis $K = 100$ and 0.632 is the proportion of the training set which is proposed by Efron and Tibshirani (1997) as an improved bootstrap estimate of the prediction error. An overview of further resampling approaches can be found in Molinaro, Simon, and Pfeiffer (2005).

In sections 3.1.3.1 and 3.1.3.2 two different measures of prediction performance for time-to-event data are presented. They are particularly important for the comparison of two or more survival models with regard to their predictive accuracy. Another important criterion for model evaluation besides prediction performance is the stability of variable selection. A desirable property is that the set of selected covariates remains stable across different resampling data sets. Different measures of feature selection stability are proposed by He and Yu (2010), Lausser et al. (2013), Bommert, Rahnenführer, and Lang (2017), and with regard to stability of ranked gene lists by Boulesteix and Slawski (2009). In this thesis, the proportion of inclusion of each predictor in different resampling data sets (*resampling inclusion frequencies*) is used to judge variable selection stability and the importance of a variable (Meinshausen and Bühlmann, 2010; Sauerbrei, Boulesteix, and Binder, 2011; Weyer and Binder, 2015).

3.1.3.1 Brier score

The Brier score is originally developed by Brier (1950) for judging the inaccuracy of probabilistic weather forecasts and is adapted for time-to-event data by Graf et al. (1999) and Gerds and Schumacher (2006). The expected Brier score can be interpreted as a mean square error of prediction. It measures the inaccuracy by comparing the estimated survival probability $\hat{S}(t|\boldsymbol{x}_m)$ of a patient m, $m = 1, .., n$, with the true survival status $Y_m(t)$ based on a loss function

$$BS(t, \hat{S}) = \mathrm{E}\left[Y_m(t) - \hat{S}(t|\boldsymbol{X}_m = \boldsymbol{x}_m)\right]^2.$$

For the estimation of the expected Brier score the true survival status is replaced by the observed status $\mathbb{1}(\tilde{t}_m > t)$

$$\widehat{BS}(t) = \frac{1}{n} \sum_{m=1}^{n} \hat{w}_m(t) \cdot \left(\mathbb{1}(\tilde{t}_m > t) - \hat{S}(t|\boldsymbol{x}_m)\right)^2$$

and the squared residuals are weighted using inverse probability of censoring weights

$$\hat{w}_m(t) = \frac{\mathbb{1}(\tilde{t}_m \leq t)\delta_m}{\hat{C}(\tilde{t}_m)} + \frac{\mathbb{1}(\tilde{t}_m > t)}{\hat{C}(t)}$$

to adjust for the bias caused by the presence of censoring in the data. $\hat{C}(t)$ is the Kaplan-Meier estimator of the censoring times (Schumacher, Binder, and Gerds, 2007; Binder, Porzelius, and Schumacher, 2011).

Important benchmark values for the Brier score are $\frac{1}{3}$, $\frac{1}{4}$ and the Brier score of the Kaplan-Meier estimator of a null model without any covariates. The first value corresponds to estimating $S(t)$ by drawing random numbers from $[0, 1]$, since the expected value of a quadratic random variable with uniform distribution on $[0, 1]$ is equal to $\frac{1}{3}$. The second benchmark value comes from predicting 50% risk for everyone ($\hat{S}(t) \equiv \frac{1}{2}\ \forall t$) (Mogensen, Ishwaran, and Gerds, 2012).

The Brier score can be used to compare the predictive performance of competing survival models over time. This can be done by plotting prediction error curves. One model is better than another when its prediction error curve lies below the other. However, sometimes this is difficult to assess, especially when curves are crossing. Prediction error curves can be summarized with the integrated Brier score as a measure of inaccuracy over a time interval rather than at single time points (Graf et al., 1999)

$$IBS(t^*) = \frac{1}{t^*} \int_0^{t^*} BS(t)\mathrm{d}t, \quad t^* > 0.$$

3.1.3.2 C-index

The C- (concordance) index is a measure of predictive discrimination and defined as the proportion of all usable pairs of patients with concordant predicted and observed survival times. Let $\tilde{t}_m$, $\tilde{t}_{m^*}$ be the observed survival times of patients m and m^*, and $r(\boldsymbol{x}_m)$, $r(\boldsymbol{x}_{m^*})$ the corresponding risk functions. $r(\boldsymbol{x}_m)$ is estimated by the risk score $\hat{\boldsymbol{\beta}}'\boldsymbol{x}_m$. A pair (m, m^*) is considered concordant if $\tilde{t}_m \lesseqgtr \tilde{t}_{m^*} \Leftrightarrow r(\boldsymbol{x}_m) \gtreqless r(\boldsymbol{x}_{m^*})$. The C-index is defined as

$$CI = \frac{1}{n_c} \sum_{\{m:\, \delta_m = 1\}} \sum_{\{m^*:\, \tilde{t}_{m^*} > \tilde{t}_m\}} \left(\mathbb{1}(r(\boldsymbol{x}_{m^*}) < r(\boldsymbol{x}_m)) + \frac{1}{2}\mathbb{1}(r(\boldsymbol{x}_{m^*}) = r(\boldsymbol{x}_m)) \right),$$

where n_c is the number of comparable pairs (m, m^*) that standardizes CI to $[0, 1]$. A patient pair is considered unusable, if both patients die at the same time, or both patients are censored, or if one is censored before the other one dies. In the latter case, it is unknown whether the censored patient will outlive the one who died. $CI \approx 1$ stands for a very good prediction and values around 0.5 suggest a random prediction (Harrell, Lee, and Mark, 1996; Heagerty and Zheng, 2005; Uno et al., 2011).

3.2 Estimation of subgroup weights

A simple approach for the definition of subgroup weights is introduced in 3.1.2.2. It assigns the observations belonging to the subgroup of interest a maximum weight of one in the subgroup-specific likelihood, while down-weighting all other observations with a constant positive weight smaller than one. A much more flexible approach with individual weights for each patient is presented in this section. The aim is to appropriately weight all observations in each subgroup model rather than using only the data from the subgroup of interest. The weights match the distribution of the entire data to the distribution in each subgroup, such that a patient who is likely to belong to the subgroup of interest receives a higher weight in the subgroup-specific model. The idea goes back to Bickel et al. (2008) who apply this approach to logistic regression for modeling the success of HIV-drug therapies. In the following, the derivation of the weights is explained.

For each patient the vector of covariates $\boldsymbol{x}_m$, the response y_m and the subgroup membership $s_m \in \{1, ..., S\}$ are observed. In the case of time-to-event data, the response corresponds to the tuple $(\tilde{t}_m, \delta_m)$, where $\tilde{t}_m$ is the observed time until an event or censoring and δ_m is the event indicator. Assume the entire training data from all subgroups are summarized in $\boldsymbol{x}$ and $\boldsymbol{y}$. Let $\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))$ be an arbitrary loss function and $f_s(\boldsymbol{x})$ the predicted response based on the observed covariates in subgroup s. $f_s(\boldsymbol{x})$ should correctly predict the true response and thus minimize the expected loss with respect to the unknown joint distribution $p(\boldsymbol{y}, \boldsymbol{x}|s)$ for each subgroup s

$$\mathrm{E}_{p(\boldsymbol{y},\boldsymbol{x}|s)}[\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))].$$

The following equation shows that the expected loss for each subgroup equals the expected weighted loss with respect to the joint distribution of the pooled data from all subgroups $p(\boldsymbol{y}, \boldsymbol{x})$

$$\begin{aligned}
\mathrm{E}_{p(\boldsymbol{y},\boldsymbol{x}|s)}[\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))] &= \int p(\boldsymbol{y}, \boldsymbol{x}|s)\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} \\
&= \int \frac{p(\boldsymbol{y}, \boldsymbol{x}|s)}{p(\boldsymbol{y}, \boldsymbol{x})} p(\boldsymbol{y}, \boldsymbol{x})\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} \\
&= \mathrm{E}_{p(\boldsymbol{y},\boldsymbol{x})}[w_s(\boldsymbol{y}, \boldsymbol{x})\ell(\boldsymbol{y}, f_s(\boldsymbol{x}))].
\end{aligned}$$

The subgroup-specific weights for each observation are defined as $w_s(\boldsymbol{y}, \boldsymbol{x}) = \frac{p(\boldsymbol{y},\boldsymbol{x}|s)}{p(\boldsymbol{y},\boldsymbol{x})}$. They match the joint distribution $p(\boldsymbol{y}, \boldsymbol{x})$ of all subgroups to the target distribution $p(\boldsymbol{y}, \boldsymbol{x}|s)$ of subgroup s. Estimation of $w_s(\boldsymbol{y}, \boldsymbol{x})$ becomes difficult when $\boldsymbol{x}$ is high-dimensional. However, with Bayes' rule the potentially high-dimensional density ratio can be reformulated in terms of a conditional distribution with a single variable

$$w_s(\boldsymbol{y}, \boldsymbol{x}) = \frac{p(\boldsymbol{y}, \boldsymbol{x}|s)}{p(\boldsymbol{y}, \boldsymbol{x})} = \frac{p(\boldsymbol{y}, \boldsymbol{x})p(s|\boldsymbol{y}, \boldsymbol{x})}{p(\boldsymbol{y}, \boldsymbol{x})p(s)} = \frac{p(s|\boldsymbol{y}, \boldsymbol{x})}{p(s)}, \quad p(s) > 0.$$

$p(s)$ can be estimated by the relative frequency of subgroup s and $p(s|\boldsymbol{y}, \boldsymbol{x})$ can be considered as a multi-class classification problem (Bickel et al., 2008). The next two subsections introduce two multi-class classification approaches, multinomial logistic regression and random forests, that will be used for the estimation of $p(s|\boldsymbol{y}, \boldsymbol{x})$.

Bickel et al. (2008) and Netzer (2013) estimate these weights based on the entire training data, which leads to overfitting (as shown in section 4.1.2). To solve this problem, cross-validation is applied to the training data to obtain predictions for $p(s|\boldsymbol{y}, \boldsymbol{x})$. In the following, the subgroup membership is considered as response and the observed data $(y_m, \boldsymbol{x}'_m)$ for patient m are summarized in a new q-dimensional covariate vector $\boldsymbol{z}_m = (y_m, \boldsymbol{x}'_m)'$.

3.2.1 Multinomial logistic regression

Assume the training data are an i.i.d. sample consisting of tuples $(s_1, \boldsymbol{z}_1)$,..., $(s_n, \boldsymbol{z}_n)$, where $s_m \in \{1, ..., S\}$ is a categorical response and $\boldsymbol{z}_m \in \mathbb{R}^q$ a vector of covariates. The aim of multi-class classification is to find a classification rule based on the training data, that assigns a new observation based on its covariate values to one of the S classes (here subgroups). Multinomial logistic regression models the posterior probability of each class, which is the conditional probability of belonging to class s given the observed covariates $\boldsymbol{z}$: $P(\mathcal{S} = s|\boldsymbol{z})$. The probabilities of the S classes are modeled via linear functions in $\boldsymbol{z}$, they have to lie in $[0, 1]$ and sum up to one. The following model equations satisfy these constraints and are termed log-odds or logit transformations

$$\log\left(\frac{P(\mathcal{S} = s|\boldsymbol{z})}{P(\mathcal{S} = S|\boldsymbol{z})}\right) = \beta_{0s} + \boldsymbol{z}'\boldsymbol{\beta}_s, \quad s = 1, ..., S-1,$$

where $\boldsymbol{\beta}_s$ is a q-dimensional vector of unknown regression coefficients. The choice of the reference class in the denominator (here S) is arbitrary and leaves the estimates of β_{0s} and $\boldsymbol{\beta}_s$ unaffected. The model can be expressed in an alternative form by transforming the log-odds

$$P(\mathcal{S} = s|\boldsymbol{z}) = \frac{\exp(\beta_{0s} + \boldsymbol{z}'\boldsymbol{\beta}_s)}{1 + \sum_{r=1}^{S-1}\exp(\beta_{0r} + \boldsymbol{z}'\boldsymbol{\beta}_r)}, \quad s = 1, ..., S-1$$
$$P(\mathcal{S} = S|\boldsymbol{z}) = \frac{1}{1 + \sum_{r=1}^{S-1}\exp(\beta_{0r} + \boldsymbol{z}'\boldsymbol{\beta}_r)}$$

(Hastie, Tibshirani, and Friedman, 2009, chapter 4.4). A symmetric version of the model as in Zhu and Hastie (2004) is used for parameter estimation

$$P(\mathcal{S} = s|\boldsymbol{z}) = \frac{\exp(\beta_{0s} + \boldsymbol{z}'\boldsymbol{\beta}_s)}{\sum_{r=1}^{S}\exp(\beta_{0r} + \boldsymbol{z}'\boldsymbol{\beta}_r)}, \quad s = 1, ..., S.$$

This parametrization requires constraints, otherwise the model equations are not uniquely identifiable. However, regularization automatically eliminates redundancy in the parametrization and forces $\sum_{s=1}^{S}\boldsymbol{\beta}_s = (0, ..., 0)'$ (Friedman, Hastie, and Tibshirani, 2010; Hastie, Tibshirani, and Friedman, 2009, chapter 18.3.2). In this thesis the entire parameter set $\boldsymbol{\theta} = (\beta_{01}, \boldsymbol{\beta}'_1, ..., \beta_{0S}, \boldsymbol{\beta}'_S)' \in \mathbb{R}^{S(q+1)}$ is estimated via cyclical coordinate descent from a penalized log-likelihood. The corresponding algorithm is implemented in the R package `glmnet` (version 2.0-13) and described in the Appendix, section A.2.

3.2.2 Classification trees and random forests

In the following, the concept of trees is explained for classification problems, but with some adaptions it can be applied to regression problems with continuous or censored response. Suppose for each patient a categorical response $s_m \in \{1, ..., S\}$ and a q-dimensional covariate vector $\boldsymbol{z}_m$ are observed, $m = 1, ..., n$. Trees are obtained by recursive binary splitting of the covariate space. At each node the aim is to find the best splitting variable and the best split point. Best means in this case that the partition results in two preferably homogeneous (pure) subsets of observations. For classification problems a standard measure of node impurity is the Gini index

$$\sum_{s=1}^{S} \hat{p}_s(1 - \hat{p}_s) = 1 - \sum_{s=1}^{S} \hat{p}_s^2,$$

where $\hat{p}_s$ is the relative frequency of class s. Starting with all observations and considering a splitting variable i and split point c, the resulting two subsets of observations are $\mathcal{Z}_1(i, c) = \{m | z_{mi} \leq c\}$ and $\mathcal{Z}_2(i, c) = \{m | z_{mi} > c\}$. Then the best pair (i, c) is sought that minimizes the Gini index

$$\min_{i,c} \left[1 - \sum_{s=1}^{S} \left(\frac{\sum_{l \in \mathcal{Z}_1(i,c)} \mathbb{1}(s_l = s)}{|\mathcal{Z}_1(i, c)|} \right)^2 + 1 - \sum_{s=1}^{S} \left(\frac{\sum_{l \in \mathcal{Z}_2(i,c)} \mathbb{1}(s_l = s)}{|\mathcal{Z}_2(i, c)|} \right)^2 \right].$$

After determination of the best split, the data are partitioned into two subsets and the splitting process is repeated on each of the two resulting nodes. This procedure, termed tree growing, continues until some stopping rule. A node that is not further partitioned is called terminal node or leaf. All observations contained in such a node are assigned to the majority class $s^* = \underset{s}{\operatorname{argmax}} \{\hat{p}_s\}$ in this node. A new observation that is predicted by the tree receives the class label of a certain terminal node when it satisfies the conditions at each split leading to this terminal node. Figure 3.2 provides an example of a classification tree. Typical stopping rules are the minimum number of observations in a node (node size) or the tree size (complexity). A large tree explains a lot of the structure in the data, but due to its complexity, it is very likely to be overfitted. In order to prevent overfitting, trees can be pruned. Therefore, leaves are iteratively removed if they do not significantly improve the prediction in cross-validation (Hastie, Tibshirani, and Friedman, 2009, chapter 9.2; Breiman et al., 1984). A review of the development and some of the major algorithms for classification and regression trees is provided by Loh (2014).

A problem with trees is their high variance. Small changes in the data may lead to a different series of splits. Random forests are a very popular approach that is less susceptible to overfitting, reduces the variance and thus increases stability. Random forest is an ensemble of trees. The prediction of a new observation is based on the majority vote from the predictions of the trees in the ensemble. This means that the class label predicted by the majority of trees is chosen. The idea is to use many weak learners to build a strong predictor that is more robust with respect to noise. The individual trees are constructed based on bootstrap samples (sampling with replacement) and for the split selection at each node of a tree a random subset of $q^* \leq q$ covariates is chosen. Consequently, the dependence among the individual trees is weakened and they become more diverse. The trees are grown large and not pruned afterward to allow

single trees to specialize (Hastie, Tibshirani, and Friedman, 2009, chapter 15; Breiman, 2001).

In this thesis, the R packages `ranger` (version 0.8.0) is used for random forests with the following default settings. Predictions are class probabilities for each observation. The estimated probabilities returned by each tree are averaged for the random forest probability estimate. The number of randomly chosen splitting variables in each node is $q^* = \left\lfloor \sqrt{q} \right\rfloor$, the minimum node size for the stopping rule is one and the number of individual trees in the ensemble of a random forest is 500 (Wright and Ziegler, 2017).

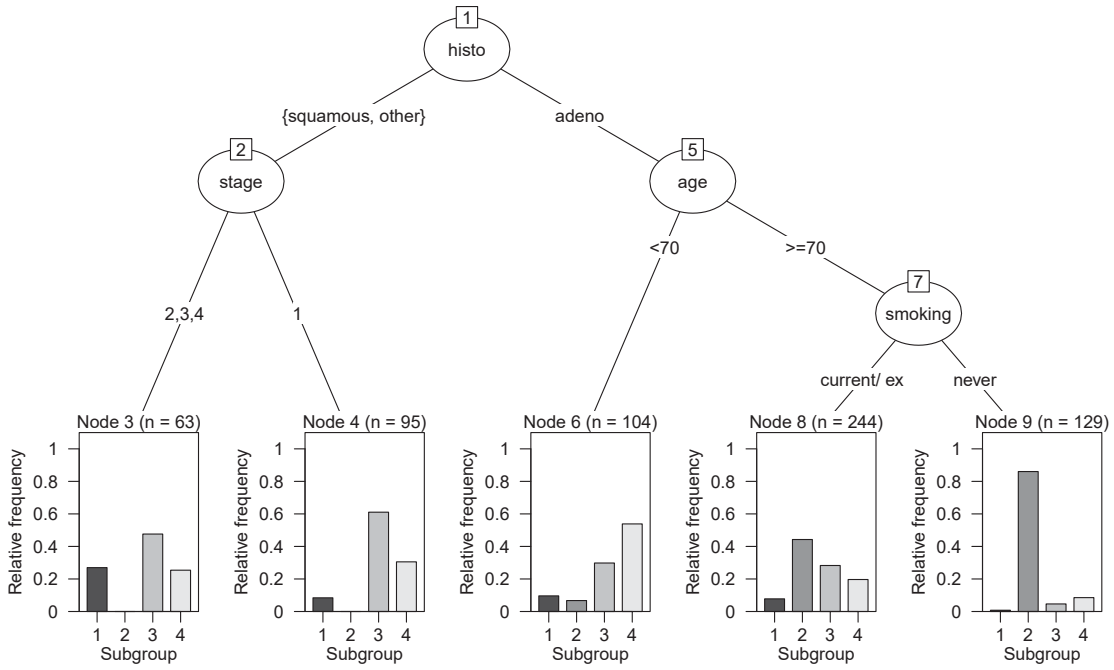


FIGURE 3.2: *Example of a classification tree to predict the subgroups (lung cancer cohorts) based on clinical covariates.*

3.2.3 Evaluation of classifier performance

In simulation studies in section 4.1.2 a distinction between two groups is of interest. For this reason, the following performance measures are introduced for the case of a binary classification problem with a binary response $\mathcal{S}$. Fawcett (2006) provides an overview of receiver operating characteristics (ROC) analysis for the assessment of classification performance and extensions for more than two classes.

Let $\hat{\pi}_s(\boldsymbol{z}_m) = \hat{P}(\mathcal{S} = s|\boldsymbol{z}_m)$ be the predicted probability of belonging to class $s \in \{0, 1\}$ for a patient with covariate vector $\boldsymbol{z}_m$. The predicted response for patient m is given by $\hat{s}_m = \underset{s\in\{0,1\}}{\text{argmax}}\{\hat{\pi}_s(\boldsymbol{z}_m)\}$. The overall *accuracy* compares the predicted response $\hat{s}_m$ with the true response s_m for each patient and is computed as the proportion of correctly classified patients

$$ACC = \frac{1}{n}\sum_{m=1}^{n} \mathbb{1}(s_m = \hat{s}_m).$$

A benchmark value is the naïve Bayes classifier that assigns all observations to the most frequent class. A disadvantage of the accuracy is that it does not distinguish between classes. Differences in the predictive performance with regard to single classes cannot be assessed.

Sensitivity and *specificity* allow a direct comparison of the predictions for each class. They compare the predicted probability rather than the predicted response with the true response. Therefore, the predicted probability has to be transformed into a binary decision rule based on a fixed threshold c

$$\hat{s}_m = \begin{cases} 1, & \text{if} \quad \hat{\pi}_1(z_m) > c \\ 0, & \text{else,} \end{cases}$$

where $s = 1$ is regarded as positive class (class of interest) and $s = 0$ is considered the negative or reference class.

The conditional probability of assigning a patient to class 1 given that he belongs to this class is denoted as sensitivity SE and can be estimated by the proportion of positive patients that are correctly identified as such (true positive rate)

$$SE = P(\hat{\mathcal{S}} = 1 | \mathcal{S} = 1), \quad \widehat{SE} = \frac{\sum_{m=1}^{n} \mathbb{1}(\hat{s}_m = 1 \ \wedge \ s_m = 1)}{\sum_{m=1}^{n} \mathbb{1}(s_m = 1)}.$$

The conditional probability of assigning a patient to class 0 given that he belongs to this class is denoted as specificity SP and can be estimated by the proportion of negative patients that are correctly classified as such (true negative rate)

$$SP = P(\hat{\mathcal{S}} = 0 | \mathcal{S} = 0), \quad \widehat{SP} = \frac{\sum_{m=1}^{n} \mathbb{1}(\hat{s}_m = 0 \ \wedge \ s_m = 0)}{\sum_{m=1}^{n} \mathbb{1}(s_m = 0)}.$$

Sensitivity and specificity depend on the choice of the threshold c. A ROC curve directly compares both measures across different thresholds. It provides a graphical overview of the diagnostic ability of a classifier and visualizes its discriminatory power with regard to each class. It is a more informative measure than overall accuracy or error rate and particularly useful in the presence of skewed class distributions or unequal misclassification costs. The ROC curve is created by plotting the true positive rate (SE) against the false positive rate (1-SP) at various thresholds. The diagonal line represents random guessing of the two classes. A good classification means both high sensitivity and high specificity, leading to a ROC curve well above the diagonal. The information of a ROC curve can be summarized in a single scalar value, the area under the ROC curve (AUC). The AUC is independent of the threshold and is suitable for the comparison of two or more classifiers. $AUC \approx 1$ stands for a very good diagnostic ability and values around 0.5 suggest a random prediction (Fawcett, 2006).

3.2.4 Imbalanced classification

In recent years, the problem of learning from imbalanced data has arisen as a new statistical challenge. He and Garcia (2009) provide a comprehensive review of the development of research on this topic, including the description of the problem, state-of-the-art solutions and evaluation metrics for their assessment. The problem generally

refers to data sets where the class distribution is heavily skewed and one class outrepresents the other(s). Learning algorithms often expect balanced class distributions and when confronted with imbalanced data, they may result in unfavorable, imbalanced accuracies across classes. The minority class tends to have much lower accuracy than the majority class. When sample size across classes is large and the minority class is not rare in its own but rather relative to the majority class, accuracy is generally not much affected by this so-called relative imbalance. However, in cases where the minority class is rare with a limited number of observations, learning becomes challenging. Rules over the sample space formed by the learning algorithms may be too specific and not generalizable, leading to overfitting. Besides new algorithms, the problem requires more informative evaluation measures than overall accuracy or error rate. He and Garcia (2009) suggest assessment metrics such as ROC curves. In the following, two solutions for imbalanced learning, *random oversampling* and *synthetic minority oversampling technique*, are described. Both are sampling methods that modify an imbalanced data set so that it becomes balanced with possibly improved classification performance. They were developed for binary classification problems and are adapted to multiple classes in the scope of this thesis. In the case of more than two classes, the largest class corresponds to the majority class and all other classes are treated as minority classes, such that the algorithms below are applied to each of the minority classes in order to balance all sample sizes.

Randomly sampling a set of observations with replacement from the minority class and adding them to the data set is called random oversampling. It increases the sample size in the minority class to balance the class distribution. In random undersampling, a set of observations from the majority class is randomly selected and removed from the data set to reduce sample size in the majority class. Both methods are very simple but have their shortcomings. Undersampling leads to a loss of information with respect to the majority class, while oversampling adds replicated data that may result in overfitting (He and Garcia, 2009).

Synthetic minority oversampling technique (SMOTE) creates artificial observations based on the feature space in the minority class. Let $\mathcal{I}_{min}$ be the index set of observations belonging to the minority class and $\boldsymbol{z}_m$ the q-dimensional feature vector of observation m, $m = 1, ..., n$. For each observation $m \in \mathcal{I}_{min}$ define the K-nearest neighbors as the K observations in $\mathcal{I}_{min}$ with the smallest distance in the q-dimensional feature space. One of the K-nearest neighbors m' is randomly chosen and a new synthetic observation $\boldsymbol{z}_{new}$ is created by interpolation

$$\boldsymbol{z}_{new} = \boldsymbol{z}_m + (\boldsymbol{z}_{m'} - \boldsymbol{z}_m) \cdot \zeta,$$

where ζ is randomly drawn from a uniform distribution in the interval [0,1] and $m, m' \in \mathcal{I}_{min}$. Thus, the new observation is a point along the line segment joining $\boldsymbol{z}_m$ and $\boldsymbol{z}_{m'}$. For categorical variables the factor level of the new artificial observation is sampled from the given levels of the two input observations m and m'. A drawback of SMOTE is that for each original minority observation the same number of synthetic observations is generated (not flexible) and the minority feature space is generalized without regard to the majority class. This may lead to class mixture or overlapping between classes (Chawla et al., 2011; He and Garcia, 2009).

In this thesis, the R package `mlr` (version 2.12) is used to apply random oversampling and SMOTE within cross-validation. When the feature space consists of continuous

covariates only, the Euclidean distance is used for K-nearest neighbor determination. K is chosen to be 5. However, in the case of both continuous and categorical covariates, Gower's dissimilarity coefficient (Gower, 1971) is applied to calculate the distance between observation m and m'

$$d(m, m') = \frac{\sum_{i=1}^{p} \zeta^i_{mm'} d^i_{mm'}}{\sum_{i=1}^{p} \zeta^i_{mm'}} \in [0, 1].$$

When variable i is missing in either or both observations m and m', then $\zeta^i_{mm'} = 0$. In all other cases $\zeta^i_{mm'} = 1$. $d^i_{mm'}$ is the distance between z_{mi} and $z_{m'i}$. For categorical variables, $d^i_{mm'} = 0$ if $z_{mi} = z_{m'i}$ and $d^i_{mm'} = 1$ otherwise. For numeric variables $d^i_{mm'}$ is the absolute difference of both values divided by the total range of variable i.

3.3 Bayesian subgroup analysis for high-dimensional survival data

In this chapter, a hierarchical Bayesian model is proposed that addresses the problem of identifying (genomic) predictors that are both relevant to the response (time-to-event endpoint) and related to each other in a conditional dependence network. The network not only links predictors within subgroups but also across different subgroups and is assumed to be unknown. The proposed model incorporates both, selection of important predictors and estimation of a graphical model describing their interdependence. Incorporating network information into the model building process can potentially increase power to detect joint effects on the survival outcome and uncover relationships among the predictors.

This chapter begins with an introduction to Bayesian inference, followed by a description of the Bayesian Cox model, Bayesian variable selection and Bayesian inference in graphical models. These methods provide the basis for the proposed Bayesian subgroup model that is explained in detail in section 3.3.5.

3.3.1 Introduction to Bayesian Inference

Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_U)'$ be the random vector of unknown and unobservable parameters of interest. The aim of Bayesian inference is to make probability statements about the parameter $\boldsymbol{\theta}$ conditional on the observed data $\mathfrak{D} = (\mathfrak{D}_1, ..., \mathfrak{D}_n)$, referred to as posterior distribution $p(\boldsymbol{\theta}|\mathfrak{D})$. $\mathfrak{D}$ will be further specified in the following chapters since it varies depending on the model. In this thesis the general assumption of exchangeability of $\mathfrak{D}_1, ..., \mathfrak{D}_n$ is made. This means that the joint distribution $p(\mathfrak{D}_1, ..., \mathfrak{D}_n)$ remains unchanged by permutations of the indices. The central quantity in Bayesian inference is the posterior distribution $p(\boldsymbol{\theta}|\mathfrak{D})$ that is obtained by Bayes' rule

$$p(\boldsymbol{\theta}|\mathfrak{D}) = \frac{p(\boldsymbol{\theta}, \mathfrak{D})}{p(\mathfrak{D})} = \frac{p(\boldsymbol{\theta})p(\mathfrak{D}|\boldsymbol{\theta})}{p(\mathfrak{D})}.$$

$p(\boldsymbol{\theta}, \mathfrak{D}) = p(\boldsymbol{\theta})p(\mathfrak{D}|\boldsymbol{\theta})$ is the joint distribution of $\boldsymbol{\theta}$ and $\mathfrak{D}$, $p(\boldsymbol{\theta})$ is the prior distribution of the parameter, $p(\mathfrak{D}|\boldsymbol{\theta})$ the likelihood and $p(\mathfrak{D}) = \int p(\boldsymbol{\theta})p(\mathfrak{D}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ the marginal likelihood. Since $p(\mathfrak{D})$ does not depend on $\boldsymbol{\theta}$ and can be considered constant with respect to $\boldsymbol{\theta}$ for fixed observations, it can be omitted in Bayes' rule and yields the unnormalized posterior distribution

$$p(\boldsymbol{\theta}|\mathfrak{D}) \propto p(\boldsymbol{\theta})p(\mathfrak{D}|\boldsymbol{\theta}).$$

The posterior distribution comprises the updated information about $\boldsymbol{\theta}$ and depends on the observed data and prior knowledge about $\boldsymbol{\theta}$ (Gelman, 2004, chapters 1.2 and 1.3). The weight that is given to the prior determines its relative influence. An informative prior has a relatively large influence on the posterior. When no prior knowledge about the parameter exists, a noninformative prior, such as a flat prior (e.g. uniform distribution) or Jeffreys' prior can be used. A more detailed discussion on noninformative priors is given in Gelman (2004, chapter 2.9), Congdon (2006, chapter

1.2), Bernardo and Smith (1994, chapter 5.6.2) and in Kass and Wasserman (1996) based on Jeffreys' rule.

Reasons for the choice of a prior can also be computational convenience and interpretability, both advantages shared by conjugate prior distributions. Conjugacy implies that the posterior distribution has the same parametric form as the prior distribution (Gelman, 2004, chapter 2.4). In practice, particularly in hierarchical models, it is not always possible to derive a conjugate model with a closed-form solution of the posterior. A hierarchical model assumes a joint probability model for the parameters θ_u, $u = 1, ..., U$, that reflects the dependency among them. Further parameters $\boldsymbol{\phi}$ (hyperparameters) are assigned to the parameters. They can be estimated from historical data but this approach is not fully Bayesian since point estimates are used rather than a probability model with a joint posterior distribution. An advantage of hierarchical models is that they have enough parameters to fit complex data well without leading to overfitting.

If no information is available to distinguish any of the θ_u's from each other and no ordering or grouping is reasonable, the parameters are assumed to be exchangeable. The parameters $\theta_1, ..., \theta_U$ are exchangeable if their joint distribution is invariant to permutations of the indices. This assumption allows the determination of a joint probability model for all parameters. In general, the exchangeable parameters $\boldsymbol{\theta}$ are modeled as independently and identically distributed given some unknown parameter $\boldsymbol{\phi}$

$$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})\mathrm{d}\boldsymbol{\phi} = \int \prod_{u=1}^{U} p(\theta_u|\boldsymbol{\phi})p(\boldsymbol{\phi})\mathrm{d}\boldsymbol{\phi}.$$

Bayesian inference in a hierarchical model yields the following joint posterior distribution

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathfrak{D}) \propto p(\boldsymbol{\theta}, \boldsymbol{\phi})p(\mathfrak{D}|\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{\theta}, \boldsymbol{\phi})p(\mathfrak{D}|\boldsymbol{\theta}),$$

where $p(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})$ is the joint prior distribution. The last part of the equation holds because $p(\mathfrak{D}|\boldsymbol{\theta}, \boldsymbol{\phi})$ depends just on $\boldsymbol{\theta}$ and $\mathfrak{D}$ is only indirectly influenced by $\boldsymbol{\phi}$ through $\boldsymbol{\theta}$ (Gelman, 2004, chapters 5.1 and 5.2).

3.3.1.1 Markov Chain Monte Carlo

In Bayesian statistics, Markov Chain Monte Carlo (MCMC) is a general computing approach to simulate an unknown target posterior distribution $p(\boldsymbol{\theta}|\mathfrak{D})$ when a direct draw out of it is not possible or computationally inefficient. The samples of $\boldsymbol{\theta}$ are drawn iteratively from an approximate distribution that is improved in each simulation step to finally converge to the target distribution. In each step, the sampled distribution depends only on the draws from the previous step. Hence, the method fulfills the property of a Markov chain: a sequence of random variables $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ...$ for which for any t the distribution of $\boldsymbol{\theta}^{(t)}$ given all previous $\boldsymbol{\theta}$'s depends only on the most recent value $\boldsymbol{\theta}^{(t-1)}$: $p(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}, ..., \boldsymbol{\theta}^{(1)}) = p(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})$. The idea is to start at some point $\boldsymbol{\theta}^{(0)}$ and after having run long enough, the Markov chain converges to a unique stationary distribution, the posterior distribution $p(\boldsymbol{\theta}|\mathfrak{D})$. In the following, two of the most widely used MCMC methods, the Gibbs sampler and the Metropolis-Hastings algorithm, are described (Gelman, 2004, chapter 11.2).

The Gibbs sampler

The Gibbs sampler is the simplest MCMC algorithm and useful for multidimensional problems where direct sampling from the full conditional posterior distribution is possible (e.g. conjugate distributions). In each iteration step t, the algorithm cycles through the U components of the parameter vector $\boldsymbol{\theta}$ and updates each component with a new sample conditional on the data $\mathfrak{D}$ and current value of all other components $\boldsymbol{\theta}_{-u}^{(t-1)}$, where $\boldsymbol{\theta}_{-u}^{(t-1)} = (\theta_1^{(t)}, ..., \theta_{u-1}^{(t)}, \theta_{u+1}^{(t-1)}, ..., \theta_U^{(t-1)})'$ contains all components of $\boldsymbol{\theta}$ at their latest values except for the u-th.

Step 0. Choose a starting point $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_U^{(0)})'$ and set $t = 1$

Step 1. For $u = 1, ..., U$ update the parameter components $\theta_u^{(t)} \sim p(\theta_u | \boldsymbol{\theta}_{-u}^{(t-1)}, \mathfrak{D})$

Step 2. Set $t = t + 1$ and go to Step 1

(Gelman, 2004, chapter 11.3; Chen, Shao, and Ibrahim, 2000, chapter 2.1).

The Metropolis-Hastings algorithm

The Gibbs sampler requires construction and direct sampling from the full conditional posterior distribution. However, the full conditional does not always reduce analytically to a well-known distribution. In this case, another MCMC technique, the Metropolis-Hastings algorithm, can be used. The algorithm works as follows:

Step 0. Choose a starting value $\boldsymbol{\theta}^{(0)}$ and set $t = 1$

Step 1. Sample a proposal $\boldsymbol{\theta}^{(prop)}$ from a proposal distribution $q(\boldsymbol{\theta}^{(prop)} | \boldsymbol{\theta}^{(t-1)})$

Step 2. Calculate the ratio of ratios $r = \frac{p(\boldsymbol{\theta}^{(prop)}|\mathfrak{D})/q(\boldsymbol{\theta}^{(prop)}|\boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)}|\mathfrak{D})/q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{(prop)})}$

Step 3. Accept the proposal $\boldsymbol{\theta}^{(prop)}$ with probability $\min\{r, 1\}$, which is equivalent to

$$\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^{(prop)} & \text{if } \min\{r,1\} > u, \text{ with } u \sim \mathcal{U}[0,1] \\ \boldsymbol{\theta}^{(t-1)} & \text{else} \end{cases}$$

Step 4. Set $t = t + 1$ and go to Step 1.

The Metropolis-Hastings algorithm is a generalization of the Metropolis algorithm in the sense, that the proposal distribution does not have to be symmetric ($q(\theta_a|\theta_b) = q(\theta_b|\theta_a)$, $\forall \theta_a, \theta_b$). To correct for the asymmetry in q, the density ratio $\frac{p(\boldsymbol{\theta}^{(prop)}|\mathfrak{D})}{p(\boldsymbol{\theta}^{(t-1}|\mathfrak{D})}$ in the Metropolis algorithm is replaced by the above ratio of ratios (Gelman, 2004, chapters 11.4 and 11.5; Gilks, Best, and Tan, 1995).

The proposal distribution can be approximated by a normal distribution with posterior mode. The posterior mode can be determined using Newton's method based on a quadratic Taylor series approximation of the log posterior distribution $l(\boldsymbol{\theta}) = \ln p(\boldsymbol{\theta}|\mathfrak{D})$. Let $\dot{l}(\boldsymbol{\theta}^{(t-1)}) = \frac{\partial l(\boldsymbol{\theta}^{(t-1)})}{\partial \boldsymbol{\theta}^{(t-1)}}$ denote the gradient and $\ddot{l}(\boldsymbol{\theta}^{(t-1)}) = \frac{\partial^2 l(\boldsymbol{\theta}^{(t-1)})}{\partial \boldsymbol{\theta}^{(t-1)} \partial \boldsymbol{\theta}^{(t-1)'}}$ the Hessian. The posterior mode at iteration step t is $\hat{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}^{(t-1)} - \ddot{l}(\boldsymbol{\theta}^{(t-1)})^{-1} \dot{l}(\boldsymbol{\theta}^{(t-1)})$. $\hat{\boldsymbol{\theta}}^{(t)}$ corresponds to the mean and $-\ddot{l}(\boldsymbol{\theta}^{(t-1)})^{-1}$ to the covariance matrix of the normal approximation (Gelman, 2004, chapters 12.1 and 12.2).

3.3.1.2 Assessing convergence

Inference is based on the assumption that for large enough t the distribution of the sampled values $\boldsymbol{\theta}^{(t)}$ is close to the target distribution $p(\boldsymbol{\theta}|\mathfrak{D})$. To minimize the influence of the starting distribution, early iterations are discarded (often the first half of a Markov chain), which is called *burn-in* (Gelman, 2004, chapter 11.6). Simple graphical tools, such as trace plots, running mean plots and plots of the autocorrelation function, can be used to assess whether a Markov chain has reached approximate convergence. In a trace plot the simulated values of a parameter are plotted against the iterations. When there is no strong pattern or trend in the plot after the burn-in period, this indicates convergence. Running mean plots plot the running mean for each parameter in a chain, which is the mean of all sampled values up to a given iteration, at different iterations. In the case of convergence, the running mean should stabilize at the posterior mean. The autocorrelation function describes the autocorrelation of a Markov chain with itself at different lags. High values indicate slow convergence, requiring a large number of iterations to be able to traverse the entire sample space.

Instead of running one long Markov chain, multiple independent Markov chains with overdispersed (relative to the posterior distribution) starting points can be simulated to monitor convergence. Gelman and Rubin (1992) propose a diagnostic that compares the variation between (B/n_c) and within (W) the chains. Let m_c be the number of sequences or Markov chains, each of length n_c (after discarding the burn-in iterations) and $\theta_c^{(t)}$ a scalar estimate of θ in chain c at iteration t. The variances between and within chains are defined as

$$\frac{B}{n_c} = \frac{1}{m_c - 1} \sum_{c=1}^{m_c} (\overline{\theta}_{c\cdot} - \overline{\theta}_{\cdot\cdot})^2$$
$$W = \frac{1}{m_c(n_c - 1)} \sum_{c=1}^{m_c} \sum_{t=1}^{n_c} (\theta_c^{(t)} - \overline{\theta}_{c\cdot})^2,$$

where $\overline{\theta}_{c\cdot}$ is the mean of sequence c and $\overline{\theta}_{\cdot\cdot}$ the sample mean of all $m_c \cdot n_c$ simulated values. The latter is an estimate of the target mean. The target variance σ^2 can be estimated by a weighted average of B and W

$$\hat{\sigma}^2 = \frac{n_c - 1}{n_c} W + \frac{1}{n_c} B$$

and is unbiased under stationarity. If the chains have not converged, W will underestimate σ^2 since the individual chains have not had time to range over the entire target distribution, while $\hat{\sigma}^2$ will overestimate σ^2 for overdispersed starting values. A pooled posterior estimate of the variance is given by $\hat{V} = \hat{\sigma}^2 + \frac{B}{m_c \cdot n_c}$. It is compared with the within-chain variance W in the *potential scale reduction factor* $\hat{R} = \sqrt{\frac{\hat{V}}{W}}$. $\hat{R}$ estimates the factor by which the scale of the current distribution for θ might be reduced if simulations were continued for $t \to \infty$. Thus, if $\hat{R}$ is large, then further simulations may improve inference about the target distribution. When $\hat{R}$ is close to one, all chains have converged. To account for sampling variability in the variance estimates, Brooks and Gelman (1998) suggest to add a correction factor based on estimated degrees of freedom $\hat{R}_c = \sqrt{\frac{df+3}{df+1} \cdot \frac{\hat{V}}{W}}$. This corrected version of the potential scale reduction factor,

as well as the other above described methods for the assessment of convergence are implemented in the R package `coda` (version 0.19-1).

3.3.2 The Bayesian Cox proportional hazards model

Assume the data consist of n independent patients and for each patient the survival or censoring time $\tilde{t}_m$, the right censoring indicator δ_m and a p-dimensional covariate vector $\boldsymbol{x}_m = (x_{m1}, ..., x_{mp})'$ are observed, $m = 1, ..., n$. Let $\boldsymbol{x} \in \mathbb{R}^{n \times p}$ denote the matrix of covariates. The Cox proportional hazards model and the partial likelihood are introduced in 3.1.2 (equations (3.1) and (3.2)). Under the Cox model the joint survival probability of n patients given $\boldsymbol{x}$ is

$$P(\tilde{\boldsymbol{T}} > \tilde{\boldsymbol{t}} | \boldsymbol{x}, \boldsymbol{\beta}, H_0) = \exp\Big(- \sum_{m=1}^{n} \exp(\boldsymbol{\beta}' \boldsymbol{x}_m) H_0(\tilde{t}_m) \Big),$$

with $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ the vector of regression parameters and $H_0(t)$ the cumulative baseline hazard function. One of the most popular choices for $H_0(t)$ is a gamma process prior

$$H_0 \sim \mathcal{GP}(a_0 H^*, a_0),$$

where $H^*(t)$ is an increasing function with $H^*(0) = 0$. H^* can be considered as an initial guess of H_0. $a_0 > 0$ describes the weight or "confidence" that is put in $H^*(t)$ (Lee, Chakraborty, and Sun, 2011). Lee, Chakraborty, and Sun (2011) propose a Weibull distribution $H^*(t) = \eta t^{\kappa}$ with fixed hyperparameters η and κ. Estimates of η and κ can be obtained from the data by fitting a parametric Weibull model without covariates to the survival data (Zucknick, Saadati, and Benner, 2015). Lee, Chakraborty, and Sun (2011) and Zucknick, Saadati, and Benner (2015) choose $a_0 = 2$. A sensitivity analysis for $a_0 \in \{0.5, 1, 2, 4, 8\}$ performed by Zucknick, Saadati, and Benner (2015) reveals that the posterior estimates of the baseline hazard are influenced by the choice of a_0. However, the posterior distribution of $\boldsymbol{\beta}$ and consequently the linear predictors $\boldsymbol{\beta}' \boldsymbol{x}_m$ used for prediction remain nearly unchanged.

In practice the presence of ties is very common, leading to the grouped data likelihood described in Ibrahim, Chen, and Sinha (2005, chapter 3.2.2). Therefore, a finite partition of the time axis is constructed with $0 = c_0 < c_1 < ... < c_J$ and $c_J > \tilde{t}_m$ for all $m = 1, ..., n$. The observed time $\tilde{t}_m$ of patient m falls in one of the J disjoint intervals $I_g = (c_{g-1}, c_g], g = 1, ..., J$. Assume the observed data $\mathfrak{D} = \{(\boldsymbol{x}, \mathcal{R}_g, \mathcal{D}_g) : g = 1, ..., J\}$ are grouped within I_g, where $\mathcal{R}_g$ and $\mathcal{D}_g$ are the risk and failure sets corresponding to interval g. Let $h_g = H_0(c_g) - H_0(c_{g-1})$ be the increment in the cumulative baseline hazard in interval I_g, $g = 1, ..., J$. From the gamma process prior of H_0 follows that the h_g's have independent gamma distributions

$$h_g \sim \mathcal{G}(\alpha_{0,g} - \alpha_{0,g-1}, a_0)\,, \quad \text{with} \quad \alpha_{0,g} = a_0 H^*(c_g)\,.$$

The conditional probability that the observed time of patient m falls in interval I_g is given by

$$\begin{aligned} P(\tilde{T}_m \in I_g|\boldsymbol{h}) &= \exp\Big(-\exp(\boldsymbol{\beta}'\boldsymbol{x}_m)H_0(c_{g-1})\Big) - \exp\Big(-\exp(\boldsymbol{\beta}'\boldsymbol{x}_m)H_0(c_g)\Big) \\ &= \exp\Big(-\exp(\boldsymbol{\beta}'\boldsymbol{x}_m)\sum_{j=1}^{g-1} h_j\Big) \cdot \Big[1 - \exp\Big(-h_g \exp(\boldsymbol{\beta}'\boldsymbol{x}_m)\Big)\Big], \end{aligned}$$

with $\boldsymbol{h} = (h_1, ..., h_J)'$. The resulting grouped data likelihood is defined as

$$L(\mathfrak{D}|\boldsymbol{\beta}, \boldsymbol{h}) \propto \prod_{g=1}^{J} \left[\exp\Big(-h_g \sum_{k \in \mathcal{R}_g - \mathcal{D}_g} \exp(\boldsymbol{\beta}'\boldsymbol{x}_k)\Big) \prod_{l \in \mathcal{D}_g} \Big[1 - \exp\Big(-h_g \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)\Big)\Big]\right]$$

(Ibrahim, Chen, and Sinha, 2005, chapter 3.2.2).

3.3.3 Bayesian variable selection

High-dimensional settings require priors that result in sparse models. One option is the use of shrinkage priors such as the Bayesian lasso as analog to the frequentist penalized likelihood approach. The Bayesian lasso prior was first applied to the Bayesian Cox model by Lee, Chakraborty, and Sun (2011). An extension of their approach is proposed by Zucknick, Saadati, and Benner (2015) to allow the mandatory inclusion of clinical covariates and to perform variable selection only for genomic predictors. The frequentist lasso penalty described in 3.1.2.3 is proportional to the (minus) log-density of independent Laplace priors for the regression parameters: $p(\boldsymbol{\beta}|\lambda_P) = \prod_{i=1}^{p} \frac{\lambda_P}{2} \exp(-\lambda_P|\beta_i|)$. An alternative conditional version of the Laplace prior that ensures unimodal posterior distributions is suggested by Park and Casella (2008)

$$p(\boldsymbol{\beta}|\sigma_\beta^2, \lambda_P) = \prod_{i=1}^{p} \frac{\lambda_P}{2\sqrt{\sigma_\beta^2}} \exp\left(-\frac{\lambda_P|\beta_i|}{\sqrt{\sigma_\beta^2}}\right).$$

More details on the full hierarchical model and prior distributions for σ_β^2 and λ_P can be found in Lee, Chakraborty, and Sun (2011) and Zucknick, Saadati, and Benner (2015). Different Bayesian variable selection techniques for the Cox model are compared by Held, Gravestock, and Sabanés Bové (2016). Alternative prior distributions for variable selection use latent indicator variables such as the stochastic search variable selection procedure explained in the following.

Stochastic search variable selection

The problem of variable selection is to find the best subset of predictors from a set of p potential candidate predictors. Considering all possible subsets would result in a comparison of 2^p possible submodels (using AIC or BIC for example). This requires a major computational challenge, especially when p is large. The stochastic search variable selection (SSVS) procedure by George and McCulloch (1993) avoids the problem of calculating the posterior probabilities of all 2^p subsets. Latent variables are introduced to identify promising subsets of covariates as those with higher posterior probability in the Gibbs sampler. For the estimation of the regression coefficients a mixture of two

normal distributions with different variances is assumed

$$\beta_i|\gamma_i \sim (1-\gamma_i)\cdot\mathcal{N}(0,\tau_i^2)+\gamma_i\cdot\mathcal{N}(0,c_i^2\tau_i^2)\,,\quad i=1,...,p\,.$$

This prior allows the β_i's to shrink towards zero. Due to the shape of the two-component mixture distribution, it is called *spike-and-slab prior*, as illustrated in Figure 3.3. An overview of spike-and-slab priors in linear regression is provided by Ishwaran and Rao (2005). The latent variable γ_i indicates the inclusion ($\gamma_i = 1$) or exclusion ($\gamma_i = 0$) of the i-th variable. It specifies the variance of the normal distribution. τ_i (> 0) is set small so that β_i is likely to be close to zero if $\gamma_i = 0$. c_i (> 1) is chosen sufficiently large to inflate the coefficients of selected variables and to make their posterior mean values likely to be non-zero. In general, the variances of the regression coefficients are assumed to be constant: $\tau_i \equiv \tau$ and $c_i \equiv c$ for all $i = 1, ..., p$.

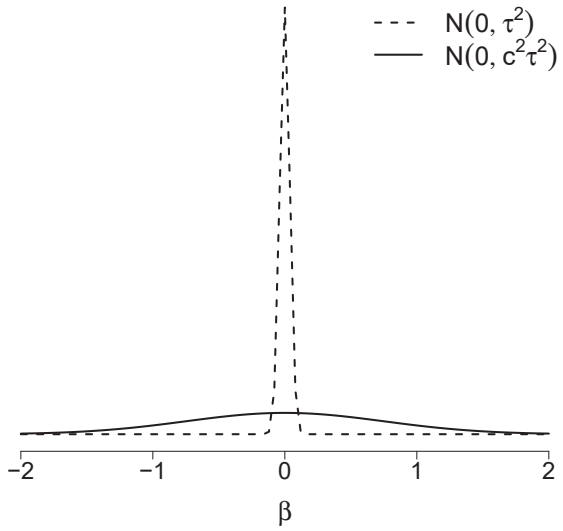


FIGURE 3.3: *Visualization of the spike-and-slab prior for $\boldsymbol{\beta}$. The dashed line represents the spike and the solid line the slab, for $\tau > 0$, $c > 1$.*

The standard prior for $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)'$ consists of a product of independent Bernoulli distributions

$$p(\boldsymbol{\gamma}) = \prod_{i=1}^{p} \pi^{\gamma_i}\cdot(1-\pi)^{1-\gamma_i},$$

with prior inclusion probability $\pi = P(\gamma_i = 1)$. Typically, these prior inclusion probabilities are chosen to be the same for all variables and often with π set to a fixed value. A common choice is $\pi = 0.5$ corresponding to the uniform prior $p(\boldsymbol{\gamma}) = 0.5^p$ and to an expected model size of $p/2$ (George and McCulloch, 1993; Chen and Ye, 2015). Alternatively, π can be defined depending on the a priori expected model size p^* with $\pi = p^*/p$ (Treppmann, Ickstadt, and Zucknick, 2017). Eicher, Papageorgiou, and Raftery (2011) compare these two fixed priors for π in combination with different g-priors on the regression coefficients and conclude that the uniform prior has the best predictive performance based on simulation studies. Ley and Steel (2009) compare both priors with random model priors drawn from a Beta distribution with fixed hyperparameters $a_\pi = 1$ and $b_\pi = (p - p^*)/p^*$. The choice $a_\pi = b_\pi = 1$ corresponds to the standard uniform distribution. The Beta distribution is the conjugate prior for the

parameter of the binomial distribution (Bernardo and Smith, 1994, chapter 3.2.2) and thus, a natural choice. Ley and Steel (2009) recommend a random π when no strong prior information on model size is available, since it is more robust to the choice of p^*. Yang et al. (2016) also assign a hierarchical Beta prior to the variable inclusion probability π with fixed hyperparameters a_π and b_π selected by cross-validation. When a Beta prior is used for π, the hyperparameters a_π and b_π of the Beta distribution are usually chosen to be constant rather than random.

However, in some situations one further hierarchical level may be desirable to allow more flexible modeling of complex data and dependencies. One important objective in this thesis is borrowing information across different subgroups of patients. This could be achieved by a hierarchical regression model with multilevel priors (e.g. multiple Beta hyperprior distributions for the prior variable inclusion probabilities) to represent heterogeneity among, or hierarchy of, the subgroups. Such a model may accommodate differences between single subgroups or groups of similar subgroups, while assuming that the underlying data are derived from a common distribution. In our situation, there is no prior information on the subgroups and similarities between them. Furthermore, we are interested in learning relations between genomic covariates. Therefore, a graphical model is more appropriate and illustrative to describe conditional dependencies among the covariates. Incorporating network structure in the variable selection prior encourages the inclusion of related variables in the graph. In the following, this approach is explained in more detail.

3.3.4 Bayesian structure learning in graphical models

A statistical model that is associated with a graph summarizing the dependence structure in the data is called a graphical model. The nodes of a graph represent the random variables of interest and the edges of a graph describe conditional dependencies among the variables. Structure learning implies the estimation of a graph. Recent applications are mainly driven by biological problems that involve the reconstruction of gene regulatory networks and the identification of pathways of functionally related genes from their expression levels. A graph is called *undirected*, when its edges are unordered pairs of nodes instead of ordered pairs with edges pointing from one node to the other (*directed* graph). When data are continuous and the underlying variables are assumed to be multivariate normal, a common choice are Gaussian models (Drton and Maathuis, 2017). In this thesis, only undirected Gaussian graphical models are considered and thus introduced in the following.

3.3.4.1 Inference in Gaussian graphical models

Assume the vector of random variables $\boldsymbol{X}_m = (X_{m1}, ..., X_{mp})'$ for observation m, $m = 1, ..., n$ follows a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The inverse of the covariance matrix is referred to as precision matrix $\boldsymbol{\Omega} = (\omega_{ij})_{i,j=1,...,p} = \boldsymbol{\Sigma}^{-1}$. $\boldsymbol{\Omega}$ is assumed to be symmetric and positive definite. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the data matrix consisting of n independent observations and $\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$ the sample covariance matrix.

In graphical models, a graph $\widetilde{G}$ is used to represent conditional dependence relationships among random variables $\boldsymbol{X}$. Let $\widetilde{G} = (V, E)$ be an undirected graph or Markov random field (MRF), where $V = \{1, ..., p\}$ is a set of nodes (e.g. genes) and $E \subset V \times V$

is a set of edges (e.g. relations between genes) with edge $(i,j) \in E \Leftrightarrow (j,i) \in E$. $\widetilde{G}$ can be indexed by a set of $p(p-1)/2$ binary variables $\boldsymbol{G} = (g_{ij})_{i<j} \in \{0,1\}^{p \times p}$ with $g_{ij} = 1$ or 0 when edge (i,j) belongs to E or not. The symmetric matrix $\boldsymbol{G}$ is termed adjacency matrix representation of the graph. The graph structure implies constraints on the precision matrix $\boldsymbol{\Omega}$ such that $g_{ij} = 0 \Leftrightarrow (i,j) \notin E \Leftrightarrow \omega_{ij} = 0$, meaning that variables i and j are conditionally independent given all remaining variables (Drton and Maathuis, 2017; Wang, 2015).

The log-likelihood in a Gaussian graphical model (up to a constant) is given by $l(\boldsymbol{\Omega}) = \frac{n}{2}(\ln|\boldsymbol{\Omega}| - \mathrm{tr}(\boldsymbol{S\Omega}))$, with $\boldsymbol{\Omega}$ being positive definite. In the frequentist model $\boldsymbol{\Omega}$ is usually estimated by maximizing the log-likelihood. However, in high-dimensional settings when $p > n$ or when a sparse estimation is desired (zeros in $\boldsymbol{\Omega}$ meaning that a pair of variables is conditionally independent), the maximum likelihood approach does not work. One popular solution is the graphical lasso that optimizes the joint Gaussian log-likelihood with lasso penalty: $\min_{\boldsymbol{\Omega}}\{-\ln|\boldsymbol{\Omega}| + \mathrm{tr}(\boldsymbol{S\Omega}) + \lambda\sum_{i,j}|\omega_{ij}|\}$ (Drton and Maathuis, 2017). An extensive review of frequentist methods for structure learning in graphical models is provided by Drton and Maathuis (2017).

Bayesian approaches to Gaussian graphical models most commonly use the G-Wishart distribution as a conjugate prior for the precision matrix. Entries corresponding to missing edges in the underlying graph are constrained to be zero. The G-Wishart prior on the precision matrix is defined as

$$p(\boldsymbol{\Omega}|\boldsymbol{G}) = C(b, \boldsymbol{D}, \boldsymbol{G})^{-1}|\boldsymbol{\Omega}|^{\frac{b-2}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{D\Omega})\}\mathbb{1}_{\{\boldsymbol{\Omega}\in\mathcal{M}^+\}},$$

where $C(b, \boldsymbol{D}, \boldsymbol{G}) = \int_{\mathcal{M}^+}|\boldsymbol{\Omega}|^{\frac{b-2}{2}}\exp\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{D\Omega})\}\mathrm{d}\boldsymbol{\Omega}$ is the normalizing constant, $b > 0$ the shape (or degrees-of-freedom) and $\boldsymbol{D}$ (positive definite) the inverse scale (or location) parameter. $\mathcal{M}^+$ is the cone of $p \times p$ symmetric positive definite matrices with entries $\omega_{ij} = 0$ whenever $(i,j) \notin E$ (Roverato, 2002; Atay-Kayis and Massam, 2005). Common choices for the hyperparameters are $b = 3$, $\boldsymbol{D} = \boldsymbol{I}_p$ (Jones et al., 2005; Dobra, Lenkoski, and Rodriguez, 2011; Wang, 2015). The edge inclusion indicators g_{ij} can be modeled through independent Bernoulli priors

$$p(\boldsymbol{G}) = \prod_{i<j}\left(\pi^{g_{ij}}(1-\pi)^{(1-g_{ij})}\right),$$

with inclusion probability π (Wang, 2015; Dobra, Lenkoski, and Rodriguez, 2011). Wang (2015) suggests the fixed hyperparameter $\pi = \frac{2}{p-1}$. Inference in high-dimensional graphical models is computationally challenging due to the approximation of the normalizing constant of the G-Wishart (Atay-Kayis and Massam, 2005; Wang, 2015).

Thus, Wang (2012) develops a Bayesian version of the graphical lasso as alternative to the G-Wishart distribution. Laplace priors are assigned to the off-diagonal elements of the precision matrix and exponential priors to the diagonal

$$p(\boldsymbol{\Omega}|\lambda) = C^{-1}\prod_{i<j}\left[\frac{\lambda}{2}\exp\{-\lambda|\omega_{ij}|\}\right]\prod_{i=1}^{p}\left[\frac{\lambda}{2}\exp\{-\frac{\lambda}{2}\omega_{ii}\}\right]\mathbb{1}_{\{\boldsymbol{\Omega}\in\mathcal{M}^+\}}.$$

The normalizing constant $C = \int_{\boldsymbol{\Omega}\in\mathcal{M}^+}\prod_{i<j}\left[\frac{\lambda}{2}\exp\{-\lambda|\omega_{ij}|\}\right]\prod_{i=1}^{p}\left[\frac{\lambda}{2}\exp\{-\frac{\lambda}{2}\omega_{ii}\}\right]\mathrm{d}\boldsymbol{\Omega}$ does not depend on λ when $\lambda \geq 0$ is fixed and the same for all ω_{ij}. In this case, the

posterior mode of $\boldsymbol{\Omega}$ coincides with the graphical lasso estimate. Wang (2012) introduces an effective block Gibbs sampler for updating $\boldsymbol{\Omega}$ one column at a time and avoiding the approximation of the normalizing constant. However, $\boldsymbol{G}$ is not used in the formulation and its posterior distribution is not considered. Another drawback of the Bayesian graphical lasso is, that its continuous prior shrinks but does not set the off-diagonal elements of the precision matrix to exact zeros, which is desired for a sparse estimation. This requires discrete and continuous mixture prior distributions such as the popular G-Wishart prior (Banerjee and Ghosal, 2015; Wang, 2012).

Inspired by the computational efficiency of continuous shrinkage priors for regression analysis (in particular the two component normal mixture for variable selection proposed by George and McCulloch, 1993), Wang (2015) introduces a new approach for structure learning with improved scalability. It is based on continuous spike-and-slab priors on the elements of the precision matrix and latent indicators for the graph structure. It induces sparsity and it is efficient due to a block Gibbs sampler and no approximation of the normalizing constant. The hierarchical model is defined as

$$p(\boldsymbol{\Omega}|\boldsymbol{G},\theta) = C(\boldsymbol{G},\nu_0,\nu_1,\lambda)^{-1}\prod_{i<j}\mathcal{N}(\omega_{ij}|0,\nu_{g_{ij}}^2)\prod_i \mathrm{Exp}(\omega_{ii}|\frac{\lambda}{2})\mathbb{1}_{\{\boldsymbol{\Omega}\in\mathcal{M}^+\}}$$

$$p(\boldsymbol{G}|\theta) = C(\theta)^{-1}C(\boldsymbol{G},\nu_0,\nu_1,\lambda)\prod_{i<j}\Big(\pi^{g_{ij}}(1-\pi)^{1-g_{ij}}\Big),$$

where $\theta = \{\nu_0,\nu_1,\lambda,\pi\}$ is the set of all parameters with $\nu_0 > 0$ small, $\nu_1 > 0$ large, $\lambda > 0$ and $\pi \in (0,1)$. A small value for ν_0 ($g_{ij} = 0$) means that ω_{ij} is small enough to bet set to zero. A large value for ν_1 ($g_{ij} = 1$) allows ω_{ij} to be substantially different from zero. The binary latent variables $\boldsymbol{G} = (g_{ij})_{i<j} \in \{0,1\}^{p(p-1)/2}$ serve as edge inclusion indicators. Wang (2015) proposes the following fixed hyperparameters $\pi = \frac{2}{p-1}$, $\nu_0 \geq 0.01$, $\nu_1 \leq 10$ and $\lambda = 1$.

3.3.4.2 Variable selection for graph-structured covariates

When the covariate space is highly structured it may be desirable to incorporate this structural information into the model building process. Several authors have considered this problem in the context of Bayesian variable selection with genomic applications (Li and Zhang, 2010; Stingo and Vannucci, 2011; Stingo et al., 2011; Peterson, Stingo, and Vannucci, 2016). They adopt a Bayesian spike-and-slab approach as in George and McCulloch (1993) for the selection of important covariates. A Markov random field (MRF) prior is used for the latent variables to incorporate information on the relationships among the covariates as described by an undirected graph. An MRF is an undirected graphical model in which the distribution of a set of random variables follows Markov properties and two unconnected covariates are considered conditionally independent given all others. This prior assumes that neighboring covariates in the network are more likely to have a common effect and encourages their joint inclusion. The MRF prior on the latent variable inclusion indicators $\boldsymbol{\gamma}$ is defined as

$$p(\boldsymbol{\gamma}|\boldsymbol{a},\boldsymbol{B}) = \frac{\exp(\boldsymbol{a}'\boldsymbol{\gamma}+\boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma})}{\sum_{\gamma\in\{0,1\}^p}\exp(\boldsymbol{a}'\boldsymbol{\gamma}+\boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma})} \propto \exp(\boldsymbol{a}'\boldsymbol{\gamma}+\boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma}),$$

where $\boldsymbol{a} = a\mathbf{1}_p$ and $\boldsymbol{B} = (b_{ij})_{p\times p}$ is a symmetric matrix with $b_{ij} = 0$ for all $(i,j) \notin E$. When there is no prior information on the strength of connection between each pair of covariates, the elements b_{ij} are usually set to some constant b for the connected nodes and to 0 for the non-connected ones. The parameter a controls the sparsity of the model, while b regulates the smoothness of the distribution of $\boldsymbol{\gamma}$ over the graph. Higher values of b encourage the selection of variables with neighbors already selected into the model. This idea becomes more evident by looking at the conditional probability

$$p(\gamma_i|\gamma_{j,j\neq i}) = \frac{\exp\left(\gamma_i(a + b\sum_{j\neq i}\gamma_j)\right)}{1 + \exp\left(a + b\sum_{j\neq i}\gamma_j\right)}.$$

If a variable does not have any neighbor, its prior distribution reduces to an independent Bernoulli with parameter $\pi = \frac{\exp(a)}{1+\exp(a)}$.

In high-dimensional settings, increasing values of b may lead to a phase transition in which the number of selected variables (model size) rises drastically. Li and Zhang (2010) provide some guidance for choosing the hyperparameters a and b to avoid a phase transition. They apply the MRF prior to variable selection in linear regression and notice that its addition implies a relatively small increase in computational cost. Stingo and Vannucci (2011) use the MRF prior for discriminant analysis and compare it to independent Bernoulli priors for the individual predictors. They show that employing the MRF prior leads to more accurate selection.

3.3.4.3 Graphical models for heterogeneous data

When data are heterogeneous and can be divided into subpopulations with different dependence structures, it is of interest to learn the structure of graphical models for subpopulations. Estimating one joint graphical model would hide the underlying heterogeneity, while estimating separate models for each subpopulation would neglect common structure. To share common structure in a frequentist setting, Guo et al. (2011) sum up the log-likelihoods of all subpopulations s including a lasso penalty on the elements of the precision matrices $\omega_{s,ij}$. The latter are reparameterized such that $\omega_{s,ij} = \theta_{ij}\tilde{\gamma}_{s,ij}$, where θ_{ij} is a common factor for the presence of edge (i,j) in all subpopulations and $\tilde{\gamma}_{s,ij}$ reflects the differences between subgroups. The lasso penalty is split according to these two factors. Sparsity in θ_{ij} results in edges being simultaneously absent from all graphs and further sparsity within each graph is induced by $\tilde{\gamma}_{s,ij}$. However, the optimization problem is not convex, which makes computation slow and might result in convergence to the wrong local maximum. To overcome this drawback, Danaher, Wang, and Witten (2014) propose another extension of the frequentist graphical lasso for the joint estimation of multiple graphical models. They consider a penalty function of the form $P(\boldsymbol{\Omega}) = \lambda_1\sum_s\sum_{i\neq j}|\omega_{s,ij}| + \lambda_2\tilde{P}(\boldsymbol{\Omega})$, where $\tilde{P}$ allows similarity across the precision matrices of all subpopulations. The group lasso penalty $\tilde{P}(\boldsymbol{\Omega}) = \sum_{i\neq j}\sqrt{\sum_s \omega_{s,ij}^2}$ encourages edges being simultaneously absent from all graphs, while the fused lasso penalty $\tilde{P}(\boldsymbol{\Omega}) = \sum_{r<s}\sum_{i,j}|\omega_{r,ij} - \omega_{s,ij}|$ yields pairwise similar edge patterns in different subpopulations. Gao et al. (2016) apply the fused graphical lasso by Danaher, Wang, and Witten (2014) to the framework of a multivariate Gaussian mixture model with unknown subpopulation membership. Saegusa and Shojaie (2016) propose a weighted Laplacian shrinkage penalty that allows some subpopulations

to be more similar to each other than others: $\tilde{P}(\boldsymbol{\Omega}) = \sum_{i \neq j} \sqrt{\sum_{r,s} \mathrm{w}_{rs}(\omega_{r,ij} - \omega_{s,ij})^2}$. The weight w_{rs} represents the degree of similarity between subpopulations r and s, with $\mathrm{w}_{rs} = 0$ if r and s are not connected.

Bayesian approaches for inferring multiple graphical models have also been developed in the last few years. Yajima et al. (2012) consider the case of two known subpopulations, treating one as the baseline and the other as the differential group. They jointly estimate a Gaussian directed acyclic graph in both subgroups as well as the strength of association. The strength of association between two variables in the differential group is expressed as the sum of the corresponding strength in the baseline group and a differential parameter. Mitra, Müller, and Ji (2016) also focus on the case of two subgroups, where one group serves as reference and the other as differential group. They consider Markov random field models (undirected graphs) and assign a uniform prior to the reference graph $\boldsymbol{G}_1$ and a mixture prior to the differential graph $\boldsymbol{G}_2$ with $g_{2,ij} = g_{1,ij}(1 - \tilde{\delta}_{ij}) + (1 - g_{1,ij})\tilde{\delta}_{ij}$. $\tilde{\delta}_{ij} = |g_{2,ij} - g_{1,ij}|$ is defined as latent indicator of a difference between the two graphs at edge (i, j). It has independent Bernoulli priors with inclusion probability π that represents the global similarity between the two networks. Peterson, Stingo, and Vannucci (2015) propose joint inference of multiple undirected networks under the assumption that some networks may be unrelated while others may have a similar, shared graph structure. The idea is to infer a separate graphical model for each subgroup but to allow for sharing information between sample groups if supported by the data. They choose a G-Wishart distribution as prior on the precision matrix and a Markov Random Field (MRF) prior on the graph structure $p(\boldsymbol{g}_{ij}|\nu_{ij}, \boldsymbol{\Theta}) \propto \exp\left(\nu_{ij}\mathbf{1}'_S\boldsymbol{g}_{ij} + \boldsymbol{g}'_{ij}\boldsymbol{\Theta}\boldsymbol{g}_{ij}\right)$, where $\boldsymbol{g}_{ij} = (g_{1,ij}, ..., g_{S,ij})'$ are the edge inclusion indicators for edge (i, j) in all S graphs, ν_{ij} includes prior knowledge on specific relations and affects sparsity of the graphs, and $\boldsymbol{\Theta} = (\theta_{rs})_{r<s}$ is a $(S \times S)$ symmetric matrix representing the pairwise relatedness of graphs for each group. The MRF prior encourages the selection of the same edge in related graphs. A spike-and-slab prior is placed on the parameters for the network similarity $p(\theta_{rs}|\tilde{\gamma}_{rs}) = (1 - \tilde{\gamma}_{rs}) \cdot \delta_0 + \tilde{\gamma}_{rs} \cdot \mathcal{G}(\theta_{rs}|\alpha, \beta)$ with independent Bernoulli priors on the latent indicators $\tilde{\gamma}_{rs}$ and fixed hyperparameters α, β and $\pi = P(\tilde{\gamma}_{rs} = 1)$.

3.3.5 The proposed Bayesian subgroup model

The methods described in the previous sections of chapter 3.3 form the basis of the proposed Bayesian model introduced below. The Bayesian Cox model is combined with the stochastic search variable selection approach that uses a spike-and-slab prior for the regression coefficients with latent indicators to represent variable inclusion. This is appropriate for situations with many covariates of which only a small number is truly associated with survival outcome and a sparse model solution is desirable. When covariates are functionally related to one another within a network, it may be of interest to learn the dependencies among them. The proposed model assumes that network information is not known a priori and allows inference of the network among the covariates. It is suitable for data consisting of multiple known subgroups that share some predictors with a similar effect on the response, while other predictors may have different effects across subgroups. A joint graph is proposed with possible edges between all pairs of covariates within each subgroup and edges between the same covariates in different subgroups. This graph structure allows sharing information between subgroups when supported by data. For each subgroup, a sparse graph

describing the conditional dependencies among the covariates is inferred and network similarities between subgroups are learned. To accomplish this, an MRF prior is used for the variable selection indicators that encourages the inclusion of predictors linked to one another within a network. The aim is to both identify the relevant predictors for each subgroup while allowing to share information between subgroups, when appropriate, and to learn a sparse network among them.

The proposed model relies on an undirected Gaussian graphical model that assumes multivariate normal data. Therefore, the covariates should be at least approximately normal. This assumption is common and appropriate for many types of biological data like gene expression data. The proposed model is applied to genomic covariates only, but can be extended to allow the mandatory inclusion of established prognostic clinical covariates.

3.3.5.1 Likelihood

Let $\boldsymbol{X}_s \in \mathbb{R}^{n_s \times p}$ be the gene expression (covariate) matrix for subgroup s, $s = 1, ..., S$, consisting of n_s independent and identically distributed observations. For observation m in subgroup s the vector of random variables $\boldsymbol{X}_{s,m} = (X_{s,m1}, ..., X_{s,mp})'$ is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and unknown precision matrix $\boldsymbol{\Omega}_s = \boldsymbol{\Sigma}_s^{-1}$, $m = 1, ..., n_s$. The sample sizes n_s are allowed to differ, but the same p genes (covariates) are assumed to be measured across all subgroups.

Both the response $\boldsymbol{Y}_s = (Y_{s,1}, ..., Y_{s,n_s})'$ with $Y_{s,m} = (\tilde{T}_{s,m}, \delta_{s,m})$ as well as the predictors $\boldsymbol{X}_s$ are considered to be random variables. Thus, the likelihood for subgroup s is the joint distribution $p(\boldsymbol{Y}_s, \boldsymbol{X}_s) = p(\boldsymbol{Y}_s|\boldsymbol{X}_s) \cdot p(\boldsymbol{X}_s)$. The conditional distribution $p(\boldsymbol{Y}_s|\boldsymbol{X}_s)$ corresponds to the grouped data likelihood in the Bayesian Cox proportional hazards model (section 3.3.2) for subgroup s

$$
\begin{aligned}
&L(\mathfrak{D}_s|\boldsymbol{\beta}_s, \boldsymbol{h}_s) \\
&\quad \propto \prod_{g=1}^{J_s} \left[\exp\Big(- h_{s,g} \sum_{k \in \mathcal{R}_{s,g} - \mathcal{D}_{s,g}} \exp(\boldsymbol{\beta}_s' \boldsymbol{x}_{s,k}) \Big) \prod_{l \in \mathcal{D}_{s,g}} \Big[1 - \exp\Big(- h_{s,g} \exp(\boldsymbol{\beta}_s' \boldsymbol{x}_{s,l}) \Big) \Big] \right],
\end{aligned}
$$

where $\mathfrak{D}_s = \{(\boldsymbol{x}_s, \mathcal{R}_{s,g}, \mathcal{D}_{s,g}) : g = 1, ..., J_s\}$ are the observed data in subgroup s, with $\mathcal{R}_g$ the risk and $\mathcal{D}_g$ the failure sets corresponding to interval $I_{s,g} = (c_{s,g-1}, c_{s,g}]$, $g = 1, ..., J_s$. The increment in the cumulative baseline hazard for subgroup s in interval $I_{s,g}$ is termed $h_{s,g} = H_0(c_{s,g}) - H_0(c_{s,g-1})$. $\boldsymbol{\beta}_s$ is the p-dimensional vector of regression coefficients for subgroup s (Lee, Chakraborty, and Sun, 2011).

The marginal distribution of $\boldsymbol{X}_s$ is multivariate normal

$$
\begin{aligned}
p(\boldsymbol{X}_s|\boldsymbol{\Omega}_s) &\propto \prod_{m=1}^{n_s} |\boldsymbol{\Omega}_s|^{1/2} \exp\Big(- \frac{1}{2} \boldsymbol{X}_{s,m}' \boldsymbol{\Omega}_s \boldsymbol{X}_{s,m} \Big) \\
&= |\boldsymbol{\Omega}_s|^{n_s/2} \exp\Big(- \frac{1}{2} \underbrace{\sum_{m=1}^{n_s} \boldsymbol{X}_{s,m}' \boldsymbol{\Omega}_s \boldsymbol{X}_{s,m}}_{= \mathrm{tr}(\boldsymbol{S}_s \boldsymbol{\Omega}_s)} \Big),
\end{aligned}
$$

with $\boldsymbol{S}_s = \boldsymbol{X}'_s\boldsymbol{X}_s$. The joint likelihood across all subgroups is the product of the subgroup likelihoods

$$\prod_{s=1}^{S} L(\mathfrak{D}_s|\boldsymbol{\beta}_s, \boldsymbol{h}_s) \cdot p(\boldsymbol{X}_s|\boldsymbol{\Omega}_s).$$

3.3.5.2 Prior specifications

Prior on the parameters $\boldsymbol{h}_s$ and $\boldsymbol{\beta}_s$ of the Cox model

The prior for the increment in the cumulative baseline hazard in subgroup s follows independent gamma distributions

$$h_{s,g} \sim \mathcal{G}(a_0(H^*(c_{s,g}) - H^*(c_{s,g-1})), a_0),$$

with a Weibull distribution $H^*(c_{s,g}) = \eta_s c_{s,g}^{\kappa_s}$, $g = 1, ..., J_s$, $s = 1, ..., S$ (Lee, Chakraborty, and Sun, 2011). More details are provided in section 3.3.2. The hyperparameters a_0, η_s and κ_s are chosen to be fixed and in accordance with Lee, Chakraborty, and Sun (2011) and Zucknick, Saadati, and Benner (2015). The certainty a_0 about the initial guess H^* of H_0 is set to $a_0 = 2$. The hyperparameters η_s and κ_s are estimated from the (training) data by fitting a parametric Weibull model without covariates to the survival data of subgroup s.

In order to perform variable selection, the SSVS approach of George and McCulloch (1993) introduced in section 3.3.3 is used. The prior of the regression coefficients $\beta_{s,i}$ in subgroup s conditional on the latent variable $\gamma_{s,i}$ is defined as a mixture of two normal distributions with small (τ^2) and large ($c^2\tau^2$) variance

$$\beta_{s,i}|\gamma_{s,i} \sim (1 - \gamma_{s,i}) \cdot \mathcal{N}(0, \tau^2) + \gamma_{s,i} \cdot \mathcal{N}(0, c^2\tau^2), \quad i = 1, ..., p.$$

The latent indicator variable $\gamma_{s,i}$ indicates the inclusion ($\gamma_{s,i} = 1$) or exclusion ($\gamma_{s,i} = 0$) of variable i in the model for subgroup s. Equal variances are assumed for all regression coefficients. The hyperparameters are fixed and set to $\tau = 0.0375$ and $c = 20$ following Treppmann, Ickstadt, and Zucknick (2017). This choice corresponds to a standard deviation of $c \cdot \tau = 0.75$ and a 95% probability interval of $[-1.47, 1.47]$ for $p(\beta_{s,i}|\gamma_{s,i} = 1)$.

Prior on $\boldsymbol{\gamma}$ linking variable and graph selection

The standard prior for the variable selection indicators $\gamma_{s,i}$ is an independent Bernoulli as utilized by Treppmann, Ickstadt, and Zucknick (2017). Here, instead of an independent prior a Markov random field (MRF) prior is chosen as introduced in section 3.3.4.2 and proposed for variable selection by, among others, Peterson, Stingo, and Vannucci (2016). The aim is to link the selection of variables to the presence of edges relating them in an undirected graph. This is achieved by an MRF prior that incorporates the network structure among the covariates and encourages the inclusion of connected variables in the network. The MRF prior for $\boldsymbol{\gamma}$ given $\boldsymbol{G}$ is defined as

$$p(\boldsymbol{\gamma}|\boldsymbol{G}) = \frac{\exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}\in\{0,1\}^{pS}} \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})} \propto \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma}),$$

where $\boldsymbol{\gamma} = (\gamma_{1,1}, ..., \gamma_{1,p}, ..., \gamma_{S,1}, ..., \gamma_{S,p})'$ is a pS-dimensional vector of variable inclusion indicators, $\boldsymbol{G}$ is a symmetric $(pS \times pS)$ adjacency matrix representation of the graph, and a, b are scalar hyperparameters. The hyperparameter a influences the overall variable inclusion probability and controls the sparsity of the model, with smaller values resulting in sparser models. Without loss of generality, $a < 0$ is assumed. The hyperparameter $b > 0$ determines the prior belief on the strength of relatedness between pairs of neighboring variables in the graph and controls the probability of their joint inclusion. Higher values of b encourage the selection of variables with neighbors already selected into the model. This supports the assumption that neighboring covariates in the network are more likely to have a common effect. A sensitivity analysis for the choice of a and b is provided in section 4.2.2.

The elements $g_{rs,ij}$ in the adjacency matrix of the graph $\boldsymbol{G}$ represent the presence $(g_{rs,ij} = 1)$ or absence $(g_{rs,ij} = 0)$ of an edge between nodes (genes) i and j in subgroups r and s. They can be viewed as latent binary indicator variables for edge inclusion. The adjacency matrix in the present model is defined as

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} & \dots & \boldsymbol{G}_{1S} \\ \boldsymbol{G}_{12} & \boldsymbol{G}_{22} & \dots & \boldsymbol{G}_{2S} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{G}_{1S} & \boldsymbol{G}_{2S} & \dots & \boldsymbol{G}_{SS} \end{pmatrix},$$

where $\boldsymbol{G}_{ss} = (g_{ss,ij})_{i<j}$ is the matrix of latent edge inclusion indicators within subgroup s and $\boldsymbol{G}_{rs} = (g_{rs,ii})_{r<s}$ is the matrix of latent edge inclusion indicators between subgroups r and s, $r, s = 1, ..., S$, $r < s$, $i, j = 1, ..., p$, $i < j$, with

$$\boldsymbol{G}_{ss} = \begin{pmatrix} 0 & g_{ss,12} & \dots & g_{ss,1(p-1)} & g_{ss,1p} \\ g_{ss,12} & 0 & \ddots & & g_{ss,2p} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ g_{ss,1(p-1)} & & \ddots & 0 & g_{ss,(p-1)p} \\ g_{ss,1p} & g_{ss,2p} & \dots & g_{ss,(p-1)p} & 0 \end{pmatrix}, \quad \boldsymbol{G}_{rs} = \text{diag}(g_{rs,11}, ..., g_{rs,pp}).$$

Thus, within each subgroup s a standard undirected graph with possible edges between all pairs of genes is assumed, whereas between different subgroups only relations between the same gene in different subgroups are allowed (different genes in different subgroups are assumed to be non-connected). To visualize this idea, Figure 3.4 shows an example network consisting of two subgroups, each with five predictors.

Graph selection prior on $\boldsymbol{\Omega}$ and $\boldsymbol{G}$

The present model does not assume that network information is available a priori and allows inference of the unknown network structure among predictors by using a Gaussian graphical model. Wang (2015) proposes a Bayesian approach for structure learning with improved scalability to larger dimensional problems. It provides a sparse and interpretable representation of the conditional dependencies found in the data. This approach is introduced in section 3.3.4.1 and used in the present model for inferring the precision matrix and network within subgroups. It is based on continuous spike-and-slab

priors on the elements of the precision matrix and latent binary indicators for edge inclusion to identify the graph structure.

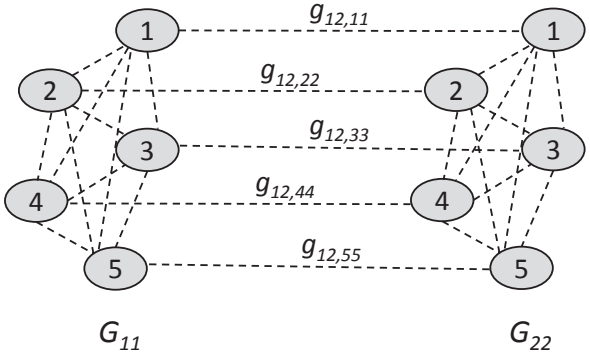


FIGURE 3.4: *Illustration of the proposed graph for $S = 2$ subgroups, each with $p = 5$ genomic predictors (nodes). Possible edges between two nodes are marked by dashed lines.*

The precision matrix of subgroup s corresponding to graph G_{ss} is given by

$$\boldsymbol{\Omega}_s = \begin{pmatrix} \omega_{s,11} & \omega_{s,12} & \dots & \omega_{s,1(p-1)} & \omega_{s,1p} \\ \omega_{s,12} & \omega_{s,22} & \dots & \omega_{s,2(p-1)} & \omega_{s,2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \omega_{s,1(p-1)} & \omega_{s,2(p-1)} & \dots & \omega_{s,(p-1)(p-1)} & \omega_{s,(p-1)p} \\ \omega_{s,1p} & \omega_{s,2p} & \dots & \omega_{s,(p-1)p} & \omega_{s,pp} \end{pmatrix}.$$

The prior on $\boldsymbol{\Omega}_s$ consists of an exponential prior on the diagonal entries and a mixture of two normal distributions with small (ν_0^2) and large (ν_1^2) variance on the off-diagonal entries

$$\begin{aligned} p(\boldsymbol{\Omega}_s|\boldsymbol{G}_{ss}, \nu_0, \nu_1, \lambda) &\propto \prod_{i<j} \mathcal{N}(\omega_{s,ij}|0, \nu^2_{g_{ss,ij}}) \prod_i \text{Exp}(\omega_{s,ii}|\frac{\lambda}{2}) \mathbb{1}_{\{\Omega_s \in \mathcal{M}^+\}} \\ &\propto \prod_{i<j} \exp\{-\frac{1}{2}\frac{\omega^2_{s,ij}}{\nu^2_{g_{ss,ij}}}\} \prod_i \exp\{-\frac{\lambda}{2}\omega_{s,ii}\} \mathbb{1}_{\{\Omega_s \in \mathcal{M}^+\}}, \end{aligned}$$

with fixed hyperparameters $\nu_0 > 0$ small, $\nu_1 > 0$ large and $\lambda > 0$. $\mathbb{1}_{\{\Omega_s \in \mathcal{M}^+\}}$ restricts the prior to the space of symmetric-positive definite matrices. The edge inclusion indicators $g_{ss,ij}$ indicate the presence ($g_{ss,ij} = 1$) or absence ($g_{ss,ij} = 0$) of edge (i, j) in subgroup s. For selected edges, a large variance ν_1^2 allows larger values for $\omega_{s,ij}$, while for non-selected edges, a small variance ν_0^2 ensures that $\omega_{s,ij}$ is likely to be close to zero.

The binary edge inclusion indicators within subgroup s ($g_{ss,ij}$) as well as between subgroups r and s ($g_{rs,ii}$) are assumed independent Bernoulli a priori

$$p(\boldsymbol{G}|\pi) \propto \prod_s \prod_{i<j} \left[\pi^{g_{ss,ij}}(1-\pi)^{1-g_{ss,ij}}\right] \cdot \prod_{r<s} \prod_i \left[\pi^{g_{rs,ii}}(1-\pi)^{1-g_{rs,ii}}\right],$$

with fixed prior probability of edge inclusion $\pi \in (0, 1)$.

Choices for the hyperparameters ν_0, ν_1, λ and π are provided by Wang (2015). He suggests $\pi = 2/(p-1)$ and $\lambda = 1$ and reports that the results are relatively insensitive to

the choice of λ. He finds that values $\nu_0 \geq 0.01$ and $\nu_1 \leq 10$ result in good convergence. In section 4.2.2 a sensitivity analysis for the choice of ν_0 and ν_1 is provided.

3.3.5.3 Posterior inference

The joint posterior distribution for the set of all parameters $\theta = \{\boldsymbol{h}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{G}, \boldsymbol{\Omega}\}$ is proportional to the product of the joint likelihood and the prior distributions of the parameters in all subgroups

$$p(\boldsymbol{h}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{G}, \boldsymbol{\Omega} | \mathfrak{D}, \boldsymbol{X}) \propto \prod_{s=1}^{S} \Big[L(\mathfrak{D}_s | \boldsymbol{\beta}_s, \boldsymbol{h}_s) \cdot p(\boldsymbol{X}_s | \boldsymbol{\Omega}_s) \Big] \cdot \prod_{s=1}^{S} \Big[p(\boldsymbol{\Omega}_s | \boldsymbol{G}_{ss}) \cdot p(\boldsymbol{G}) \cdot p(\boldsymbol{\gamma} | \boldsymbol{G}) \cdot \prod_{i=1}^{p} p(\beta_{s,i} | \gamma_{s,i}) \cdot \prod_{g=1}^{J_s} p(h_{s,g} | \boldsymbol{\beta}_s) \Big].$$

Markov Chain Monte Carlo (MCMC) simulations are required to obtain a posterior sample of the parameters. The different parameters are updated iteratively according to their conditional posterior distributions using a Gibbs sampler. A brief outline of the MCMC sampling scheme is given in the following. More details are provided in the Appendix, section A.3.

1. For subgroup $s = 1, ..., S$ update $\boldsymbol{\Omega}_s$ with the block Gibbs sampler proposed by Wang (2015).

2. Update all elements in $\boldsymbol{G}$ iteratively with Gibbs sampler from the conditional distributions $p(g_{rs,ii} = 1 | \boldsymbol{G}_{-rs,ii}, \boldsymbol{\gamma})$ and $p(g_{ss,ij} = 1 | \boldsymbol{G}_{-ss,ij}, \omega_{s,ij}, \boldsymbol{\gamma})$, where $\boldsymbol{G}_{-rs,ii}$ ($\boldsymbol{G}_{-ss,ij}$) denotes all elements in $\boldsymbol{G}$ except for $g_{rs,ii}$ ($g_{ss,ij}$).

3. Update all elements in $\boldsymbol{\gamma}$ iteratively with Gibbs sampler from the conditional distributions $p(\gamma_{s,i} = 1 | \boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}, \beta_{s,i})$, where $\boldsymbol{\gamma}_{-s,i}$ denotes all elements in $\boldsymbol{\gamma}$ except for $\gamma_{s,i}$.

4. Update $\beta_{s,i}$ from the conditional distribution $p(\beta_{s,i} | \boldsymbol{\beta}_{s,-i}, \boldsymbol{\gamma}_s, \boldsymbol{h}_s, \mathfrak{D}_s)$, $s = 1, ..., S$, $i = 1, ..., p$, using a random walk Metropolis-Hastings algorithm with adaptive jumping rule as proposed by Lee, Chakraborty, and Sun (2011). $\boldsymbol{\beta}_{s,-i}$ includes all elements in $\boldsymbol{\beta}_s$ except for $\beta_{s,i}$.

5. The conditional distribution $p(h_{s,g} | \boldsymbol{h}_{s,-g}, \boldsymbol{\beta}_s, \boldsymbol{\gamma}_s, \mathfrak{D}_s)$ for the update of $h_{s,g}$ can be well approximated by the gamma distribution

$$h_{s,g} | \boldsymbol{h}_{s,-g}, \boldsymbol{\beta}_s, \boldsymbol{\gamma}_s, \mathfrak{D}_s \overset{\text{approx.}}{\sim} \mathcal{G}\Big(a_0 (H^*(c_{s,g}) - H^*(c_{s,g-1})) + d_{s,g}, a_0 + \sum_{k \in \mathcal{R}_{s,g} - \mathcal{D}_{s,g}} \exp(\boldsymbol{\beta}_s' \boldsymbol{x}_{s,k}) \Big),$$

where $d_{s,g}$ are the number of events in interval g for subgroup s and $\boldsymbol{h}_{s,-g}$ denotes the vector $\boldsymbol{h}_s$ without the g-th element, $g = 1, ..., J_s$, $s = 1, ..., S$ (Ibrahim, Chen, and Sinha, 2005, chapter 3.2.2).

Starting with an arbitrary set of initial values for the parameters, the MCMC algorithm runs with a reasonably large number of iterations to obtain a representative

sample from the posterior distribution. After removing the burn-in samples, the remaining samples are used for computing the posterior estimates. There are different opinions on using a single long-run Markov chain or multiple short-run chains with overdispersed starting values. A single long chain may be closer to the target distribution at the end of the run and provide more accurate inference compared to several shorter chains. Particularly in situations that require longer burn-in periods, multiple shorter chains suffer from a large number of discarded burn-in samples and may be a waste of resources (Geyer, 1992; Chen, Shao, and Ibrahim, 2000). Gelman and Rubin (1992) propose the use of multiple short chains because they may have better exploratory power. A single chain may not explore the entire parameter space when it gets caught at an attractive mode and remains in its neighborhood. In this thesis, both types of MCMC approaches are used to assess convergence and mixture of the chains. In a preliminary analysis in section 4.2.3 several independent MCMC chains are run with different starting values. This serves as additional information to confirm convergence and to ensure that the chains are not too short. In all subsequent analyses, a single MCMC chain is used and the initial values are chosen as follows:

$$\boldsymbol{G}^{(0)} = \mathbf{0}_{pS \times pS}$$

$$\boldsymbol{\Sigma}_s^{(0)} = \boldsymbol{I}_{p \times p} \text{ and } \boldsymbol{\Omega}_s^{(0)} = (\boldsymbol{\Sigma}_s^{(0)})^{-1} \text{ for } s = 1, ..., S$$

$$\boldsymbol{\gamma}_s^{(0)} = (0, ..., 0)' \text{ for } s = 1, ..., S$$

$$\beta_{s,i}^{(0)} \sim \mathcal{U}[-0.02, 0.02] \text{ for } i = 1, ..., p,\ s = 1, ..., S$$

$$h_{s,g}^{(0)} \sim \mathcal{G}(1, 1) \text{ for } s = 1, ..., S,\ g = 1, ..., J_s.$$

Chapter 4

Results

This thesis is motivated by the assumption that the present data are heterogeneous due to known patient subgroups whose covariates differ in their relation to survival outcome. Some subgroups may be closer related to one another while others are not. Thus, sharing information between subgroups in order to increase sample size is reasonable when supported by data. Main objectives of this thesis are the prediction of a patient's survival function based on potentially high-dimensional covariates such as gene expression data, and simultaneously, the inclusion of variable selection and consideration of heterogeneity in modeling. The aim is to provide a separate prediction model for each subgroup that allows the identification of common as well as subgroup-specific effects and has improved prediction accuracy over standard approaches. To accomplish this, a classical frequentist and a Bayesian Cox regression model are proposed and compared to a standard subgroup model and to a standard combined model. The former is based only on patients of the subgroup of interest, while the latter incorporates patients from all subgroups and thus, benefits from the increased sample size. However, pooling data without taking into account heterogeneity may result in biased estimates, and subgroup-specific effects may remain undetected. All models are applied to both simulated and real data with different lung cancer studies as subgroups, including survival endpoint and clinical as well as gene expression data as covariates. Information on these lung cancer studies is provided in section 2.2.

The statistical software R is used for all computations (version 3.4.1 for sections 4.1.2 and 4.1.2.2, and version 3.4.3 for all remaining analyses), and the R package `batchtools` (version 0.9.6 for sections 4.1.2 and 4.1.2.2, and version 0.9.8 for all remaining analyses) for parallelization.

Section 4.1 presents the results of the frequentist models. In subsection 4.1.2 the proposed model is evaluated through simulations and compared to standard Cox models and a weighted Cox model with fixed weights. In subsection 4.1.2.1 different types of Cox models with lasso penalty are compared to componentwise likelihood-based boosting for Cox models with regard to prediction performance based on simulated data. The case of unbalanced subgroup sizes is considered in one simulation study in subsection 4.1.2.2, including the comparison of oversampling techniques in classification to potentially improve performance of weights estimation in the proposed weighted Cox model. Finally, the frequentist Cox models are applied to real lung cancer studies (subsection 4.1.3). Results of the Bayesian approaches are reported in section 4.2, beginning with two preliminary analyses based on simulated data for the choice of hyperparameters and assessment of convergence of the proposed model. This is followed by an evaluation of the Bayesian Cox models in simulation studies (subsection 4.2.4) and in application to real lung cancer studies (subsection 4.2.5).

4.1 Frequentist subgroup model

The proposed frequentist subgroup model uses patients from all subgroups for training but assigns them individual weights in the likelihood based on their subgroup affiliation. Weights for a specific subgroup s are defined as $\frac{p(s|\boldsymbol{z})}{p(s)}$, with $\boldsymbol{z}$ being the observed set of covariates, survival time and status (event indicator). $p(s)$ is estimated by the relative frequency of subgroup s and $p(s|\boldsymbol{z})$ by classification, such that a patient who is likely to belong to the subgroup of interest receives a higher weight in the subgroup-specific likelihood. More details are provided in section 3.2. Different classification methods that are appropriate for multi-class problems and high-dimensional covariates are compared with respect to their predictive quality. In this context, sparsity and interpretability of the classification model are unimportant. The focus is on prediction performance only, and the ability to discriminate between differing subgroups. In the following simulation studies similarities between subgroups are known and performance of weights estimation is assessed by the accuracy (ACC) and the area under the ROC curve (AUC). Multinomial logistic regression with lasso and ridge penalty, and random forest are considered as classification methods. To potentially improve prediction performance, two further parameters for weights estimation are compared. First, the inclusion of pairwise interactions between each (genomic) covariate and the survival time, which corresponds to the assumption that covariates of different subgroups may be related to different prognosis. Second, replacing the survival time with the Nelson–Aalen estimator of the cumulative hazard rate (HR), which has been recommended by White and Royston (2009) to improve multiple imputation of missing data. In summary, the following three parameters resulting in 12 different combinations for the estimation of subgroup weights are studied:

- Method: multinomial logistic regression with lasso (*lasso*) or ridge (*ridge*) penalty, or random forest (*rf*)
- Interactions: including (*intera.*) or excluding (*no intera.*) interactions between covariates and survival time
- Cumulative HR: replacing the survival time with the cumulative HR (*cumHR*) or not (*no cumHR*).

The proposed weighted approach is compared to the standard combined and subgroup model, as well as a weighted Cox regression model with different fixed weights as proposed by Weyer and Binder (2015). Observations belonging to a certain subgroup are assigned a weight of 1 in the subgroup-specific likelihood, while all other observations are down-weighted with a constant weight $w \in (0, 1)$. Since sparsity and interpretability of the resulting Cox models are important besides good prediction performance, all Cox models use a lasso penalty for variable selection. To sum up, the following types of Cox regression models are compared:

- Weighted model with estimated weights (different parameters for weights estimation)
- Weighted model with fixed weights $w = 0.1, 0.2, ..., 0.9$
- Standard subgroup model (*sub*), using only patients of a specific subgroup

- Standard combined model (*all*), using patients of all subgroups. The subgroup indicator is included as additional covariate.

Figure 4.1 provides a schematic representation of the analysis pipeline. First, the whole data are randomly split into a training (with proportion 0.632) and a test data set. Subsampling is done stratified by subgroup and event indicator, to take different subgroup sizes and censoring proportions into account. This procedure is repeated 100 times. Numeric covariates, in particular gene expression variables, are standardized before model fitting and evaluation to have zero mean and unit variance. Parameters of the training data set (mean and standard deviation of each variable) are used to scale the training and test data set. For the standard subgroup model each subgroup is standardized separately, whereas for the weighted model and combined model training data of all subgroups are pooled. Individual subgroup weights are estimated from the training data with 10-fold cross-validation (CV). Next, the combined and weighted Cox models are fitted based on the training data of all subgroups, while the standard subgroup model is based on the training data of the respective subgroup only. Finally, the prediction performance of the estimated Cox models with respect to a certain subgroup is evaluated using only the test data of this particular subgroup. The R package `mlr` (version 2.11 for section 4.1.2, and version 2.12 for all remaining analyses) is used as a framework for weights estimation, Cox model fitting and evaluation by C-index.

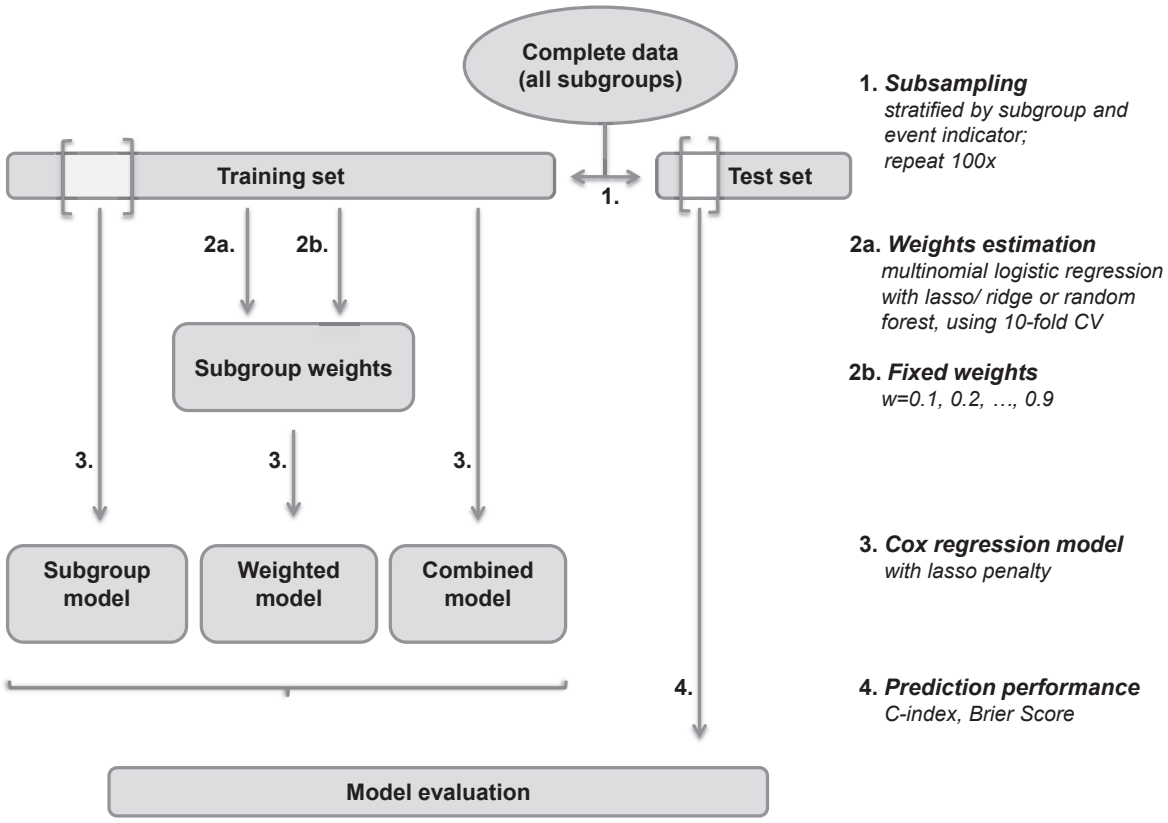


FIGURE 4.1: *Analysis pipeline for the frequentist setting.*

4.1.1 Simulation setup

In the following simulation studies, four subgroups (1A, 1B, 2A, 2B) are considered that belong to two differently distributed groups: group 1 including subgroups 1A and 1B, and group 2 including subgroups 2A and 2B. Within each group, the same parameters are chosen for data simulation. In the following, the index $g^* = 1, 2$ denotes the group. Survival data are simulated from a Weibull distribution according to Bender, Augustin, and Blettner (2005), with scale η_{g^*} and shape κ_{g^*} parameters estimated from two independent lung cancer data sets (GSE37745 and GSE50081). Therefore, survival probabilities at 3 and 5 years are computed using the Kaplan-Meier estimator for both lung cohorts separately. The corresponding probabilities are 57% and 75% for 3-years survival, and 42% and 62% for 5-years survival, respectively. Individual event times for group g^* are simulated as

$$T_{g^*} \sim \left(-\frac{\log(U)}{\eta_{g^*} \exp(\boldsymbol{x}_{g^*} \boldsymbol{\beta}_{g^*})} \right)^{1/\kappa_{g^*}}, \quad U \sim \mathcal{U}[0, 1],$$

with true effects $\boldsymbol{\beta}_{g^*} \in \mathbb{R}^p$, $g^* = 1, 2$. Noninformative censoring times C_{g^*} are randomly drawn from a Weibull distribution with the same parameters as for the event times, resulting in approximately 50% censoring rates in both groups. The individual observed event indicators and times until an event or censoring are defined as $\delta_{g^*} = \mathbb{1}(T_{g^*} \leq C_{g^*})$ and $\widetilde{T}_{g^*} = \min(T_{g^*}, C_{g^*})$.

The following simulation studies only consider covariates that represent characteristics of gene expression data, particularly the relatively small sample size with respect to the number of variables. Gene expression data $\boldsymbol{x}_{g^*} \in \mathbb{R}^{n \times p}$ are simulated from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_{g^*}$ and covariance matrix $\boldsymbol{\Sigma}$. The same p genes are assumed to be measured in all subgroups with an equal number of observations n. In section 4.1.2.2 the case of differing sample sizes across subgroups is studied. In all simulated scenarios the first 12 genes are assumed to be prognostic in at least one of the two groups and subsequently termed prognostic genes. Their true effects on the survival outcome are

	Gene											
	1	2	3	4	5	6	7	8	9	10	11	12
$\boldsymbol{\beta}_1$	1	1	0	0	-0.5	0.5	0.75	0.25	-1	-1	-0.75	-0.25
$\boldsymbol{\beta}_2$	0	0	1	1	0.5	-0.5	0.25	0.75	-1	-1	-0.75	-0.25

including subgroup-specific effects (genes 1 to 4), opposite effects (genes 5 and 6), effects in the same direction but of different size (genes 7 and 8), and joint effects of varying sizes (genes 9 to 12). These effects are chosen with alternate signs so that they sum up to zero, resulting in reasonable simulated survival times. In settings where $p > 12$, all remaining genes are assumed to be noise and unrelated to survival outcome in both groups ($\beta_{13} = ... = \beta_p = 0$). The amount of added noise is varied to test the ability of the proposed model to identify important covariates in the presence of noise. Elements of the mean vector $\boldsymbol{\mu}_{g^*}$ are defined by a linear function with parameter $\epsilon \in [0, 1]$ that reflects the degree of similarity between the two groups. $\mu = 4 + 4 \cdot \epsilon$ is assigned to genes with a strong effect on the response ($|\beta| = 1$), $\mu = 4 + 2 \cdot \epsilon$ corresponds to genes with a moderate effect ($|\beta| = 0.5, 0.75$), and $\mu = 4$ to genes with a weak or no effect

($|\beta| = 0, 0.25$). This choice relies on the assumption that prognostic genes have a higher expression level than noise genes. The magnitude of μ is chosen in accordance with real Microarray gene expression data, where gene expression values typically range from 4 to 12 after transformation to $\log_2$ scale.

For one specific simulation setting survival and gene expression data of all subgroups are illustrated in Figure 4.2 by means of a Kaplan-Meier plot of estimated survival functions, and PCA (Principal Component Analysis) plots based on gene expression data with $\epsilon = 0$ or $\epsilon = 1$ and survival times. When $\epsilon = 0$, gene expression data of all subgroups are simulated from the same distribution and groups 1 and 2 differ only in their simulated survival times. When $\epsilon = 1$, both groups also differ in gene expression. Specifically, different mean values are used to simulate expression data of prognostic genes. The PCA plots indicate that all subgroups are inseparable in the directions of the first two principal components (PCs) when $\epsilon = 0$, in contrast to $\epsilon = 1$, where subgroups 1A and 1B cluster together, as do subgroups 2A and 2B. According to the second PC, all subgroups cannot be distinguished from each other. But in the direction of the first PC, group 1 and 2 are clearly distinguishable. This effect becomes more evident when the sample size is large compared to the number of genes. However, the proportion of explained variance in both PCA plots is very small.

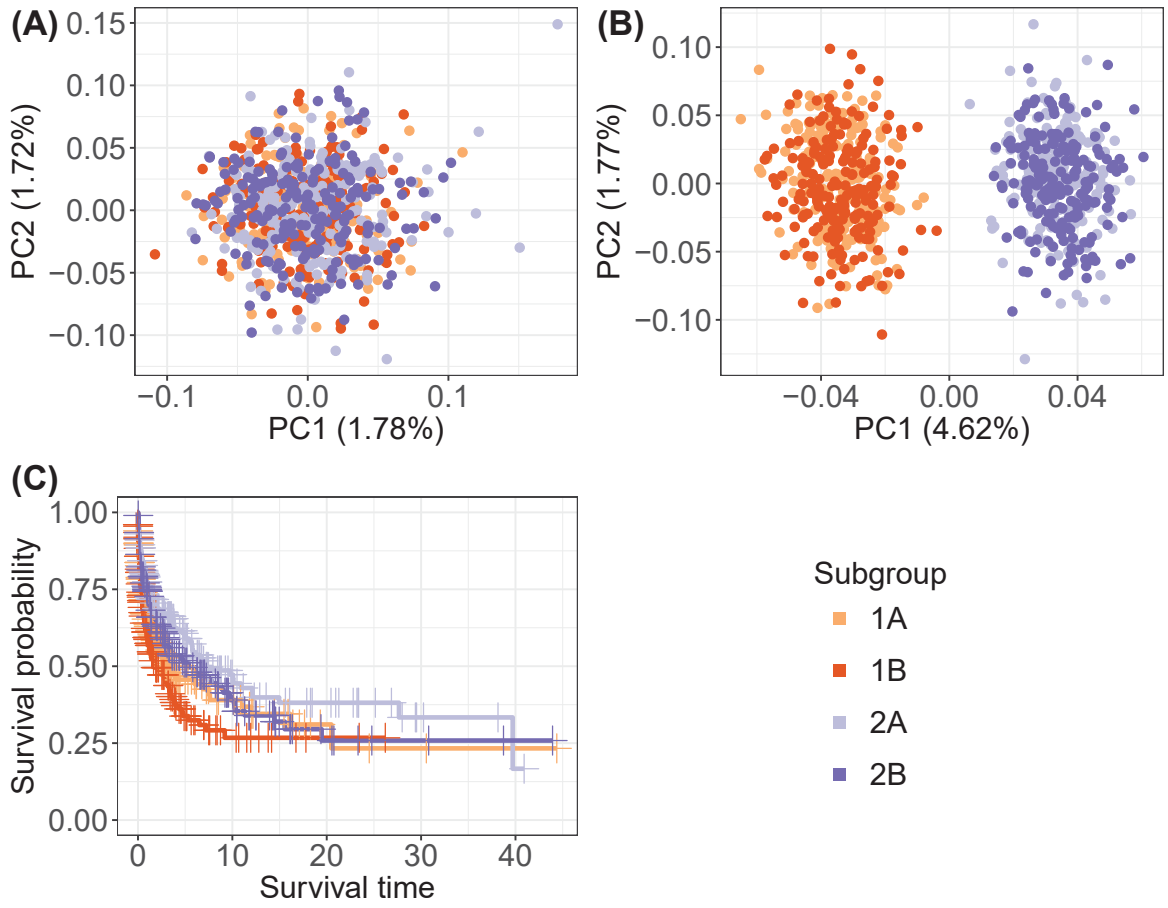


FIGURE 4.2: *Descriptive analysis of simulated subgroup data. PCA plot based on uncorrelated gene expression data and survival times with $n = 200$, $p = 100$, and (A) $\epsilon = 0$, (B) $\epsilon = 1$. (C) Kaplan-Meier plot of estimated survival functions for all subgroups.*

All subgroups share the same covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$. Five different types of covariance matrices are considered in the following. Assuming that all genes within each subgroup are independently $\mathcal{N}(0,1)$ distributed, corresponds to $\boldsymbol{\Sigma} = \boldsymbol{I}_{p \times p}$ (termed *uncorrelated*). Alternatively, a blockwise autoregressive correlation structure is generated (termed *block*), where each of the 12 prognostic genes is assigned to a block of size $q = \left\lfloor \frac{p-12}{12} \right\rfloor$ of non-prognostic genes. If $p > 12q + 12$, the remaining non-prognostic genes are uncorrelated. Prognostic genes are uncorrelated with each other and have correlation 0.5^i, $i = 1, ..., q$ with their respective non-prognostic genes. Within each block of non-prognostic genes, genes i and j are pairwise correlated with $0.5^{|i-j|}$, $i, j = 1, ..., q$. For the purpose of illustration, assume there are $p = 12$ genes whereof the first 3 genes are prognostic and the last 9 genes are non-prognostic ($q = \frac{12-3}{3} = 3$). Then the correlation structure would be as follows

$$\left(\begin{array}{ccc:ccc:ccc:ccc}
1 & 0 & 0 & 0.5 & 0.5^2 & 0.5^3 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0.5 & 0.5^2 & 0.5^3 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.5^2 & 0.5^3 \\
\hdashline
0.5 & 0 & 0 & 1 & 0.5 & 0.5^2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.5^2 & 0 & 0 & 0.5 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.5^3 & 0 & 0 & 0.5^2 & 0.5 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hdashline
0 & 0.5 & 0 & 0 & 0 & 0 & 1 & 0.5 & 0.5^2 & 0 & 0 & 0 \\
0 & 0.5^2 & 0 & 0 & 0 & 0 & 0.5 & 1 & 0.5 & 0 & 0 & 0 \\
0 & 0.5^3 & 0 & 0 & 0 & 0 & 0.5^2 & 0.5 & 1 & 0 & 0 & 0 \\
\hdashline
0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.5 & 0.5^2 \\
0 & 0 & 0.5^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 1 & 0.5 \\
0 & 0 & 0.5^3 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5^2 & 0.5 & 1
\end{array}\right).$$

Further, two adaptions of this *block* correlation structure are used, by removing the correlations between each prognostic gene and its corresponding non-prognostic genes (*blockdiag*), and by setting all pairwise correlations smaller than 0.1 to 0.1 (*min01*). In addition to the four artificial correlation structures, the shrinkage correlation estimator by Schäfer and Strimmer (2005) is applied (termed *shrinkage*). The empirical correlation matrix obtained from the combined gene expression data of two lung cancer cohorts (GSE37745 and GSE50081) and 1000 randomly selected genes is shrunken towards the identity matrix. The shrinkage correlation estimate is defined once (see Figure B.1 in the Appendix) and when $p = 100$, only the first 100 row and column entries are used.

4.1.2 Simulation studies

This section reports the results of two simulation studies that examine the effect of varying number of genomic covariates p and sample sizes n, as well as different covariance structures $\boldsymbol{\Sigma}$ and degrees of similarity ϵ between the two distinct groups. The focus is on high-dimensional settings where sample size is small compared to the number of variables, a typical characteristic of gene expression data. Table 4.1 presents the parameter combinations of the first simulation study. For $p = 12$ only uncorrelated covariates are considered ($\boldsymbol{\Sigma} = \boldsymbol{I}_{p \times p}$). This results in 66 parameter combinations for data simulation.

Netzer (2013) estimates the probability $p(s|\boldsymbol{z})$ in the numerator of the weights from the complete training data, which leads to overfitting. This effect is particularly pronounced when using random forest, as illustrated in Figure B.2 in the Appendix.

For multinomial logistic regression with lasso or ridge penalty, the overfitting effect is much weaker but becomes more evident for larger differences between subgroups ($\epsilon = 1$). As a solution to this problem, 10-fold cross-validation is applied to the training data to obtain predictions for $p(s|\boldsymbol{z})$. Figure 4.3 displays boxplots of the estimated weights across all training sets for two selected simulation scenarios with $\epsilon = 0$ and $\epsilon = 1$. The x-axis represents the true subgroup membership of each observation, and each of the four plots shows the estimated weights in each of the four subgroup models (predicted probabilities of belonging to a certain subgroup divided by the subgroup proportion). Results of all methods (random forest, lasso, ridge) and parameters (with/without interactions/cumulative HR) for weights estimation do not differ much. The largest difference is due to ϵ. When $\epsilon = 0$, multi-class classification fails to distinguish the two groups. All observations are assigned a weight of around one in all subgroup models. When $\epsilon = 1$, however, classification succeeds in providing an almost perfect separation between both groups (Figure 4.3).

Parameter	Values (per subgroup)
n	50, 200, 1000
p	12, 100, 1000
$\boldsymbol{\Sigma}$	uncorrelated, block, blockdiag, min01, shrinkage
ϵ	0, 1

TABLE 4.1: *Parameter values in the first simulation study.*

The area under the ROC curve (AUC) and the accuracy (ACC) are used to assess the performance of weights estimation in all simulation settings. A distinction is made between groups 1 and 2 only (not between subgroups A and B). Both performance measures are computed based on cross-validated training data and test data. Results of both measures based on training and test data are very similar. Hence, only AUC results from cross-validated training data are shown. Mean AUC values are depicted in Figure 4.4 for $\epsilon = 0$, and in Figure B.3 in the Appendix for $\epsilon = 1$. When $\epsilon = 0$ the AUC mostly lies between 0.5 and 0.6, indicating that prediction performance is not much better than random and that discriminatory power with regard to both groups is low. The performance is better in low-dimensional settings ($p = 12$) and for increasing sample size in high-dimensional settings ($p = 100, 1000$). Random forest including interactions and cumulative HR is the best method in terms of the highest AUC values. In contrast, lasso and ridge exhibit better performance without cumulative HR. For $\epsilon = 1$ the mean AUC is one, with the exception of ridge for $p = 1000$ and $n = 50$. This suggests an (almost) perfect discrimination between the two groups, independent of the parameter settings and methods.

To further examine the effect of the parameter settings on classification performance, a regression tree is computed with AUC as response and all parameters for data simulation and weights estimation as covariates (see Figure 4.5). The most important splitting variable leading to the largest difference in AUC is ϵ. The performance of ridge is slightly worse than lasso and random forest. When $\epsilon = 0$ and $n < 1000$, random forest outperforms both multinomial logistic regression approaches. Larger sample size n and smaller number of covariates p also result in better prediction performance. The same regression tree is obtained for AUC based on all test sets, and for ACC.

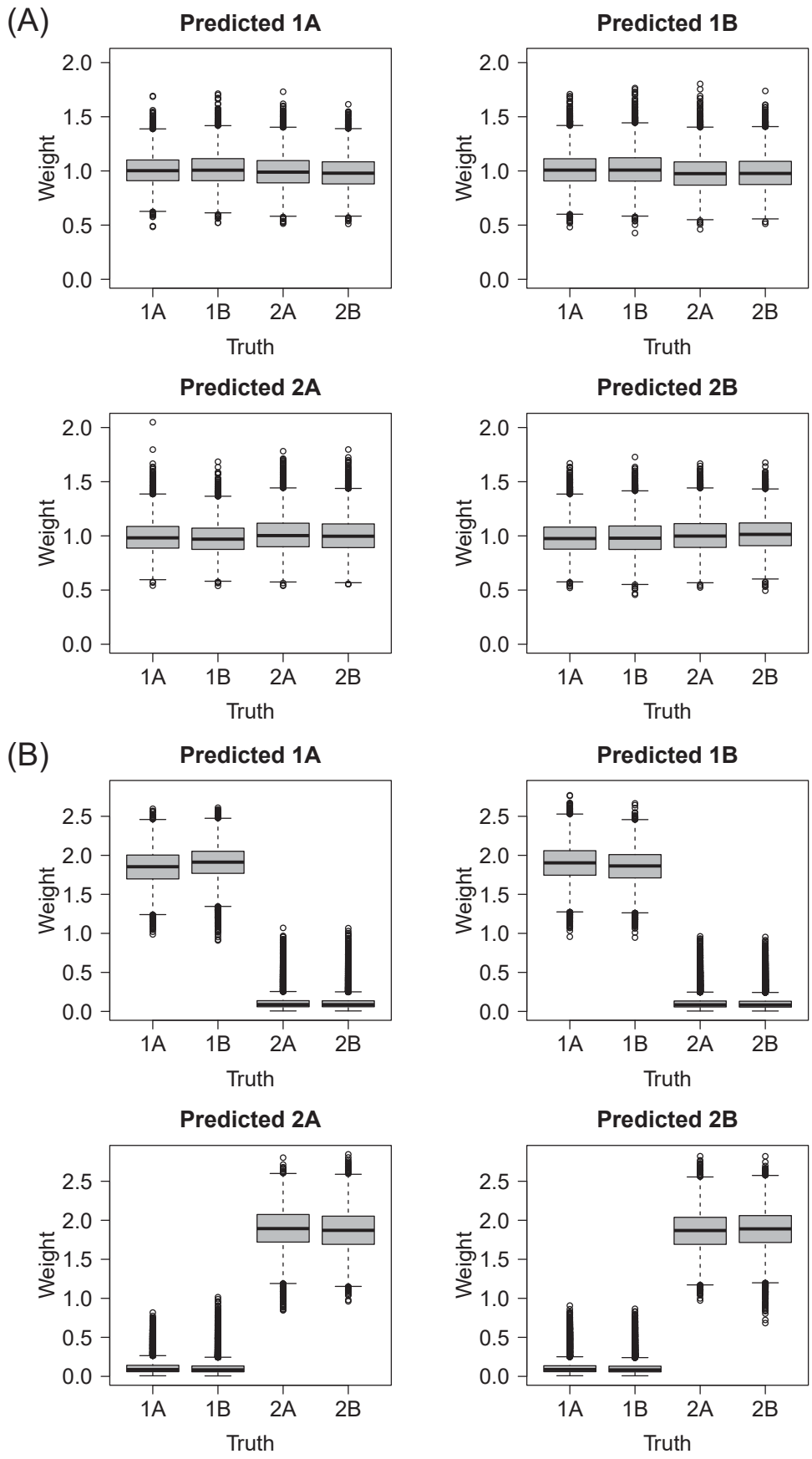


FIGURE 4.3: *Weights estimated with random forest without interactions and cumulative HR, for simulated data with $n = 200$, $p = 100$, block correlation, and (A) $\epsilon = 0$, (B) $\epsilon = 1$.*

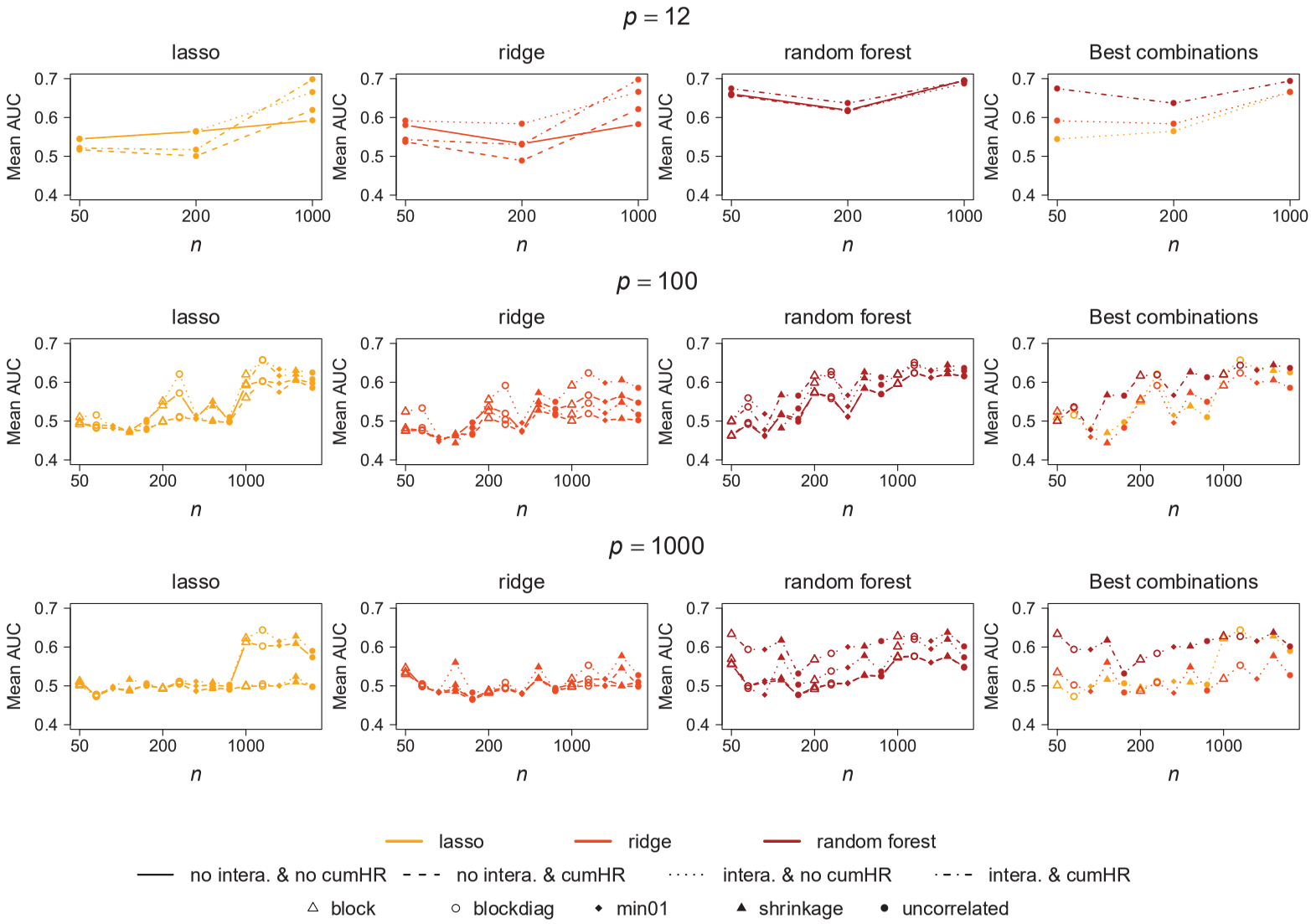


FIGURE 4.4: *Mean AUC for weights estimation based on cross-validated training sets. Comparison of different parameters (line type) and methods (line color) for weights estimation, as well as different parameters for data simulation with $\epsilon = 0$. The first three columns show all parameter combinations for the three classification methods, and the fourth column for each method the best combination (mean across different n and $\boldsymbol{\Sigma}$).*

For each parameter combination for data simulation, the top five combinations of methods and parameter settings for weights estimation in terms of highest mean AUC (averaged over all cross-validated training sets) are summarized in Table C.2 in the Appendix. When $\epsilon = 0$, random forest with interactions and cumulative HR provides almost always the best prediction performance. When $\epsilon = 1$, all combinations perform equally well with a mean AUC of one.

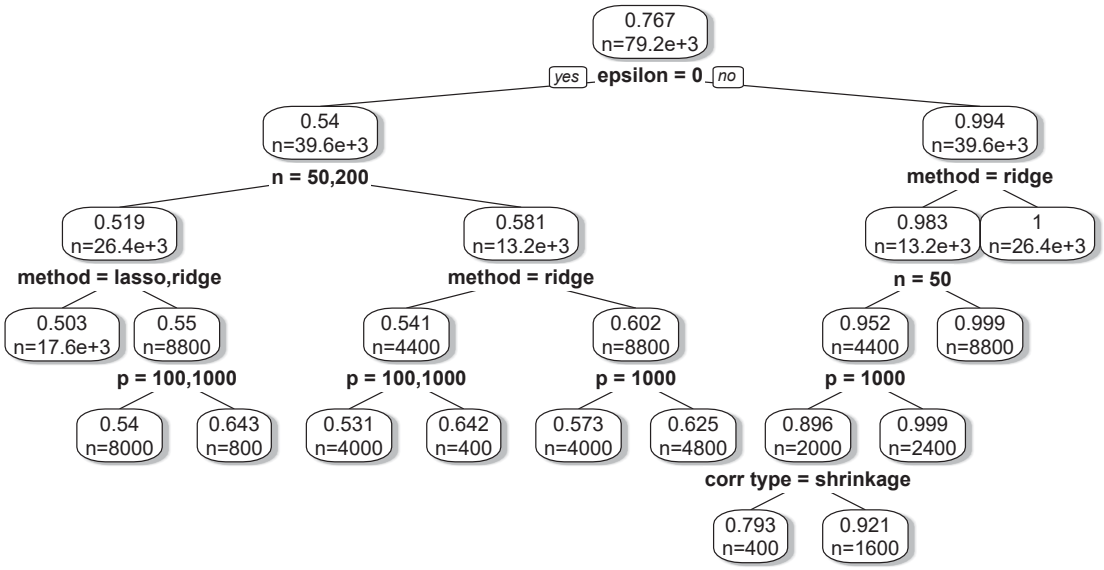


FIGURE 4.5: *Regression tree with AUC based on cross-validated training data as response and all parameters for data simulation and weights estimation as covariates. Each box shows mean AUC and sample size in the corresponding node.*

Next, results of all Cox models are regarded and weighted Cox models, including fixed or estimated weights (with different methods for weights estimation), are compared to the standard combined and subgroup model. To begin with, Cox model fits are evaluated by looking at the estimated regression coefficients and the corresponding mean inclusion frequencies for variable selection stability. Scatterplots of mean estimated regression coefficients of the first 12 prognostic genes (mean across all training sets and subgroups A, B) for $p = 100$ uncorrelated covariates, $n = 50, 200$ and $\epsilon = 0, 1$ are displayed in Figure 4.6. Similar results are obtained for other correlation structures. When $\epsilon = 0$, the combined and weighted model with estimated weights provide very similar results. They identify common effects better than the subgroup model when the sample size is small ($n = 50$) and $p = 12, 100$. When $n = 50$ and $p = 1000$, all models fail to correctly estimate the coefficients of the prognostic genes and estimate all effects close to zero. For $n \geq 200$ the subgroup model detects common effects at least as well as the other two model approaches and outperforms the latter in identifying subgroup-specific effects. The combined and estimated weights approach tend to average effects across all subgroups, which becomes apparent with regard to subgroup-specific effects. When $\epsilon = 1$, the weighted model with estimated weights improves in detecting subgroup-specific effects and, in this regard, moves further away from the combined model. It performs similarly well as the subgroup model when the sample size is large ($n \geq 200$), and outperforms the latter when the sample size is small compared to the number of covariates ($n = 50$, $p > 12$, or $n = 200$, $p = 1000$). The best parameter

setting for weights estimation is random forest including interactions and cumulative HR, followed by the lasso. Using fixed weights in the Cox model results in subgroup-specific effects that mostly lie between the subgroup and combined model (depending on the weight size).

Corresponding mean inclusion frequencies (MIFs) of the prognostic genes are shown in Figure B.4 in the Appendix. The MIF of a certain covariate is defined as the proportion of subsampling data sets in which that covariate is included in the model ($\hat{\beta}_j \neq 0$). MIFs agree with the results of estimated regression coefficients. For joint effects, MIFs of the combined and weighted model with estimated weights are higher than MIFs of the subgroup model when the sample size is small ($n = 50$). For increasing sample size the MIFs of all models increase too. When $\epsilon = 0$, the MIFs of the estimated weights approach are similar to the combined model. For subgroup-specific effects and small sample size, they are higher than the MIFs of the subgroup model. However, this also means that subgroup-specific effects that are present in only one group and null in the other group are more often erroneously selected in the models of the latter group by the combined and weighted approach. In contrast, when $\epsilon = 1$, MIFs of the estimated weights approach are closer to the subgroup model regarding subgroup-specific effects and closer to the combined model with respect to common effects.

The Manhattan and Euclidean distance between the estimated and true regression coefficients are computed to summarize information on the fitted Cox models across all covariates. Similar results are obtained for both distance measures. Thus, only results of the Manhattan distance are shown. Table C.3 in the Appendix outlines the top five methods with the best Cox model fit in terms of smallest mean Manhattan distance (averaged over all training sets and subgroups) for each parameter combination for data simulation. In all situations where $p > 12$ the distance is computed twice, either based on the first 12 prognostic genes or based on all p genes. When $\epsilon = 0$, $p \leq 100$, $n \geq 200$, or $p = n = 1000$, then standard subgroup analysis performs best, followed by fixed weights of increasing size. When sample size is small relative to the number of covariates ($p \leq 100$, $n = 50$, or $p = 1000$, $n < 1000$), mainly fixed weights of different size are among the top five methods. For $\epsilon = 1$, $n = 50$, $p \leq 100$ mostly estimated weights perform best, followed by fixed weights $w > 0.5$. When $p = 1000$, fixed weights $w \geq 0.5$, as well as the combined model, are most frequently among the best methods, followed by estimated weights with random forest and ridge. For increasing sample size the standard subgroup model and estimated weights approach outperform the combined model and fixed weights. The standard subgroup model provides the best model fit for $\epsilon = 1$, $n = 1000$, $p \leq 100$.

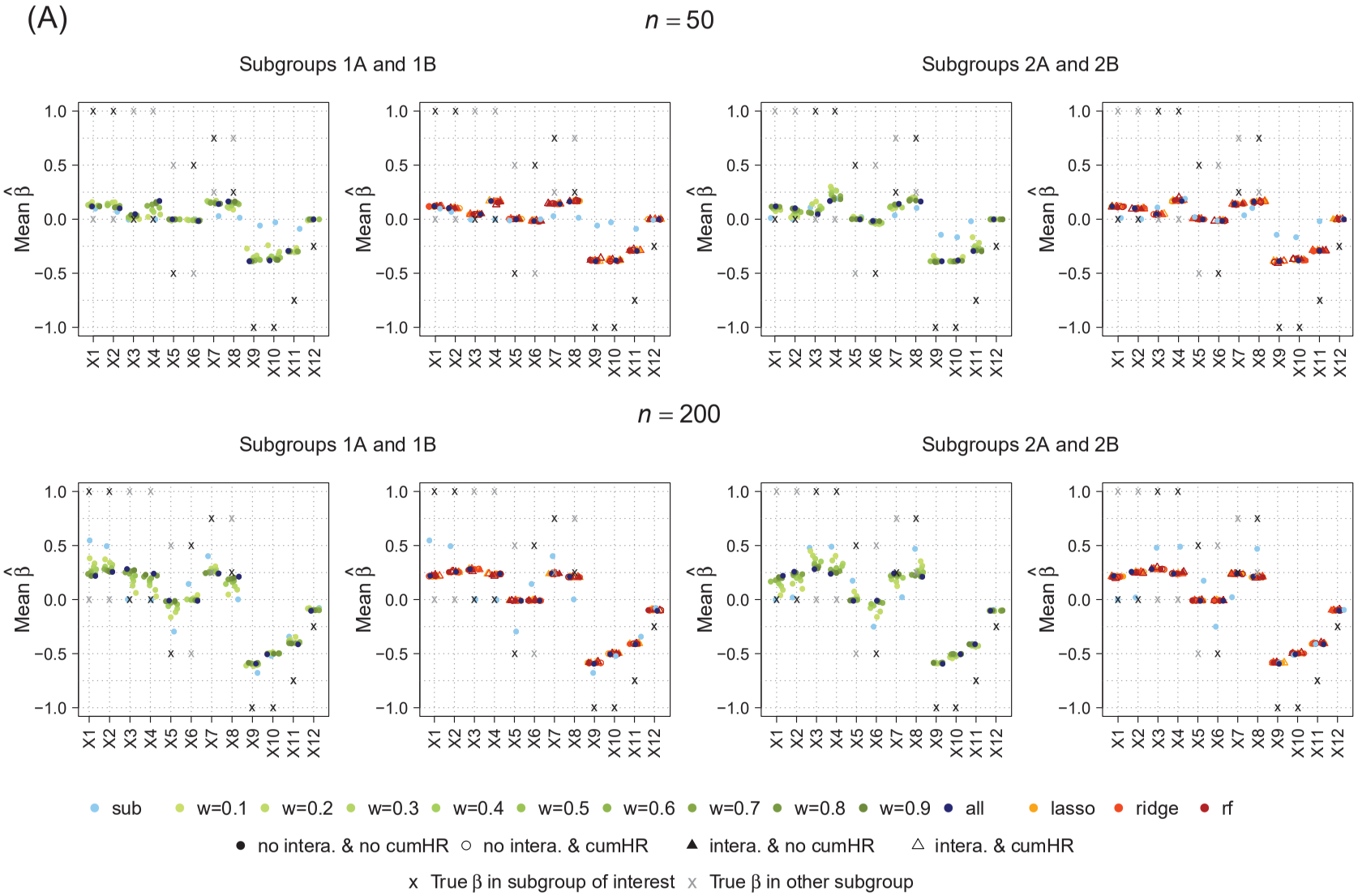


FIGURE 4.6: *Mean estimated regression coefficients of the Cox model for different model types (colors) and parameter settings for weights estimation (point symbols). Results are based on simulated data with* $p = 100$ *uncorrelated predictors,* $n = 50, 200$*, and (A)* $\epsilon = 0$*, (B)* $\epsilon = 1$*.*

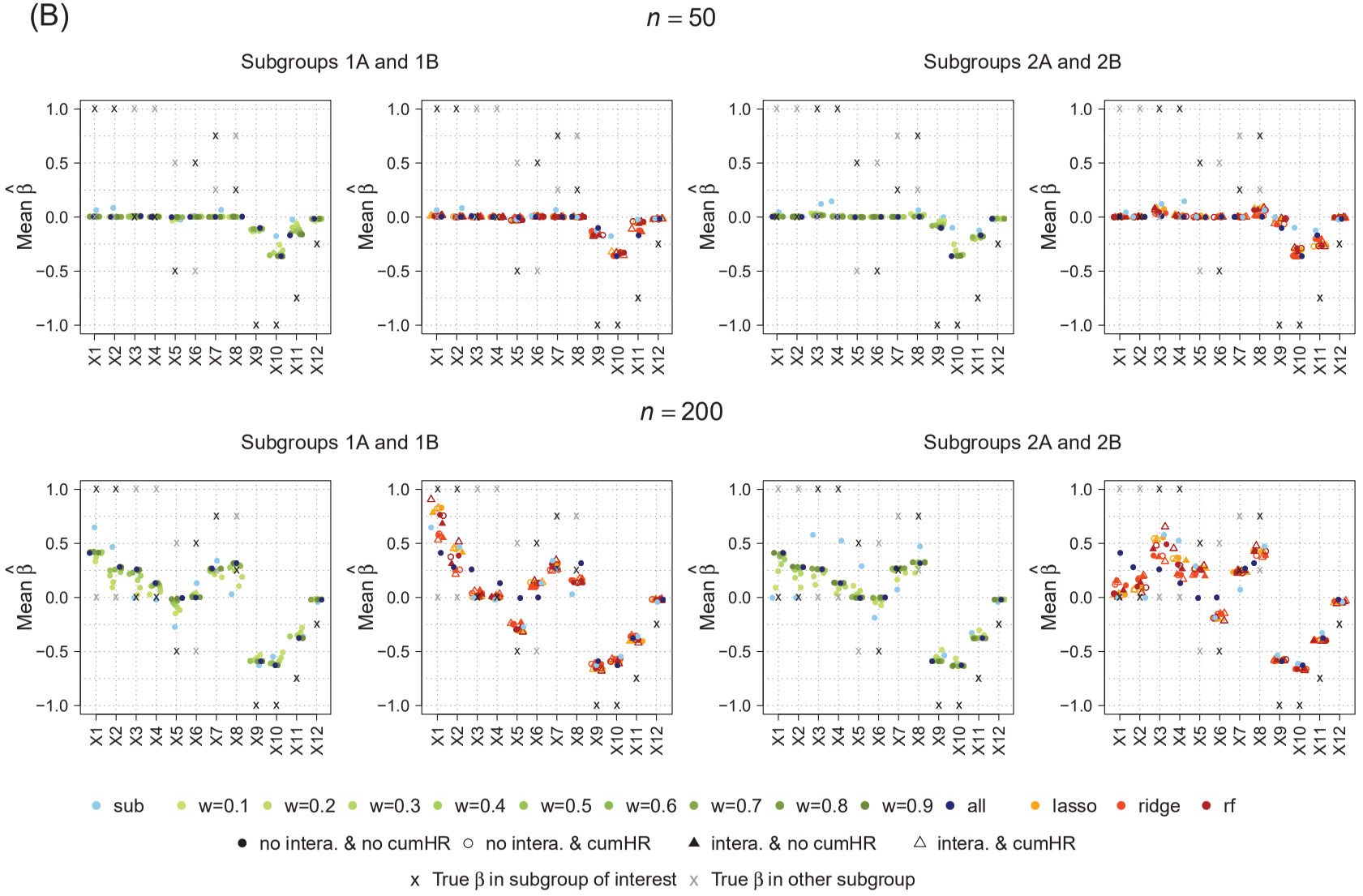


FIGURE 4.6: *Mean estimated regression coefficients of the Cox model for different model types (colors) and parameter settings for weights estimation (point symbols). Results are based on simulated data with $p = 100$ uncorrelated predictors, $n = 50, 200$, and (A) $\epsilon = 0$, (B) $\epsilon = 1$ (cont.).*

Finally, the prediction performance of all Cox models is assessed in terms of C-index and integrated Brier score (IBS). For all parameter settings and model types, the mean C-index (averaged across all test sets and subgroups) is displayed in Figure 4.7. The corresponding plot for the mean IBS is depicted in Figure B.5 in the Appendix. High values of the C-index (close to one) indicate a good predictive discrimination, in contrast to the IBS, where small values (close to zero) speak for a high predictive accuracy. Benchmark values for the Brier score are $\frac{1}{3}$, $\frac{1}{4}$, and the Brier score of the Kaplan-Meier estimator of a null model without covariates. For the C-index a benchmark value is 0.5 corresponding to random prediction. Figure 4.7 and Figure B.5 (in the Appendix) show that for $\epsilon = 0$ the combined model and estimated weights approaches have the same prediction performance, that is better compared to the standard subgroup model when sample size is small (C-index: $n = 50$, $p > 12$; IBS: $n = 50$, $p < 1000$, or $n < 1000$, $p = 1000$). However, when the sample size increases ($n > 50$, $p < 1000$, or $n = p = 1000$), the standard subgroup model outperforms the other methods. When $\epsilon = 1$ and $p \leq n$, the combined model performs worse than the weighted model with estimated weights. In all other situations, both approaches perform similarly well. The estimated weights approach performs better than the standard subgroup model when $n = 50$ and otherwise provides comparable predictive ability.

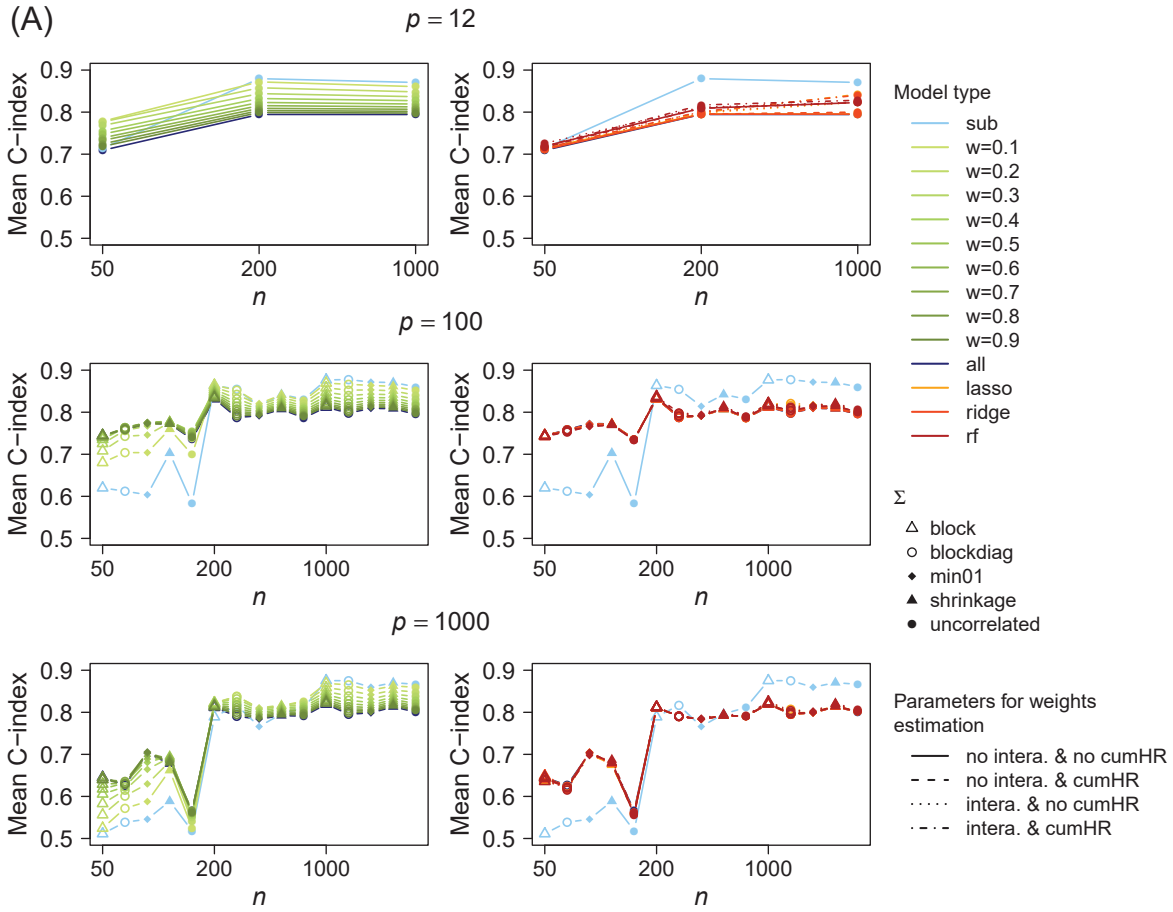


FIGURE 4.7: *Mean C-index for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$) with (A) $\epsilon = 0$, (B) $\epsilon = 1$.*

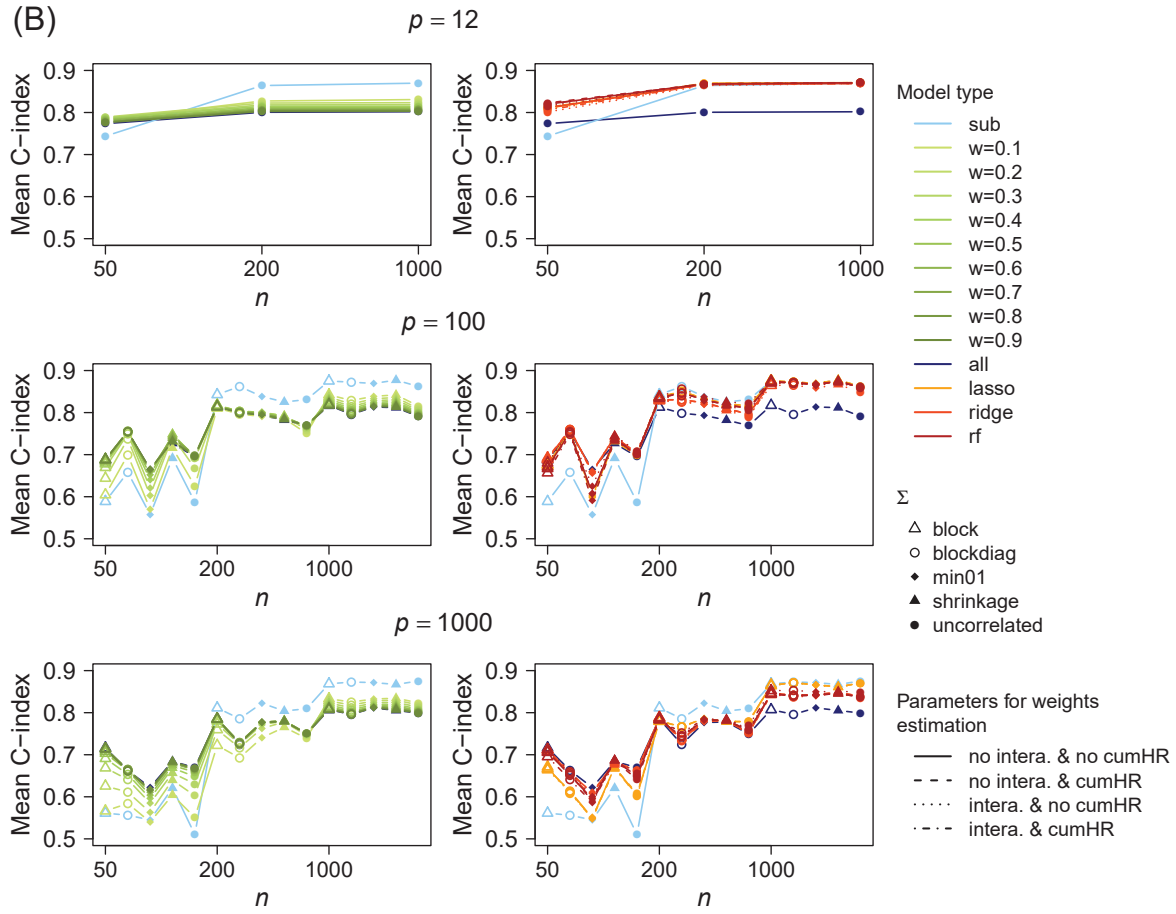


FIGURE 4.7: *Mean C-index for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$) with (A) $\epsilon = 0$, (B) $\epsilon = 1$ (cont.).*

The most important parameter settings and model types that exhibit the largest differences in prediction performance are summarized in a regression tree with C-index and IBS, respectively as response (Figures B.6 and B.7 in the Appendix). Apart from subgroup membership, all parameters for data simulation, weights estimation and Cox model types are included as covariates. The most important splitting variable is the sample size ($n \lessgtr 50$), followed by the number of covariates ($p \gtrless 1000$). The larger n and smaller p, the better the predictive ability. When $n = 50$ and $p = 1000$, the subgroup model and small fixed weights (C-index: $w < 0.3$, IBS: $w < 0.2$) perform worst. The effect of different correlation structures among covariates is ambiguous. The largest C-index is obtained for $n = 1000$, $\epsilon = 1$ and the subgroup or weighted Cox model with estimated or fixed weights $w < 0.4$. The IBS is smallest for $n > 50$.

For each parameter combination for data simulation, the top five methods with respect to prediction performance in terms of highest mean C-index or smallest mean IBS (averaged over all test sets and subgroups) are summarized in Tables C.4 and C.5 in the Appendix. When $\epsilon = 0$, $n \geq 200$, $p < 1000$, or $n = p = 1000$ the performance of the standard subgroup model is almost always best and followed by fixed weights of increasing size. For $n = 200$ and $p = 1000$, or $n = 50$ and $p = 12$ fixed weights mainly with $w < 0.6$ have the best predictive ability. For $n = 50$ and $p > 12$ fixed

weights with a trend towards larger values and estimated weights perform better than standard approaches. When $\epsilon = 1$ and $n \geq 200$ the standard subgroup model still outperforms the other models in most cases but is closely followed by the estimated weights approach. For $n = 50$ all weighted Cox models provide the best predictions, however, when $p = 1000$ the combined model and larger fixed weights ($w > 0.5$) improve over the other approaches. Comparing the prediction performance of different methods for weights estimation, lasso and random forest tend to be better than ridge.

Finally, all Cox models are compared regarding their computation time and memory. Figure B.8 in the Appendix shows that computation time and memory increase slightly for larger n when the number of covariates is small to moderate ($p = 12, 100$). However, in high-dimensional settings ($p = 1000$) both numbers rise exponentially for increasing n. Estimated weights approach with lasso takes the most time, followed by ridge, random forest, and fixed weights. Standard, unweighted approaches require the least time and memory, with only small differences between the subgroup and combined model. The largest amount of memory is consumed by ridge, followed by lasso, random forest, and all remaining models.

In summary, estimation of subgroup weights with regard to discrimination between differing groups works well for larger differences between the two groups ($\epsilon = 1$). When all subgroups are very similar ($\epsilon = 0$), classification methods fail to distinguish the two groups and all observations are similarly weighted with a weight around one, which corresponds to the standard combined model. Random forest including interactions and cumulative HR performs better than multinomial logistic regression, with a tendency of lasso towards improved results. Results of different Cox models in terms of estimated regression coefficients, variable selection stability and prediction performance indicate that for $\epsilon = 0$ the combined and weighted model with estimated weights perform very similar and have larger power to detect common effects than the standard subgroup model when the sample size is small ($n = 50$). However, they tend to average subgroup-specific effects across subgroups which results in biased estimates. When sample size is large ($n \geq 200$) the standard subgroup model outperforms the other approaches regarding predictive ability and identification of important covariates. For $\epsilon = 1$ the weighted model with estimated weights improves in correctly estimating subgroup-specific effects. In situations where sample size is small ($n = 50$) or smaller than the number of covariates ($n < p$), the weighted and combined model provide better prediction performance than the standard subgroup model. For $n \geq 200$ the proposed approach with estimated weights as well as the standard subgroup model perform best.

Results of the first simulation study have revealed that sample size n and degree of dissimilarity between groups ϵ have the strongest impact on the performance of the weights estimation and Cox models. Thus, a second simulation study is performed to investigate the effect of both parameters more closely by choosing a larger number of different parameter values. Different correlation structures have shown some variation, however, no clear conclusions can be drawn. Therefore, only block correlation (for $p > 12$) and uncorrelated covariates are subsequently considered. Due to computation time, only $p < 1000$ is further examined. All parameter combinations in the second simulation study (in total 252) are summarized in Table 4.2.

Parameter	Values (per subgroup)
n	20, 30, ..., 100, 200, 500, 1000
p	12, 100
$\boldsymbol{\Sigma}$	uncorrelated, block
ϵ	0, 0.1, ..., 0.5, 1

TABLE 4.2: *Parameter values in the second simulation study.*

First, the performance of weights estimation is assessed in terms of AUC from cross-validated training data. As before, results of AUC and ACC based on cross-validated training data and test data are very similar. Mean AUC values (averaged over all cross-validated training sets) for different parameter settings are depicted in Figure B.9 in the Appendix. For $\epsilon = 0$ the highest AUC is obtained with random forest including interactions and cumulative hazard rate. In contrast to random forest, ridge and lasso perform best without cumulative HR. The same applies to $\epsilon = 0.1$ and $n \leq 100$. For increasing sample size n, the performance of ridge and lasso becomes better compared to random forest, with a trend towards slightly improved results for ridge over lasso. Including interactions is no longer improving classification performance. For larger values of ϵ performance tends to be better without interactions and without cumulative HR, however, these parameters result in only minor differences. In low-dimensional settings ($p = 12$), ridge exhibits the highest AUC, followed by lasso. When $p = 100$ and $n < 50$ random forest performs almost always best. For $n \geq 50$ lasso performs best and for large sample size ($\epsilon = 0.2$ and $n = 500$, or $\epsilon = 0.3, 0.4$ and $n = 200$, or $\epsilon = 0.5$ and $n = 100$) ridge is equally good as lasso. For $\epsilon = 0.5$ the AUC is close to one except for $p = 100$ and $n < 50$, and for $\epsilon = 1$ the AUC is approximately one in almost all cases. A regression tree with AUC as response and all parameters for data simulation and weights estimation as predictors shows that $\epsilon \geq 0.3$ leads to an almost perfect discrimination between both groups (AUC>0.9). $\epsilon = 0$, or $\epsilon = 0.1$ and $n < 50$ display the worst performance (AUC<0.6) (Figure 4.8).

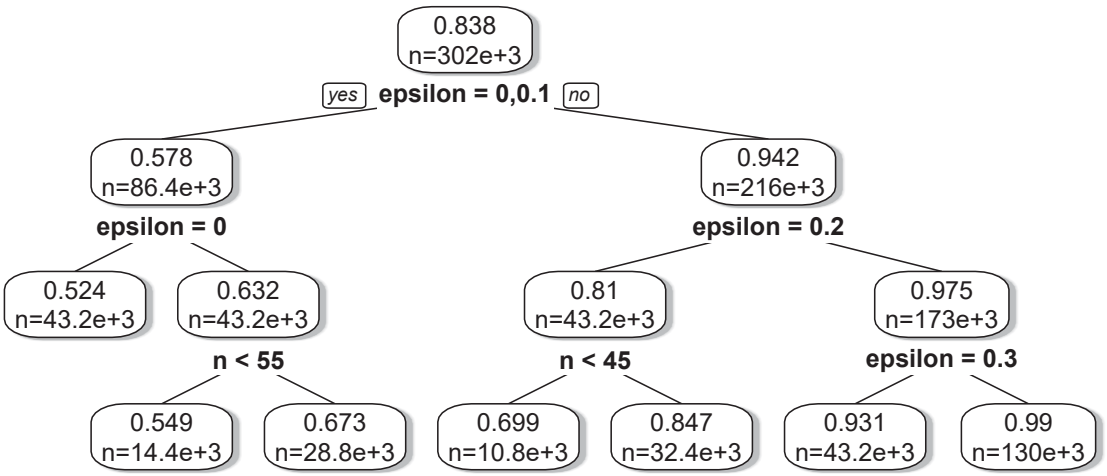


FIGURE 4.8: *Regression tree with AUC based on cross-validated training data as response and all parameters for data simulation and weights estimation as covariates. Each box shows mean AUC and sample size in the corresponding node.*

Next, results of all Cox models are evaluated with respect to estimated regression coefficients. Scatterplots of mean values (averaged across all training sets) indicate that the quality of the model fit strongly depends on the parameters ϵ, n and p. The standard combined model and the estimated weights approach identify common effects equally well with larger power compared to the standard subgroup model in all scenarios with $p = 100$ and $n \leq 100$. Results for fixed weights lie between the standard subgroup and combined model. The subgroup model tends to estimate subgroup-specific effects better than the estimated weights approach, especially for increasing sample size, and for $n \geq 200$ the subgroup model always outperforms the other model approaches. For increasing values of ϵ and n the model with estimated weights detects subgroup-specific effects increasingly better than the combined model, and similarly well or even better than the standard subgroup model when $\epsilon \geq 0.3$, $p = 100$, $n \leq 60$. These findings agree with the corresponding mean inclusion frequencies (MIFs), that increase with growing sample size. The standard combined model and the estimated weights approach mainly have larger MIFs than the standard subgroup model. This has a positive impact on the detection of common effects, but subgroup-specific effects that are present in only one group may be biased towards and overestimated in the other group. For increasing ϵ, the MIFs of the model with estimated weights move closer to the MIFs of the subgroup model. This effect is stronger for lasso compared to ridge and random forest.

Prediction performance of the Cox models is assessed by C-index and integrated Brier score (IBS). Mean values of the C-index (averaged over all test sets and subgroups) for all model types and parameter settings are displayed in Figure 4.9 for $\epsilon = 0.2$ and in Figure B.10 in the Appendix for all remaining values of ϵ. For $\epsilon = 0$ the combined and weighted model with estimated weights exhibit the same performance. For increasing ϵ and n the weighted approach performs better than the combined model. Lasso and ridge improve compared to random forest for larger n. The standard subgroup model has worse predictive ability than the combined and weighted model when $p = 100$ and $n \leq 100$. For $n > 100$ the subgroup model performs similarly well or better ($\epsilon < 0.2$) than the model with estimated weights. The prediction performance of fixed weights lies between the standard combined and subgroup model. Similar results are obtained for the mean IBS (Figure B.11 in the Appendix).

These findings are confirmed when looking at the top five methods with the highest prediction performance in terms of mean C-index and mean IBS for each simulated data setting. For $n \geq 200$ the standard subgroup model performs best, followed by fixed small weights $w = 0.1, 0.2$ ($\epsilon < 0.2$) or lasso weights ($\epsilon \geq 0.2$). When n is small, estimated or fixed weights perform best, with fixed weights more frequently for $\epsilon < 0.2$ and estimated weights for larger ϵ. Random forest and ridge tend to perform better than lasso for small sample sizes, and lasso vice versa for larger n (results not shown).

Regression trees with C-index and IBS as response, and all model types and parameter settings as covariates are shown in Figures B.13 and B.12 in the Appendix. As mentioned before, the most important splitting variable is sample size ($n \lessgtr 50$). The best prediction quality is achieved for $n > 100$ under the subgroup or weighted model with estimated or fixed weights $w < 0.4$. For $50 < n \leq 100$ performance is better for small p. The standard subgroup model performs worst for $p = 100$ and $n = 30, 40, 50$. In most cases, the predictive ability of uncorrelated covariates is worse than for the block correlation structure, which is also supported by a scatterplot of the mean C-index (Figure B.14 in the Appendix).

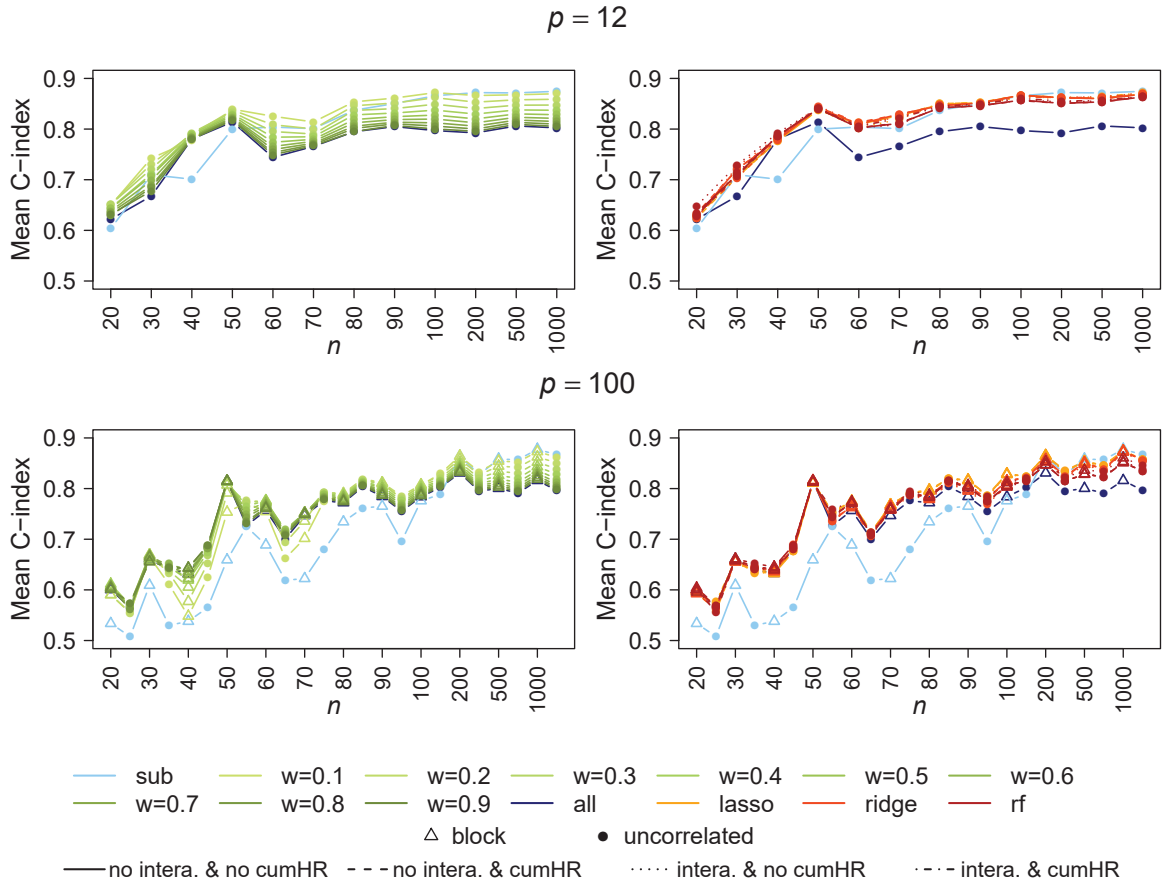


FIGURE 4.9: *Mean C-index for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$) with* $\epsilon = 0.2$.

In conclusion, estimation of subgroup weights improves with increasing differences between groups ($\epsilon > 0$) and increasing sample size. Random forest performs better than multinomial logistic regression for small ϵ or low sample sizes ($p = 100$ and $n < 50$). The larger n the better the discriminative ability of lasso and ridge. The inclusion of interactions and cumulative HR in the classification model only improves predictive quality of random forest for $\epsilon \leq 0.1$. Lasso and ridge tend to perform better without cumulative HR. The prediction performance is high for $\epsilon \geq 0.3$ and almost perfect discrimination between both groups is reached for $\epsilon \geq 0.5$. Main results of the Cox models can be summarized as follows. When sample size is low compared to the number of covariates ($n \leq 100$ and $p = 100$), the standard combined model and the weighted model with estimated weights identify common effects better than the standard subgroup model and have a higher predictive ability. However, for increasing sample size subgroup-specific effects are more precisely estimated by the standard subgroup model and for $n \geq 200$ this approach exhibits the best performance. When differences between subgroups become larger, the proposed approach with estimated weights improves over the combined model in identifying subgroup-specific effects and prediction performance. It performs similarly well as the standard subgroup model when $\epsilon \geq 0.2$ and even better for small sample sizes. Results of the weighted Cox model

with fixed weights lie between the standard subgroup and combined model. Fixed weights tend to perform better than estimated weights when $\epsilon < 0.2$. With regard to correlations between covariates, there is a trend towards improved results under the block correlation structure compared to uncorrelated covariates.

4.1.2.1 Results with CoxBoost

Results of the first simulation study in the previous section, obtained by Cox regression models with lasso penalty, are now compared to componentwise likelihood-based boosting for Cox models implemented in the R package `CoxBoost` (version 1.4). This approach also performs variable selection by using a penalized partial log-likelihood. The algorithm starts with all parameter estimates (regression coefficients) set to zero. In each boosting step, univariate candidate models are considered for each covariate and the best candidate variable is determined that improves the fit most with regard to the penalized score statistic. Only this variable is updated in the corresponding boosting step by adding the current parameter estimate to the estimate from the previous boosting step. Parameter estimates of all other covariates remain unchanged.

There are two tuning parameters: the number of boosting steps corresponding to the number of updates for the estimated regression coefficients, and the penalty parameter in the partial log-likelihood controlling the size of the updates. The former is optimized by 10-fold cross-validation. In accordance with Weyer and Binder (2015) and Matsui, Buyse, and Simon (2015), the penalty parameter is set to $\sum_m \delta_m(1/v - 1)$ which results in updates approximately the size of v times the maximum partial likelihood estimates. δ_m is the binary event indicator for observation m. The parameter v is selected in increments of 0.01 from 0.1 up to 0.01, such that the optimal number of boosting steps determined by cross-validation is larger than 50 as suggested in Tutz and Binder (2006) and Binder et al. (2009). The range of values v is chosen according to Matsui, Buyse, and Simon (2015) with lower limit $v = 0.1$ being the default value in the corresponding R package.

The componentwise likelihood-based boosting algorithm (CoxBoost) is applied to the same training and test data, and uses the same estimated weights as before in the Cox lasso model. Thus, differences in the results are only due to different Cox model algorithms and not due to variation in the underlying data. For reasons of computing time, only simulation settings with $p < 1000$ covariates are considered for subsequent comparisons. Results of both Cox algorithms are averaged over all subgroups and subsampling data sets. They are compared with respect to prediction accuracy (mean C-index and mean IBS), as well as the mean distance between true and estimated regression coefficients. Prediction performance results are displayed in Figure 4.10. They show that Cox lasso outperforms CoxBoost when including weights (particularly estimated weights) in settings with $\epsilon = 1$. This refers to situations where both groups are better distinguishable and the proposed Cox model with estimated weights improves over the other approaches. In all other cases, both algorithms have similar performance. These findings are confirmed when looking at Manhattan distances of the estimated regression coefficients in Figure B.15 (in the Appendix). Here it becomes more clear that differences between both methods also depend on the dimension of data, with large sample size compared to the number of covariates leading to larger differences. Results based on the Euclidean distance are similar.

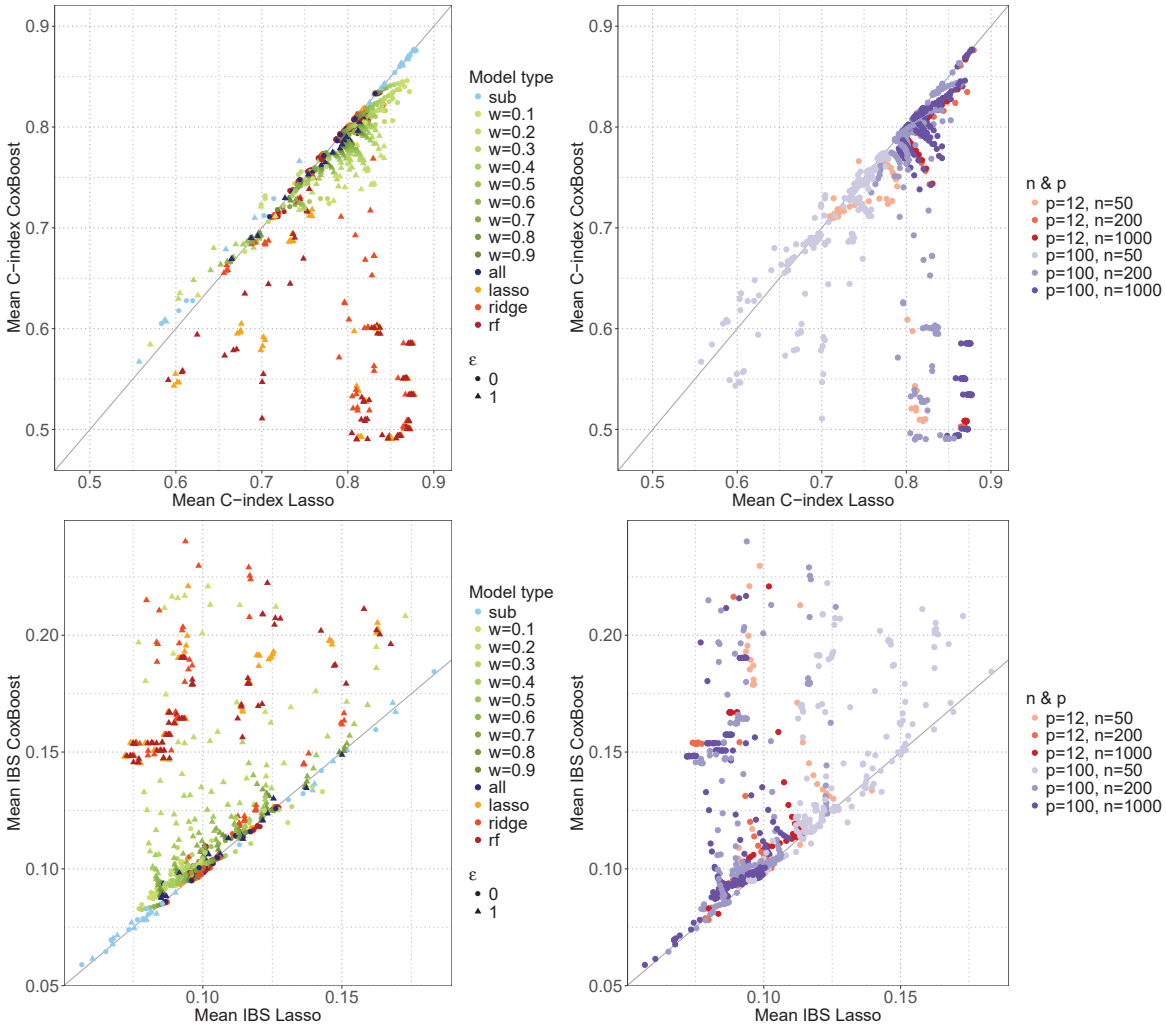


FIGURE 4.10: *Mean C-index (top row) and mean IBS (bottom row) (averaged over all subgroups and test sets) for CoxBoost vs. Cox model with lasso penalty.*

4.1.2.2 Unbalanced subgroup sizes

Hitherto, the same sample size has been assumed for all subgroups. However, this is rarely the case in practice. In this section, the effect of unbalanced subgroup sizes on prediction performance of weights estimation and Cox models is investigated. Therefore, subgroup 1A is simulated with $n^{(1)} = 50$ observations and the remaining three subgroups are generated with increased sample size of $n^{(2)} = 100$ and $n^{(2)} = 200$, respectively. The degree of similarity between group 1 and 2 is varied by choosing $\epsilon = 0, 0.2$. In all settings, the number of covariates is $p = 100$ with underlying block correlation structure. Two oversampling methods (random oversampling and synthetic minority oversampling technique,
estimation of subgroup weights, and results are compared to classification without oversampling. Oversampling increases sample size of the minority class 1A so that it is balanced with respect to the other subgroups.

Predicted probabilities of subgroup membership depend on sample sizes as shown in Figure B.16 (in the Appendix). As expected, predicted probabilities for subgroup 1A are always smaller compared to the other subgroups. However, this effect is compensated

for when predicted probabilities are divided by the relative frequencies of each subgroup to obtain the estimated weights (Figure B.17 in the Appendix). Similar results are obtained for $n^{(2)} = 100, 200$, and oversampling techniques seem to have no effect on classification performance (see Figure 4.11, and Figure B.18 in the Appendix). Figure 4.11 shows that when $\epsilon = 0$, random forest including interactions and cumulative HR has the best predictive quality in weights estimation, while lasso (without interactions and cumulative HR) has competitive ($n^{(2)} = 200$) or better performance ($n^{(2)} = 100$) than random forest when $\epsilon = 0.2$. As before, results of ACC and AUC based on cross-validated training and test data are very similar.

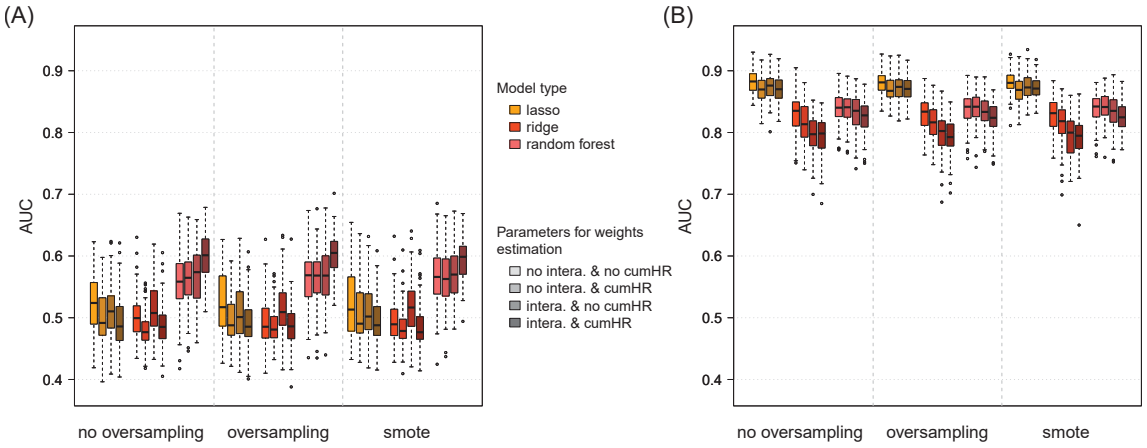


FIGURE 4.11: *Boxplots of AUC values based on cross-validated training data for* $n^{(2)} = 100$, *and (A)* $\epsilon = 0$, *(B)* $\epsilon = 0.2$.

Prediction performance of the Cox models is assessed in terms of C-index (see Figure 4.12) and integrated Brier score (IBS) (see Figure B.19 in the Appendix). Both performance measures indicate that prediction accuracy of the standard subgroup model for prediction of subgroup 1A is much worse compared to the other Cox models that benefit from increased sample size. With regard to prediction of the remaining three subgroups, the combined and proposed weighted Cox model exhibit similar performance when $\epsilon = 0$ and outperform the standard subgroup model when $n^{(2)} = 100$. When $\epsilon = 0.2$, the proposed weighted model tends to have the highest predictive quality, followed by small fixed weights. These findings agree with previous simulation studies.

In summary, unbalanced subgroup sizes affect prediction performance of the standard subgroup Cox model, with worse predictive accuracy regarding the minority subgroup. The standard combined and weighted Cox models are hardly influenced by unequal sample sizes. Differences in estimated weights due to sample size cannot be established, and oversampling techniques show no effect in the present simulation study.

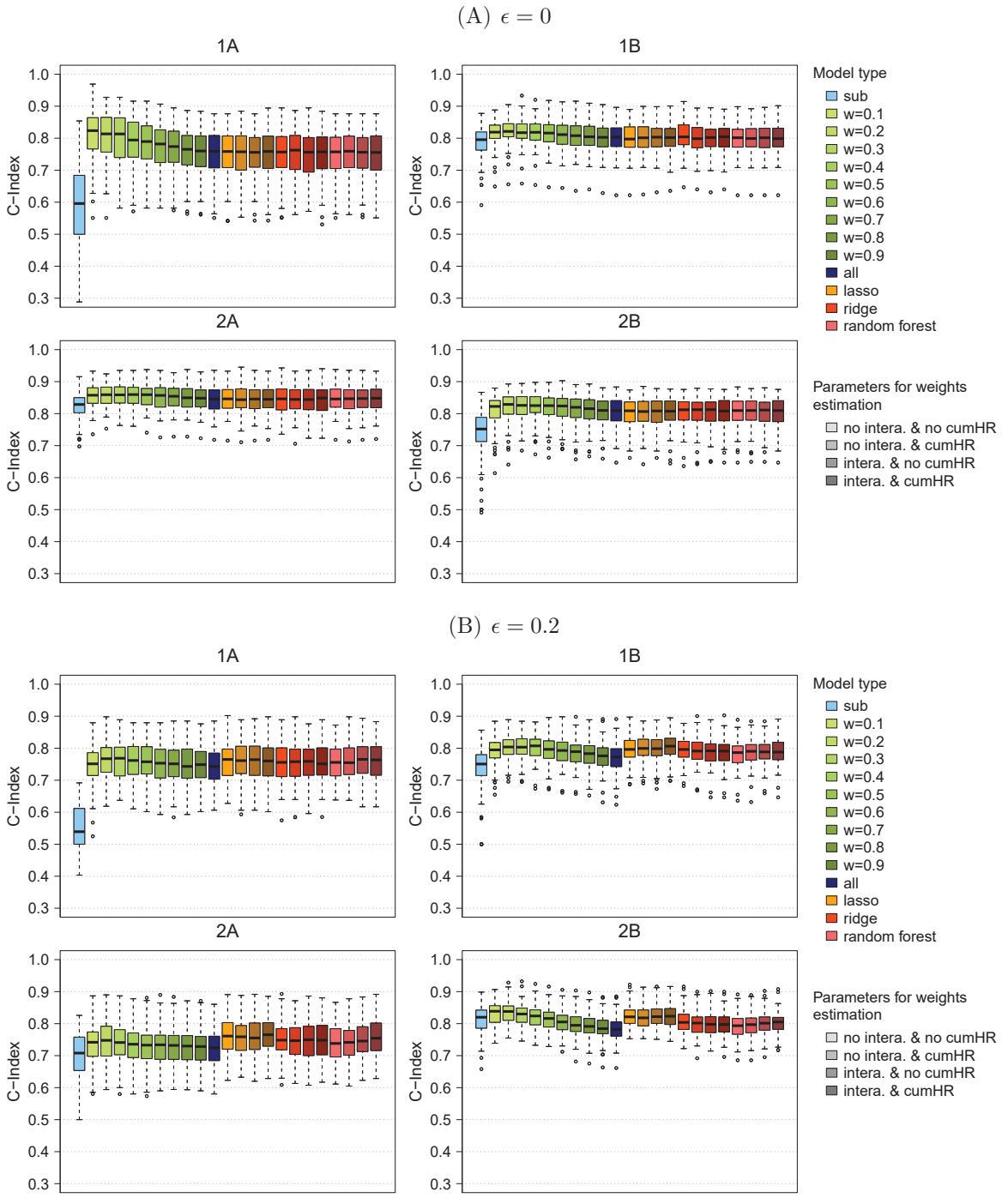


FIGURE 4.12: *Boxplots of C-index based on all test sets for the prediction of each subgroup, with $n^{(2)} = 100$ and weights estimation without oversampling.*

4.1.3 Application to lung cancer studies

In this section, the proposed weighted Cox model is applied to four lung cancer studies (see chapter 2.2) that are considered as subgroups. According to the preceding simulation studies, different parameters for weights estimation are investigated and Cox model results are compared to the standard subgroup and combined model, as well as to a weighted model with fixed weights. All Cox models are assessed with regard to prediction performance and variable selection. Another objective is the examination of the additional predictive value of genomic predictors over the following five established clinical predictors: age, sex, pTNM stage, histology, and smoking status. Therefore, all models are fitted comprising only genomic covariates, a combination of genomic and clinical covariates, and only clinical covariates. Three different pre-specified sets of genes are considered for analysis: all available genes ($p = 54675$), top-1000-variance genes ($p = 1000$), and a literature-based selection of prognostic genes ($p = 3429$) (see chapter 2.2). Genomic covariates are penalized and subject to variable selection, whereas clinical information is included as mandatory covariates when combined with gene expression variables.

First, results of weights estimation are reported, including three different classification methods (logistic regression with lasso or ridge penalty and random forest) with/without interactions between genes and survival time, with/without cumulative HR instead of survival time, and with/without oversampling to balance sample sizes. Interactions are only considered for the top-1000-variance and prognostic genes due to the already large number of covariates accompanied by high computation time. Oversampling is applied to all subgroups apart from the largest subgroup (GSE31210) to make all sample sizes equal.

Figure 4.13 shows boxplots of the estimated weights for all subgroups, based on either genomic or clinical covariates only. When genomic covariates are used for weights estimation, patients belonging to the subgroup of interest receive a relatively large weight in the respective subgroup-specific model, while the contribution of all other subgroups is close to zero. Interestingly, by far the smallest subgroup GSE29013 has the highest weight in the corresponding model (on average 11 whereas medium weights of the other subgroups are 3 or 4). When clinical covariates are used for weights estimation, subgroups become more similar and seem to benefit from each other, particularly GSE29013, GSE37745, and GSE50081. There are no apparent differences between the various parameters used for weights estimation (classification methods, oversampling, interactions, cumulative HR). Results based on only genomic or the combination of genomic and clinical covariates are also very similar, regardless of the gene filter. Distinct differences in estimated weights exist exclusively between usage of clinical covariates only and inclusion of genomic covariates.

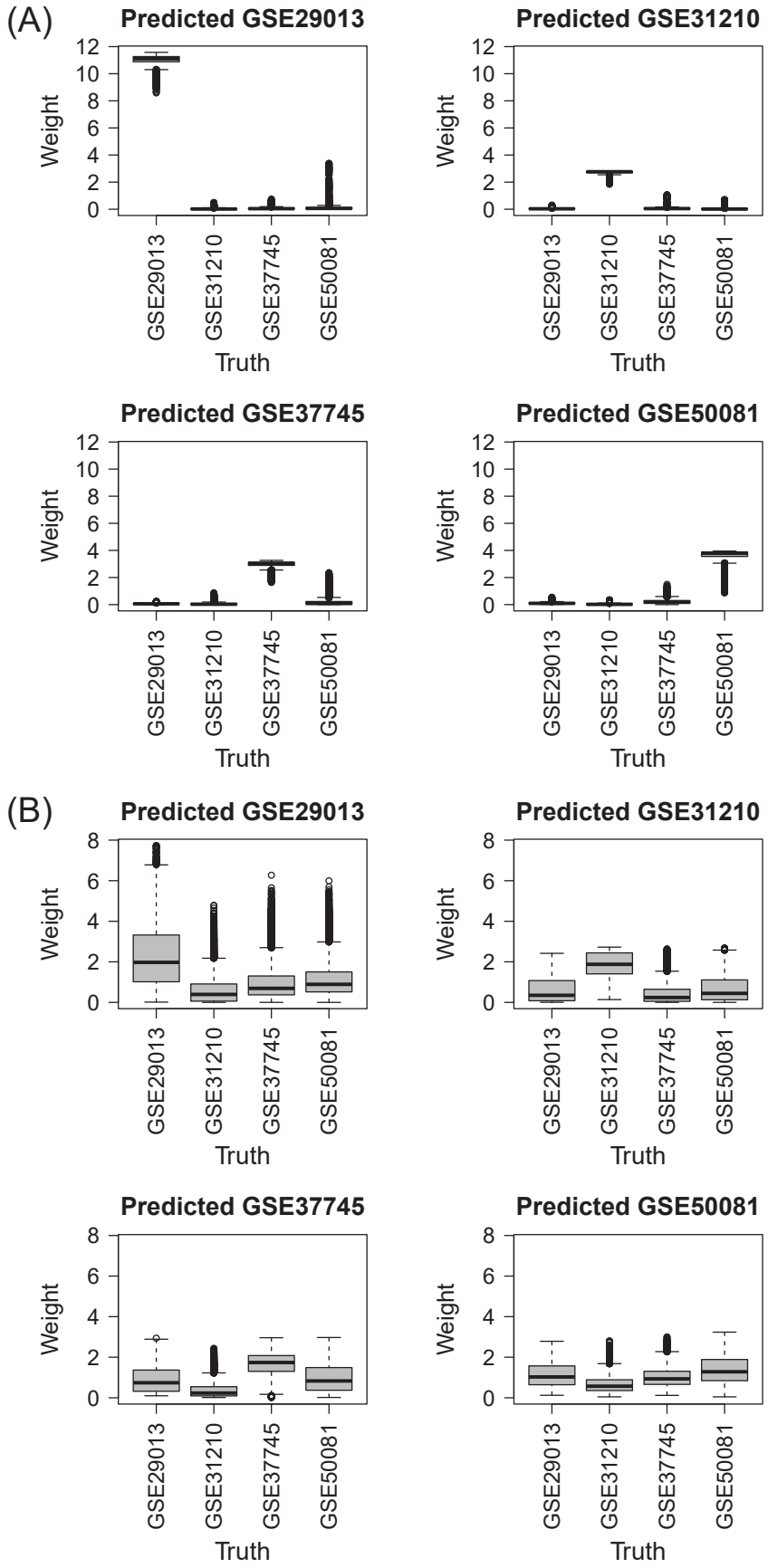


FIGURE 4.13: *Weights estimated with random forest without interactions, cumulative HR and oversampling, for all lung cancer studies including (A) only prognostic genes as covariates, (B) only clinical covariates.*

These findings are mostly confirmed by measures of prediction performance in the corresponding Cox models. Boxplots of C-index and integrated Brier score (IBS) for the comparison of different parameters for weights estimation are displayed in Figures B.20 and B.21 (in the Appendix). Mean performance values of both measures are summarized in scatterplots in Figure 4.14. In the first scatterplot, colors correspond to different covariate sets and plot symbols to the three classification methods. In the second plot, colors and symbols refer to different parameter settings. These results indicate no remarkable differences between oversampling techniques and inclusion of interactions or cumulative HR. However, IBS and C-index lead to slightly different conclusions with respect to the additional predictive value of genomic covariates. Looking at C-index, clinical covariates perform similarly well compared to the combination of clinical and genomic covariates, and better than genomic predictors only, except for GSE29013. In contrast, IBS suggests slightly improved prediction under genomic features compared to their combination with clinical covariates, apart from GSE31210. Ridge regression along with top-1000-variance genes shows improved predictive ability over the other two classification methods, whereas random forest tends to be the best method in the presence of all genes and prognostic genes.

Next, different Cox model approaches are compared with regard to their predictive quality. For the proposed weighted approach only results without interactions, cumulative HR and oversampling are considered, since these parameters have demonstrated no distinct differences. Mean C-index and mean IBS are summarized in a scatterplot in Figure 4.15, where colors correspond to Cox model types and plot symbols to covariate sets. The distribution of both performance measures across all test sets is displayed in boxplots in Figures B.22 and B.23 (in the Appendix), choosing the top-1000-variance genes as representative gene filter. Results involving genomic covariates show the highest predictive accuracy for the combined model and fixed weights of increasing size, while the estimated weights approach and standard subgroup model perform similarly bad. Random forest tends to be the best classification method in combination with prognostic and all genes, whereas ridge tends to outperform the other classification methods along with top-1000-variance genes. When only clinical covariates are included, the subgroup model performs almost always worst and all other models have similar prediction performance. In most situations, predictions of subgroups are not much better compared to random or Kaplan-Meier estimators of models without any covariates.

To assess the additional predictive value of genomic covariates over established prognostic clinical covariates, the mean prediction performance of all models including only clinical covariates is compared to models including the combination of clinical and genomic covariates. Mean C-index suggests that adding genomic covariates increases prediction accuracy of both standard models and the weighted model with fixed weights, whereas the proposed model does not seem to benefit from it. Looking at IBS leads to the conclusion that predictions for subgroup GSE37745 under the weighted model with fixed weights are improved by the inclusion of genes. But for all other subgroups prediction performance is worse when genomic covariates are added. Thus, the additional predictive value of genomic covariates remains unclear and depends on the performance measure. In standard models and models with fixed weights, genes may contribute to an improvement of prediction performance (Figure B.24 in the Appendix).

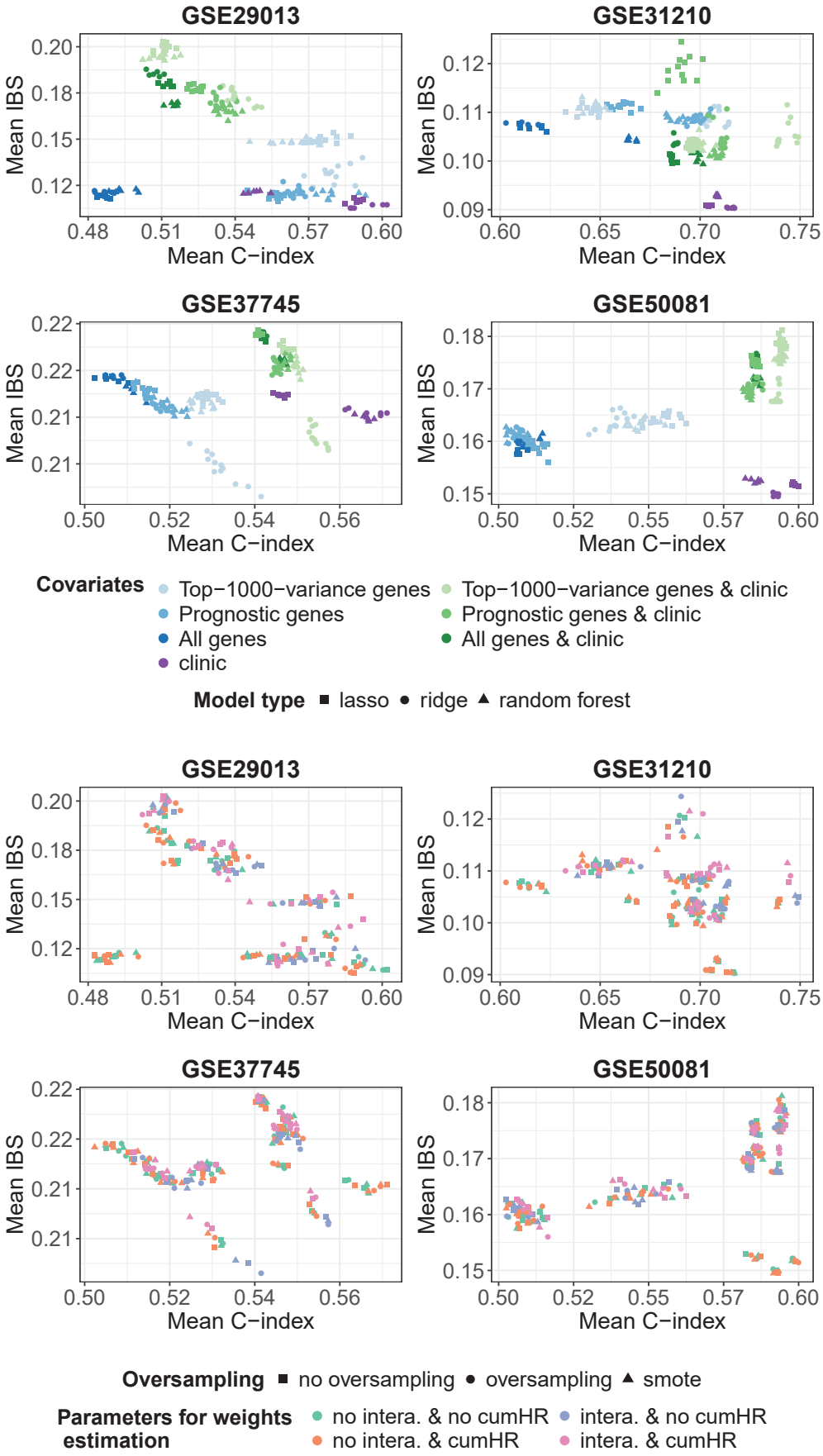


FIGURE 4.14: *Mean C-index and IBS (averaged across all test sets) for the prediction of each subgroup, under varying covariates and parameters for weights estimation.*

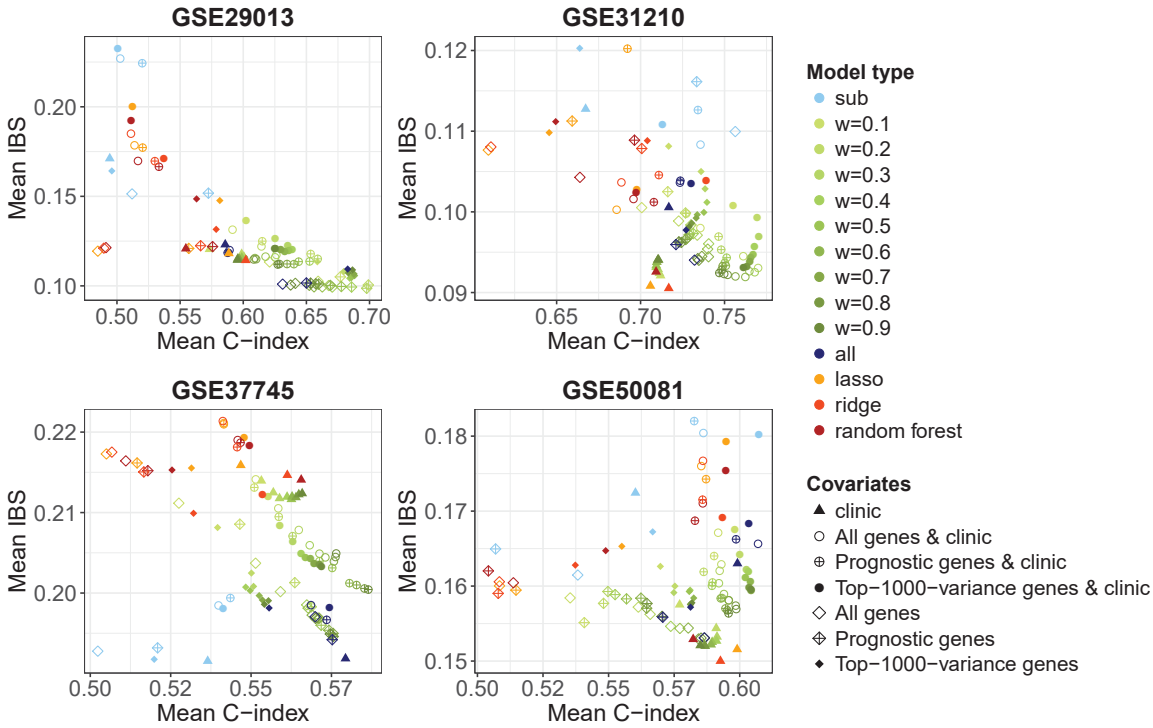


FIGURE 4.15: *Mean C-index and IBS (averaged across all test sets) for the prediction of each subgroup, comparing different types of Cox models and covariates.*

Finally, variable selection results of all Cox models are compared with regard to mean inclusion frequencies (MIFs) and estimated regression coefficients. MIFs of mandatory clinical covariates are approximately one, in contrast to the much smaller MIFs of genomic covariates that are zero for the majority of genes. The percentage of genes with MIF larger than 0.1 in any Cox model ranges from 0.1% to 0.4% for all genes, from 1.6% to 4.6% for prognostic genes, and from 6.1% to 13.6% for top-1000-variance genes (Figure B.25 in the Appendix). Analyses are based on probe set level of gene expression data, but for the illustration of variable selection results, probe set IDs are translated into gene symbols using the R/Bioconductor annotation package `hgu133plus2.db` (version 3.2.3). In case of missing gene symbols, original probe set IDs are retained such as 215780_s_at in Figure 4.16. Corresponding gene annotation is retrieved from the Ensembl website (Zerbino et al., 2018) in order to obtain gene-specific information on encoded proteins, related pathways, Gene Ontology (GO) annotations, associated diseases, and related articles in PubMed. This information is retrieved from the NCBI Gene (Brown et al., 2015) and GeneCards (*GeneCards®: The Human Gene Database*) databases.

Figure 4.16 shows, separately for each subgroup, MIFs and mean estimated regression coefficients of genes with MIF larger than 0.4 in any model type including top-1000-variance genes and mandatory clinical covariates. Eight genes are in the overlap of all subgroups illustrated by the Venn diagram, among them an immune-related gene (DEFB1) and genes (CDKN3, 215780_s_at/SET) that were reported to be associated with worse prognosis in different types of cancer such as NSCLC. These genes are most frequently selected by the combined (DEFB1) and weighted Cox model with large fixed weights (DEFB1, CDKN3, 215780_s_at/SET) and have positive effect estimates

in all subgroups. Seven of these eight jointly selected genes are also in the overlap of all subgroups based on the top-1000-variance genes only (disregarding any clinical covariates) (see Figure B.30 in the Appendix). Subgroup-specific genes with high MIFs in the standard subgroup model are, for example, cancer-related genes SPRR3 in GSE31210 and CLPTM1L in GSE50081. Publications suggest an association of SPRR3 with worse prognosis, while study results for CLPTM1L are inconsistent, reporting associations with a decreased risk of lung cancer and with lung cancer susceptibility. Contrary associations with survival may depend on genotype. The present results indicate that SPRR3 is negatively and CLPTM1L is positively correlated with longer overall survival. Other subgroup-specific, cancer-related genes with more stable selection in weighted Cox models compared to standard models include SLC7A11 and CST1 in subgroup GSE37745, WDR66 in GSE50081, as well as ADH1C and CHGB in GSE31210.

Variable selection results of all remaining covariate sets are displayed analogously in Figures B.26 to B.31 in the Appendix. Selected genes for all covariate sets are summarized in Tables C.6 to C.11 in the Appendix. Interestingly, almost all selected genes are either in the overlap of all subgroups or specific for only one subgroup, as shown in the Venn diagrams. There are hardly any genes selected by two or three subgroups, which may be due to the fact that these lung cancer studies are heterogeneous. The top-1000-variance genes not adjusted for clinical covariates have 14 genes in the overlap of all subgroups, whereof seven genes are also in the overlap of the combination of top-1000-variance genes and clinical covariates. The majority of the other seven genes (BCHE, GLS, KLF6, PLOD2) are associated with different cancers and function as a tumor suppressor (KLF6) or as support of tumor cell growth and metastasis (GLS, PLOD2). They are most frequently selected by the combined and weighted model with large fixed weights, however, corresponding estimated regression coefficients are relatively small suggesting weak effects on survival outcome compared to the other genes included in the multivariate models. Subgroup-specific genes with strong effects on overall survival and high MIFs in the proposed weighted model involve the following cancer-related genes: ADH1C and BMP5 in GSE31210, as well as AREG and COL4A3 in GSE29013. Interestingly, BMP5 was reported to be more strongly expressed in lung adenocarcinoma, constituting the entire population in GSE31210, compared to lung squamous cell carcinoma (Figure B.30 in the Appendix).

Cox models including all genes identify fewer genes compared to the other two gene filters which is likely caused by the large amount of noise genes. There are two cancer-related genes most frequently selected across all subgroups by the combined and weighted model with large fixed weights: ERN1 and MAGEH1. The latter was reported to act as a tumor suppressor by inhibiting cell proliferation, implying correlation with improved prognosis which agrees with present findings. MIFs and estimated regression coefficients of the subgroup and proposed model are mainly close to zero, except for PTGER3 that has high MIFs in GSE31210 for the proposed model with estimated weights. PTGER3 induces tumor progression in different cancer types including adenocarcinoma of the lung. This may explain the specific association with GSE31210 being the only subgroup comprising exclusively adenocarcinoma (Figure B.27 in the Appendix).

There is one gene (SPP1) that is in the overlap of all subgroups and all six covariate sets including gene expression data. SPP1 - also known as Osteopontin (OPN) - is involved in inflammatory response, osteoblast differentiation for bone formation and attachment of osteoclasts to the mineralized bone matrix for bone resorption.

It is associated with several malignant diseases and reported to promote tumor cell proliferation and worse prognosis in NSCLC. This relationship with survival is consistent with the present results of all model types, subgroups and covariate sets. SPP1 has highest MIFs in the combined and weighted Cox model with large fixed weights and smallest MIFs in the subgroup model.

Cox models including only clinical covariates that are not subject to penalization indicate a higher mortality risk of males compared to females and of current/former smokers compared to never-smokers. Advanced tumor stage (stage II-IV) is also related to worse prognosis, as well as increasing age. Adenocarcinoma (ADC) is associated with slightly better survival outcome compared to other NSCLC (Figure B.31 in the Appendix). These findings seem plausible.

In summary, application to lung cancer studies shows no distinct differences between various parameters of weights estimation (interactions, cumulative HR) and oversampling seems to have no effect on prediction in classification. Estimated weights from classification models including genomic covariates suggest that subgroups are very different from each other and resemble the standard subgroup model, where only the subgroup of interest is assigned a high weight and all other subgroups have weights close to zero. In contrast, estimated weights from classification models based on clinical covariates only suggest that all subgroups are more or less similar.

Prediction performance of Cox models indicate that logistic regression with ridge penalty and top-1000-variance genes outperforms the other two classification methods, while random forest tends to perform best in combination with all genes and prognostic genes. Cox models comprising genomic covariates show the highest predictive accuracy for the combined and weighted model with fixed weights of increasing size, while the estimated weights approach and standard subgroup model perform similarly bad. The inclusion of clinical covariates only shows a similar performance of the combined and all weighted models being superior to the subgroup model. Overall prediction performance is mostly moderate and not much better than random or reference models without any covariates.

The additional predictive value of genomic covariates over clinical covariates is unclear and may only exist in standard models and models with fixed weights. The latter exhibit in most cases the highest variable selection stability, followed by the combined model. Genes identified most frequently by these models are often present in all subgroups and some of them were reported to be associated with prognosis in various cancers. Few cancer-related genes with subgroup-specific effects are detected exclusively by the subgroup model and more stable by the proposed model with estimated weights.

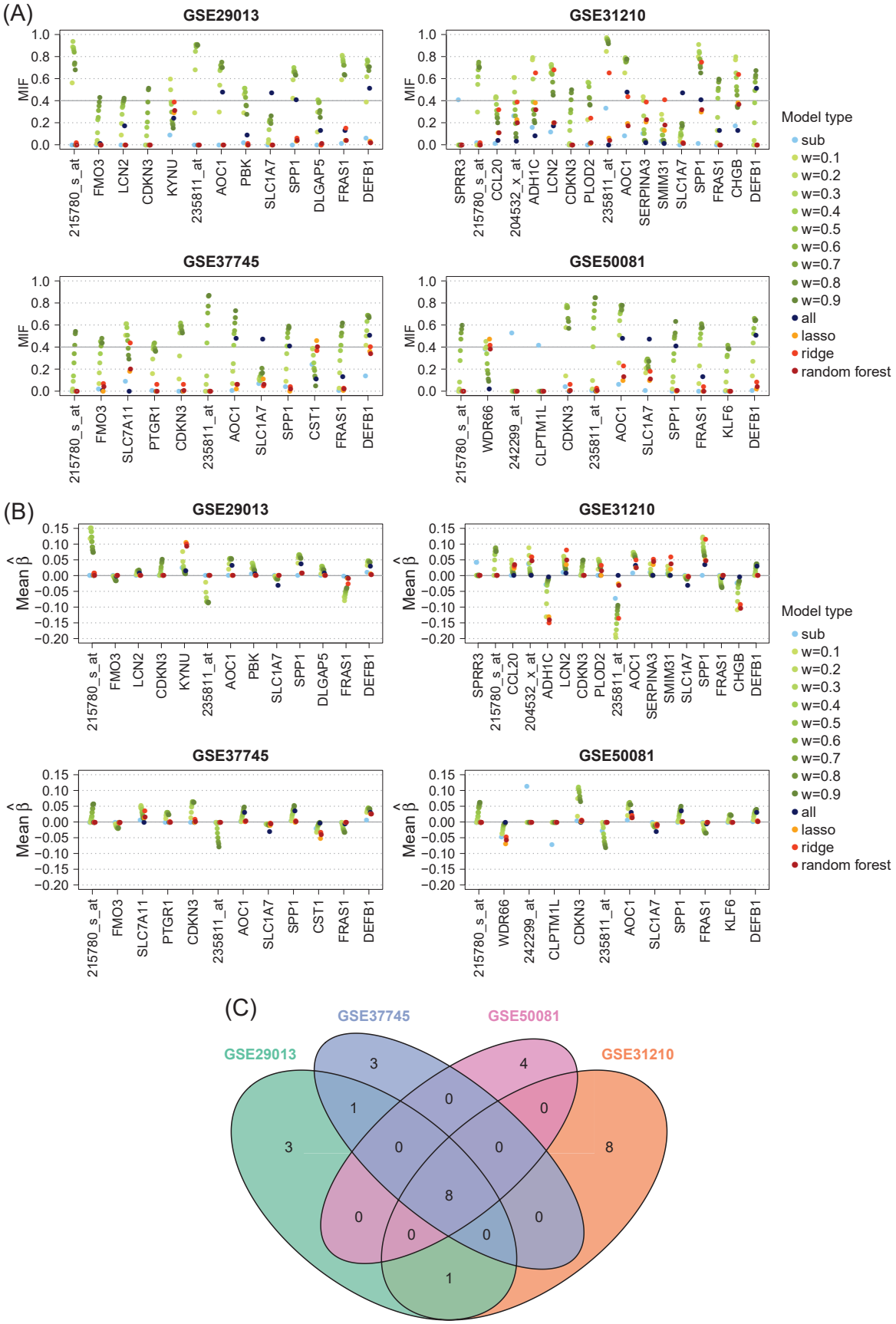


FIGURE 4.16: *Results of variable selection for Cox models including top-1000-variance genes and clinical covariates. For each subgroup genes with a mean inclusion frequency (MIF) larger than 0.4 in any model type are selected. (A) Mean inclusion frequencies, (B) mean estimated regression coefficients, and (C) Venn diagram of selected genes in all subgroups.*

4.2 Bayesian subgroup model

The proposed Bayesian Cox model is an extension of the model by Treppmann, Ickstadt, and Zucknick (2017) in the sense, that it uses a graphical model for the variable selection prior rather than an independent Bernoulli prior. An unknown graph structure linking genomic covariates within and across multiple subgroups is assumed and inferred simultaneously with the important variables in the Cox model. This encourages the selection of covariates that are both relevant to the survival outcome and related to each other in the graph. The subgraph $\boldsymbol{G}_{ss}$ within subgroup s is associated with the corresponding precision matrix $\boldsymbol{\Omega}_s$, representing conditional dependencies among the covariates. The subgraph $\boldsymbol{G}_{rs}$ linking subgroups r and s helps to identify genes that are prognostic in both subgroups.

The proposed model is compared to a standard subgroup model and to a combined model. The combined model pools data from all subgroups and treats them as one homogeneous cohort, whereas the subgroup model only uses information in the subgroup of interest and ignores the other subgroups. Both standard approaches follow the Bayesian Cox model proposed by Treppmann, Ickstadt, and Zucknick (2017) with stochastic search variable selection and independent Bernoulli priors for the variable inclusion indicators $\boldsymbol{\gamma}$. The hyperparameter of the Bernoulli distribution is chosen as $\pi = p^*/p$, where $p^* = 2$ is the a priori expected model size, resulting in an uninformative selection prior (for more details see chapter 3.3.3).

First, two preliminary analyses are conducted based on simulated data for the choice of hyperparameters and assessment of convergence of the proposed model. Section 4.2.2 reports a sensitivity analysis of variable selection for the choice of hyperparameters a, b and ν_0, ν_1. In section 4.2.3 four independent MCMC chains with different initial values of the parameters are run to assess convergence and mixing properties of the Markov chains. Afterward the proposed model is validated through simulations and applied to four lung cancer studies.

Genomic covariates are standardized before model fitting and evaluation to have zero mean and unit variance. Parameters of the training data set (mean and standard deviation of each variable) are used to scale the training and test data set. For the standard subgroup model and the proposed model each subgroup is standardized separately, whereas for the combined model training data of all subgroups are pooled.

Results of the Cox models are reported in terms of marginal posterior means and standard deviations of the estimated regression coefficients, as well as posterior selection probabilities. After removal of the burn-in samples, the remaining MCMC samples serve as draws from the posterior distribution to calculate the empirical estimates. The strategy for variable selection follows Treppmann, Ickstadt, and Zucknick (2017). First, the mean model size m^* is computed as the average number of included variables across all MCMC iterations after the burn-in. Then the m^* variables with the highest inclusion frequency (posterior selection probability) are considered as the most important variables and selected in the final model. Variable selection accuracy is assessed with regard to the number of correctly identified prognostic variables (true positives), the number of correctly rejected non-prognostic variables (true negatives), the number of incorrectly selected non-prognostic variables (false positives), and the number of incorrectly rejected prognostic variables (false negatives). Prediction performance of the Cox models is evaluated by prediction error curves, integrated Brier score and C-index.

4.2.1 Simulation setup

For the preliminary analyses and simulation study in the following three sections, a training and a test data set consisting of n observations and p genomic covariates, respectively, are simulated from the same distribution for each subgroup as described in the following. Training data are used for parameter estimation, and model performance is evaluated based on independent test data. Two subgroups are considered that differ only in the relation between genomic covariates and survival time ($\boldsymbol{\beta}_s$, $s = 1, 2$), and in the parameters for the simulation of survival times. Gene expression data are generated from the same multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The corresponding precision matrix is defined as

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 1 & 0.5 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0.5 & 1 & 0.5 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0.25 & 0.5 & 1 & 0.5 & 0.25 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0.25 & 0.5 & 1 & 0.5 & 0.25 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0.25 & 0.5 & 1 & 0.5 & 0.25 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0.25 & 0.5 & 1 & 0.5 & 0.25 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0.25 & 0.5 & 1 & 0.5 & 0.25 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.25 & 0.5 & 1 & 0.5 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0.5 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & 1 \end{pmatrix},$$

with partial correlations only between the first nine prognostic genes, the remaining non-prognostic noise genes are assumed to be uncorrelated. Survival times are simulated from a Weibull distribution with scale η_s and shape κ_s parameters, estimated from two independent lung cancer cohorts (GSE37745 and GSE50081). Event times T_s are simulated according to section 4.1.1. Noninformative censoring times C_s are randomly drawn from a Weibull distribution with different parameters as for the event times. The parameters are calculated from the Kaplan-Meier estimators of the censoring times in both lung cancer cohorts. The censoring rates at 5 years are 98% and 72%, and 81% and 19% at 7 years, respectively. The individual observed event indicators and times until an event or censoring are defined as $\delta_s = \mathbb{1}(T_s \leq C_s)$ and $\widetilde{T}_s = \min(T_s, C_s)$, $s = 1, 2$. This results in approximately 40% and 60% censoring rates in both subgroups. True effects of the genomic covariates on survival outcome are chosen as

	Gene											
	1	2	3	4	5	6	7	8	9	10	...	p
$\boldsymbol{\beta}_1$	1	1	1	0	0	0	-1	-1	-1	0	...	0
$\boldsymbol{\beta}_2$	0	0	0	1	1	1	-1	-1	-1	0	...	0

The first six genes are subgroup-specific, while genes 7, 8 and 9 have the same effect on the response in both subgroups. All remaining genes have no effect in both subgroups. The focus is on situations where most of the variables are noisy ones, to test the ability of the proposed model to identify important covariates in the presence of a varying amount of noise.

In preliminary analyses in sections 4.2.2 and 4.2.3 the same training and test data sets are used comprising $n = 100$ observations and $p = 100$ genes. In subsequent

simulation studies, data are simulated for varying n and p. Each MCMC chain is run for 20 000 iterations, where the first 10 000 are discarded as burn-in.

4.2.2 Sensitivity analysis

In the following, a sensitivity analysis of the proposed model to the choice of fixed hyperparameters a, b and ν_0, ν_1 is conducted to examine their effect on the number of selected variables (model size). Therefore, inference under varying parameter settings is performed using a simulated data set as described in section 4.2.1. First, different values of a and b are considered, while fixing $\nu_0 = 0.1$ and $\nu_1 = 10$ in accordance with Peterson, Stingo, and Vannucci (2016), and Wang (2015). Inference for all combinations of the following parameter settings is performed: $a \in \{-4, -3.5, -3, -2.5, -2\}$ and $b \in \{0.25, 0.5, 0.75, 1\}$. Results of the number of selected variables are shown in Figure 4.17 and Table C.12 (Appendix). Among the best combinations with the smallest average number of incorrectly selected variables (FP+FN) (see Table C.12 in the Appendix), the parameter setting $a = -4$ and $b = 1$ provides the best convergence and will be used henceforth. The hyperparameters a and b specify the prior probability of variable inclusion in the MRF prior of $\boldsymbol{\gamma}$. Thus, it is not surprising that increasing values of a or b result in larger models. These findings agree with Li and Zhang (2010) and Peterson, Stingo, and Vannucci (2016).

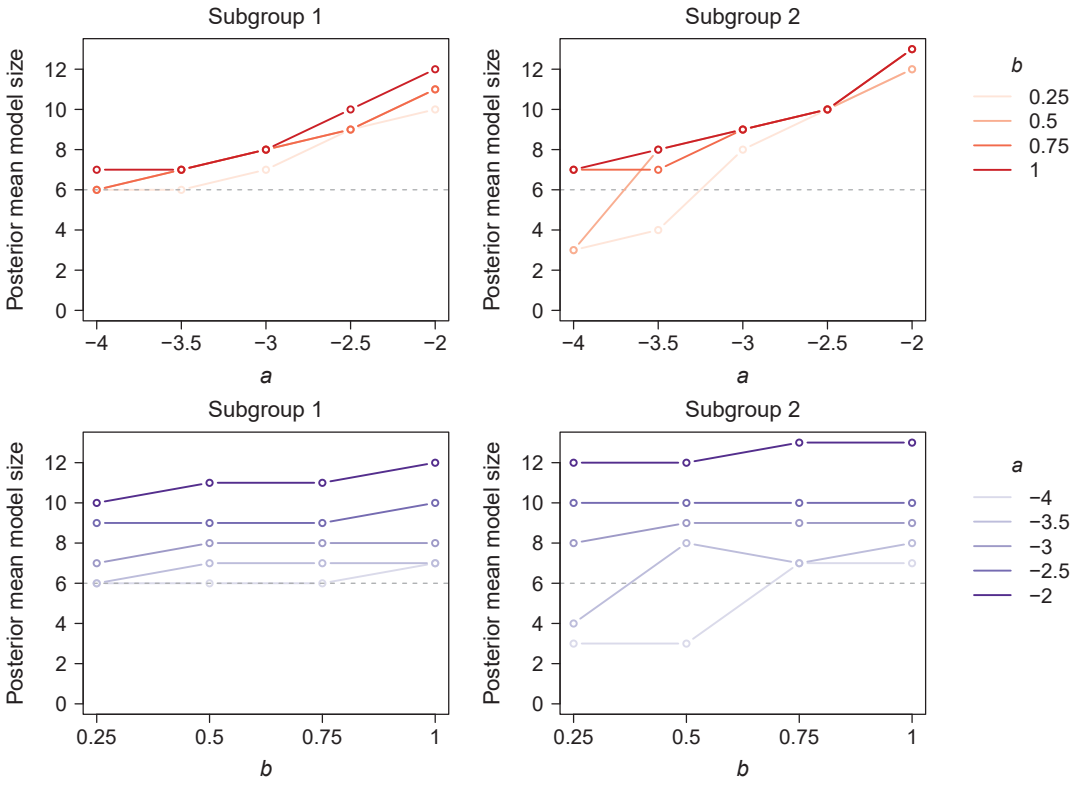


FIGURE 4.17: *Sensitivity analysis showing the effect of varying hyperparameters a and b on the average number of selected variables. The horizontal dashed gray line represents the true model size.*

Next, the choice of ν_0 and ν_1 is considered which determine the prior variances of the off-diagonal elements in $\mathbf{\Omega}_s$. All combinations of values $h \in \{10, 50, 100\}$ and $\nu_0 \in \{0.01, 0.02, 0.05, 0.1\}$ are studied, with $\nu_1 = h \cdot \nu_0$, similarly as in Wang (2015). Results are shown in Figure 4.18 and Table C.13 (Appendix), analogously to the sensitivity analysis of a and b. Among the two best combinations with the smallest average number of incorrectly selected variables (FP+FN) (see Table C.13 in the Appendix), the parameter setting $\nu_0 = 0.1$ and $h = 50$ provides the best convergence and will be used henceforth. Compared to the choice of a and b, it seems that the results are less sensitive to the choice of ν_0 and $\nu_1 = h \cdot \nu_0$.

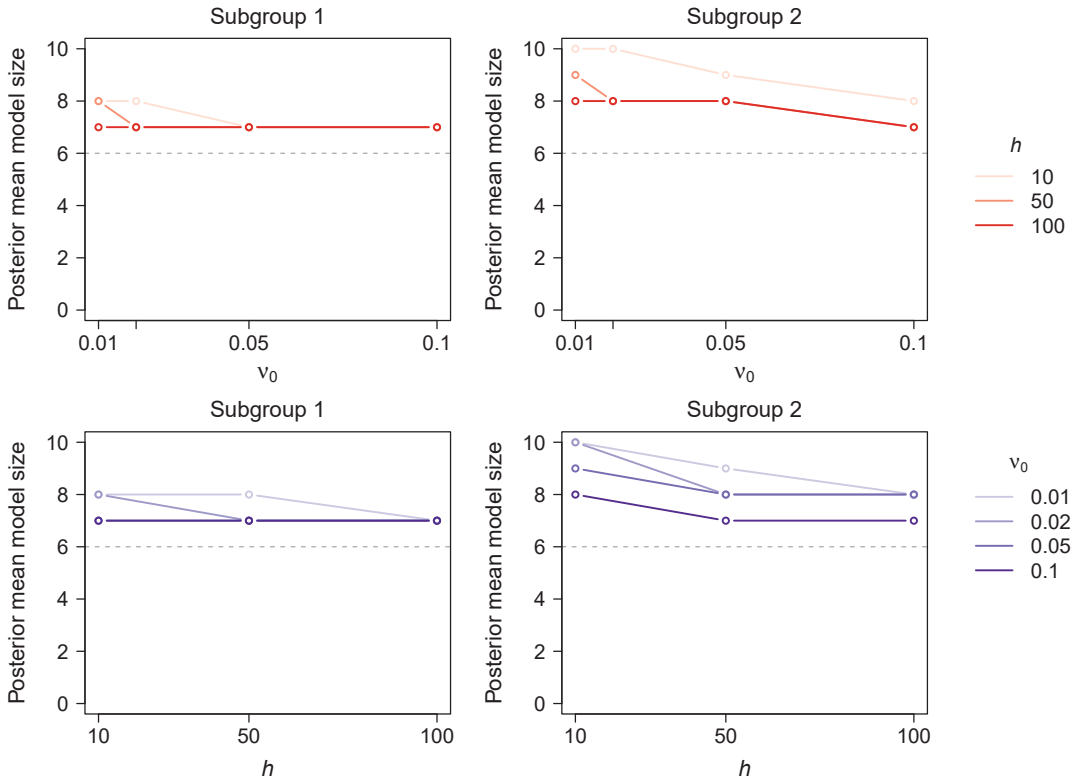


FIGURE 4.18: *Sensitivity analysis showing the effect of varying hyperparameters ν_0 and $\nu_1 = h \cdot \nu_0$ on the number of selected variables. The horizontal dashed gray line represents the true model size.*

4.2.3 Multiple Markov chains to assess convergence

To assess convergence and mixing properties of the Markov chains, four independent MCMC chains with different initial configurations are run using the same simulated data as in section 4.2.2. Starting values are chosen to be overdispersed with regard to the posterior distribution. Specifically, one chain is started from an empty model (all variable and edge inclusion indicators set to zero), one is started from a full model (all variable and edge inclusion indicators set to one), and two chains are started with 50% or 20% of the variables and edges selected at random in each subgroup. The same initial values are used for all subgroups. Starting values of the regression coefficients are chosen depending on the corresponding variable inclusion indicators. When the

latter is zero, $\beta^{(0)}$ is drawn from a uniform distribution in the interval $[-0.02, 0.02]$, and otherwise from a uniform distribution in the interval $[-2, 2]$. MCMC convergence and mixing are assessed graphically with trace plots, and plots of the autocorrelation function and the corrected version of the potential scale reduction factor (PSRF) by Gelman and Rubin (1992) of the individual regression coefficients (for more details see section 3.3.1.2). Plots of the first nine prognostic variables are shown, as well as the tenth variable as representative of the remaining noise variables, which provide very similar results.

Trace plots of the model size (Figure B.32 in the Appendix) and estimated regression coefficients (Figure B.33 in the Appendix) indicate that convergence takes place rapidly within a few iterations. The output from all Markov chains exhibit fast up-and-down variation without long-term trends or distinct patterns, and trace plots of different chains do not look much different from each other. In Figure B.34 (in the Appendix) the autocorrelation function is plotted separately for each variable in each chain. Autocorrelation values decrease rapidly and become very small in most cases, suggesting good convergence. PSRF is a diagnostic that compares the variation between and within chains and is applied to each variable separately. Figure B.35 (in the Appendix) illustrates how this shrink factor evolves when the number of iterations increases. After 10 000 iterations, point estimates of the PSRF and their upper confidence limits are always smaller than 1.05, indicating that all chains were run for a satisfactory number of iterations and have converged to a common target distribution. The Metropolis-Hastings acceptance rates for the update of β are between 0.48 and 0.52 for all chains. Furthermore, agreement of the results between the four chains is assessed in terms of pairwise correlations between the marginal posterior probabilities of variable and edge selection. Overall, the results confirm that 20 000 iterations in total with a burn-in period of 10 000 iterations seem to be sufficient to reach (approximate) convergence.

4.2.4 Simulation results

This section reports the results of two simulation studies to compare the prediction performance of the proposed model, referred to as *CoxBVSSL* (for Cox model with Bayesian Variable Selection and Structure Learning, as an extension of the model by Treppmann, Ickstadt, and Zucknick (2017)), with the standard subgroup model and the combined model. Varying number of genomic covariates and sample sizes are examined, with a focus on small sample sizes relative to the number of variables which is characteristic for gene expression data. For Bayesian inference, typically one training data set is used for parameter estimation and an independent test data set for model evaluation. However, results of the present models have shown some variation due to the data draw. Therefore, simulation of training and test data in the following is repeated ten times for each simulation scenario.

In the first simulation study two low-dimensional settings are considered with $p = 20$ genes and $n = 50, 100$ observations in each subgroup, as well as high-dimensional settings with $p = 100$ and sample sizes $n = 25, 50, 75, 100, 150$. Trace plots, running mean plots and prediction error curves are shown for the first training and test data set in an exemplary way. Results based on the remaining nine training and test data sets vary more or less. Mean posterior probabilities of variable and edge inclusion, and

posterior estimates of regression coefficients are averaged across all ten training data sets.

Convergence of each chain is assessed by looking at autocorrelations, trace plots and running mean plots of regression coefficients. In most cases, autocorrelation values decrease rapidly and become very small. Only when $n < 100$ there are single training data sets where autocorrelations of prognostic genes decrease slowly and remain relatively large in both the subgroup and CoxBVSSL model. Trace plots and running mean plots of the regression coefficients are depicted in Figures B.36 and B.37 (in the Appendix). In the low-dimensional setting for $n = 50$ MCMC convergence of the CoxBVSSL model is worse than for the standard subgroup model and estimates are less stable. For $n = 100$ both models perform well and running means stabilize already after a few iterations at the posterior means. The combined model only correctly estimates the joint effects of genes 7, 8 and 9, and with a slightly worse performance, the subgroup-specific effects of subgroup 2 (genes 4, 5 and 6). In high-dimensional settings with $p = 100$ and $n = 25$, all models fail to identify prognostic genes and mixing performances are bad. As expected, for increasing n model performance improves. The CoxBVSSL model tends to have higher power in estimating the prognostic effects than the subgroup model when $n \leq 100$. For $n = 150$ both models perform similarly well. The combined model only identifies the common effects.

These findings are in accordance with mean posterior probabilities of variable inclusion, and posterior estimates (mean and standard deviation) of regression coefficients. The CoxBVSSL model has larger power to detect true effects than the subgroup model when $p = 100$ and $50 \leq n \leq 100$. In these situations, mean posterior selection probabilities of prognostic genes are higher and regression coefficients are estimated more accurately in the CoxBVSSL model compared to the standard subgroup model. The combined model correctly identifies joint effects of genes 7, 8, and 9, but fails to detect subgroup-specific effects (Table 4.3; Figures B.38 and B.39, and Table C.14 in the Appendix). Findings support the assumption that incorporating network information into variable selection may increase power to detect associations with survival outcome.

Results are further confirmed when looking at the models' prediction performance. Prediction error curves in Figure B.41 (in the Appendix) suggest that the subgroup model and the CoxBVSSL model mainly have better prediction performance than the reference models (Kaplan-Meier estimates) and the combined model, with CoxBVSSL tending to outperform the subgroup model. Mean C-index and mean integrated Brier score values indicate that CoxBVSSL outperforms the subgroup model when $p = 100$ and $50 \leq n \leq 100$. When sample size is large ($p = 20$ and $n = 100$, or $p = 100$ and $n = 150$) the CoxBVSSL and standard subgroup model perform similarly well. When $p = 100$ and $n = 25$ all models perform badly and not much better than random or empty models without covariates. The CoxBVSSL model performs slightly worse than the other two approaches when $p = 20$ and $n = 50$ (Table 4.4; Table C.15 in the Appendix). Trace plots of the log-likelihood are displayed in Figure B.40 (in the Appendix).

n	p	Model	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$	$X8$	$X9$
50	20	CoxBVSSL	0.206	0.216	0.298	0.088	0.012	0.093	0.315	0.118	0.211
50	20	Subgroup	0.383	0.446	0.551	0.163	0.045	0.087	0.659	0.361	0.488
50	20	Combined	0.137	0.055	0.064	0.073	0.139	0.213	0.962	0.832	0.867
100	20	CoxBVSSL	0.999	1.000	1.000	0.150	0.062	0.121	1.000	1.000	1.000
100	20	Subgroup	1.000	1.000	1.000	0.058	0.039	0.029	1.000	0.988	0.991
100	20	Combined	0.492	0.414	0.435	0.376	0.365	0.463	1.000	1.000	1.000
25	100	CoxBVSSL	0.031	0.020	0.040	0.075	0.123	0.021	0.038	0.017	0.018
25	100	Subgroup	0.045	0.029	0.041	0.080	0.094	0.021	0.065	0.021	0.024
25	100	Combined	0.011	0.009	0.036	0.011	0.073	0.102	0.095	0.066	0.034
50	100	CoxBVSSL	0.076	0.111	0.230	0.074	0.012	0.045	0.441	0.342	0.252
50	100	Subgroup	0.070	0.096	0.206	0.118	0.009	0.034	0.322	0.252	0.201
50	100	Combined	0.028	0.023	0.007	0.050	0.058	0.142	0.822	0.705	0.603
75	100	CoxBVSSL	0.782	0.768	0.843	0.027	0.011	0.025	0.950	0.896	0.883
75	100	Subgroup	0.550	0.562	0.723	0.010	0.009	0.015	0.624	0.448	0.445
75	100	Combined	0.156	0.053	0.077	0.014	0.113	0.143	0.901	0.825	0.800
100	100	CoxBVSSL	0.997	1.000	0.998	0.061	0.019	0.040	0.982	0.965	0.965
100	100	Subgroup	0.825	0.833	0.975	0.007	0.009	0.009	0.967	0.859	0.868
100	100	Combined	0.068	0.034	0.013	0.007	0.012	0.083	0.997	0.982	0.985
150	100	CoxBVSSL	1.000	1.000	1.000	0.078	0.023	0.064	1.000	1.000	1.000
150	100	Subgroup	1.000	1.000	1.000	0.006	0.005	0.006	1.000	1.000	1.000
150	100	Combined	0.276	0.184	0.190	0.104	0.023	0.044	1.000	1.000	1.000

n	p	Model	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$	$X8$	$X9$
50	20	CoxBVSSL	0.016	0.013	0.041	0.096	0.132	0.281	0.331	0.157	0.128
50	20	Subgroup	0.064	0.073	0.084	0.362	0.332	0.430	0.559	0.440	0.444
50	20	Combined	0.137	0.055	0.064	0.073	0.139	0.213	0.962	0.832	0.867
100	20	CoxBVSSL	0.029	0.056	0.120	0.998	0.998	1.000	0.988	0.976	0.975
100	20	Subgroup	0.053	0.029	0.032	0.983	0.976	0.977	0.921	0.917	0.977
100	20	Combined	0.492	0.414	0.435	0.376	0.365	0.463	1.000	1.000	1.000
25	100	CoxBVSSL	0.014	0.012	0.017	0.011	0.024	0.081	0.068	0.120	0.020
25	100	Subgroup	0.016	0.014	0.020	0.010	0.016	0.095	0.066	0.097	0.020
25	100	Combined	0.011	0.009	0.036	0.011	0.073	0.102	0.095	0.066	0.034
50	100	CoxBVSSL	0.015	0.010	0.017	0.086	0.145	0.324	0.300	0.142	0.111
50	100	Subgroup	0.016	0.012	0.018	0.053	0.100	0.251	0.327	0.066	0.067
50	100	Combined	0.028	0.023	0.007	0.050	0.058	0.142	0.822	0.705	0.603
75	100	CoxBVSSL	0.008	0.012	0.141	0.636	0.642	0.650	0.651	0.468	0.470
75	100	Subgroup	0.007	0.006	0.104	0.510	0.551	0.639	0.515	0.355	0.376
75	100	Combined	0.156	0.053	0.077	0.014	0.113	0.143	0.901	0.825	0.800
100	100	CoxBVSSL	0.013	0.012	0.073	0.931	0.926	0.940	0.994	0.905	0.892
100	100	Subgroup	0.010	0.008	0.022	0.839	0.765	0.803	0.740	0.394	0.378
100	100	Combined	0.068	0.034	0.013	0.007	0.012	0.083	0.997	0.982	0.985
150	100	CoxBVSSL	0.006	0.023	0.058	1.000	1.000	1.000	1.000	1.000	1.000
150	100	Subgroup	0.005	0.006	0.007	1.000	1.000	0.997	1.000	1.000	0.995
150	100	Combined	0.276	0.184	0.190	0.104	0.023	0.044	1.000	1.000	1.000

TABLE 4.3: *Mean posterior inclusion frequencies (averaged over all training sets) of the prognostic variables for subgroup 1 (top) and subgroup 2 (bottom). Variables included on average are highlighted in red.*

Estimation of subgraphs $\boldsymbol{G}_{ss}$ and corresponding precision matrices $\boldsymbol{\Omega}_s$ in subgroups $s = 1, 2$ works relatively well and improves for increased sample size (Figures B.43 and B.42 in the Appendix). Mean posterior probabilities of edge inclusion in subgraph $\boldsymbol{G}_{12}$ are relatively small in all simulation settings (≤ 0.45). They become larger for increasing n. For edges relating genes 7, 8, and 9 with joint effects in both subgroups probabilities are considerably higher than for all remaining edges, except for $n = 25$ and $p = 100$ (Figure B.44 in the Appendix).

n	p	s	Combined	Subgroup	CoxBVSSL
50	20	1	0.74 (0.04)	0.76 (0.07)	0.72 (0.06)
50	20	2	0.79 (0.04)	0.77 (0.05)	0.74 (0.05)
100	20	1	0.74 (0.05)	0.84 (0.03)	0.84 (0.02)
100	20	2	0.76 (0.06)	0.84 (0.04)	0.84 (0.04)
25	100	1	0.60 (0.10)	0.56 (0.07)	0.54 (0.05)
25	100	2	0.57 (0.10)	0.59 (0.11)	0.56 (0.13)
50	100	1	0.67 (0.05)	0.65 (0.06)	0.66 (0.09)
50	100	2	0.75 (0.09)	0.67 (0.11)	0.68 (0.11)
75	100	1	0.70 (0.04)	0.75 (0.06)	0.81 (0.07)
75	100	2	0.76 (0.04)	0.77 (0.05)	0.78 (0.05)
100	100	1	0.74 (0.04)	0.82 (0.02)	0.83 (0.02)
100	100	2	0.76 (0.04)	0.80 (0.07)	0.84 (0.04)
150	100	1	0.74 (0.04)	0.84 (0.02)	0.84 (0.02)
150	100	2	0.76 (0.03)	0.86 (0.02)	0.86 (0.02)

TABLE 4.4: *Mean (standard deviation) of the C-index (computed over all test sets) for the prediction of subgroup $s = 1, 2$.*

In a second simulation study, the influence of finer increments of true effects on survival outcome is investigated. True effects of genomic covariates in both subgroups are chosen as follows:

	Gene														
	1	2	3	4	5	6	7	8	9	10	11	12	13	...	p
$\boldsymbol{\beta}_1$	1	1	0	0	-0.5	0.5	0.75	0.25	-1	-1	-0.75	-0.25	0	...	0
$\boldsymbol{\beta}_2$	0	0	1	1	0.5	-0.5	0.25	0.75	-1	-1	-0.75	-0.25	0	...	0

including subgroup-specific effects (genes 1 to 4), opposite effects (genes 5 and 6), effects in the same direction but of different size (genes 7 and 8), and common effects of varying sizes (genes 9 to 12). The first 12 genes are referred to as prognostic genes since they are associated with outcome in at least one subgroup. All remaining genes are considered to be noise. Gene expression and survival data are simulated analogously to the first simulation study as described in section 4.2.1. The only difference is that the precision matrix $\boldsymbol{\Omega}$ consists of an AR(2) correlation structure between the first 12 instead of the first 9 genes; all other genes are uncorrelated. High-dimensional settings with $p = 100$ and $n = 50, 75, 100, 200$ are studied, as well as two low-dimensional settings with $p = 20$ and $n = 50, 100$. Since computation time of the CoxBVSSL model increases drastically with the number of covariates, only up to $p = 100$ covariates are considered. As before,

ten independent training and test data sets are generated from the same distribution for each subgroup.

Convergence is assessed in terms of autocorrelation plots, trace plots and running mean plots of the regression coefficients, here only the latter two are shown representatively for the first training data set. Results of posterior means and measures of prediction performance are averaged across all data sets. Findings mainly agree with those of the first simulation study. Convergence of all three models is poor when $n = 50$ and improves with increasing sample size. Estimates of regression coefficients in the CoxBVSSL model tend to stabilize faster compared to the standard subgroup model (Figures B.45 and B.46 in the Appendix).

Results of variable selection (see Tables C.16, C.17 and C.20 in the Appendix) and posterior estimates of regression coefficients (Tables C.18 and C.19 in the Appendix) indicate that the combined model tends to have larger power to identify joint effects when the sample size is small but fails to detect subgroup-specific effects. In all settings, except for $n = 50$, the CoxBVSSL model has higher posterior probabilities of variable inclusion compared to the standard subgroup model. This applies to both prognostic and non-prognostic genes (genes 3 and 4 in subgroup 1, genes 1 and 2 in subgroup 2). It leads to a more stable selection of prognostic genes, larger power for identification and more accurate estimation of true effects, in particular weak effects. However, a potential downside may be a tendency of CoxBVSSL towards more false positives. In the present simulation study, CoxBVSSL selects only one false positive in subgroup 1 when $n = 200$. With regard to the posterior means of the estimated regression coefficients, the subgroup model underestimates the majority of effects, whereas CoxBVSSL leads to more precise estimates when $n < 200$. For $n = 200$ both models show a tendency towards overestimation of true effects, which is more pronounced in CoxBVSSL.

Prediction performance in terms of mean C-index and IBS is summarized in Table 4.5 and Table C.21 in the Appendix. All three models perform better than chance and reference models without covariates (Kaplan-Meier estimators). The combined model is competitive when $n = 50$ but inferior to the standard subgroup and CoxBVSSL model in all other situations. The subgroup and CoxBVSSL model perform similarly in low-dimensional settings and in high-dimensional settings for $n = 50, 200$. However, in all other situations CoxBVSSL has the best predictive ability.

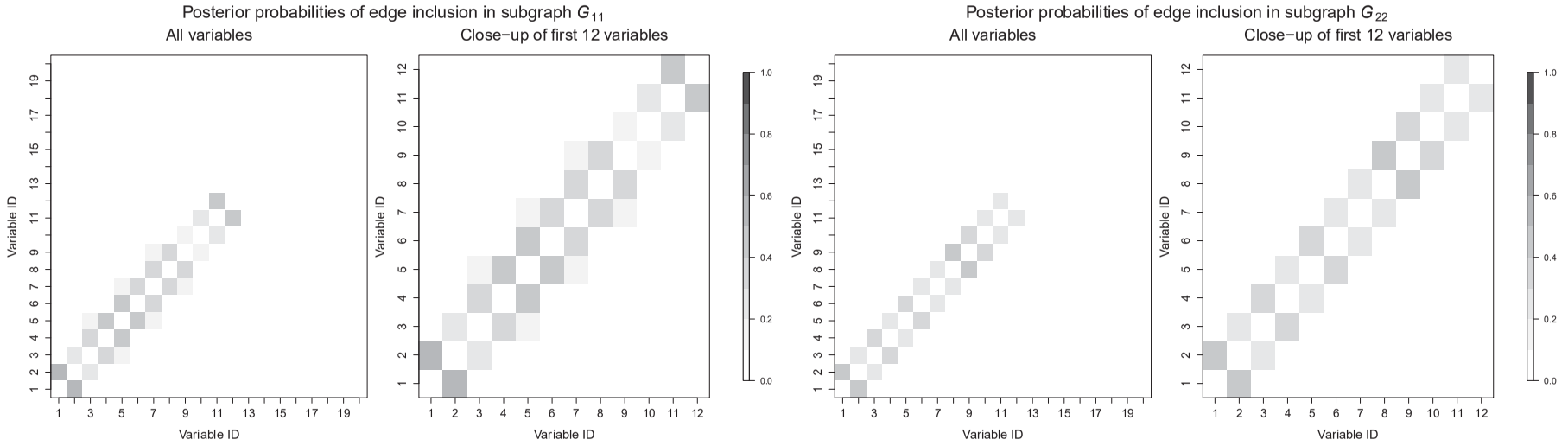
Results of the inferred network among covariates are very similar to those in the first simulation study, and therefore only shown for selected simulation settings. Mean posterior probabilities of edge inclusion in subgraphs $\boldsymbol{G}_{ss}$ and posterior means of elements in the corresponding precision matrices $\boldsymbol{\Omega}_s$ for subgroups $s = 1, 2$ indicate that estimation of the underlying dependence structure between genes within each subgroup improves with increasing sample size and works well for $n \geq 100$. Posterior probabilities of edge inclusion become larger, leading to a more stable identification of true edges between the prognostic genes (see Figure 4.19). Posterior probabilities of edge inclusion in subgraph $\boldsymbol{G}_{12}$, relating the same genes across both subgroups, are smaller than in subgraphs $\boldsymbol{G}_{ss}$. However, as sample size grows it becomes increasingly apparent that edges between genes 5, 6, 9, 10, and 11 are most frequently selected, followed by genes 7 and 8. Thus, edge selection in $\boldsymbol{G}_{12}$ indicates common prognostic genes in both subgroups (not necessarily with effects in the same direction) and depends on the effect size, with larger effects leading to more stable edge selection (Figure 4.20).

n	p	s	Combined	Subgroup	CoxBVSSL
50	20	1	0.66 (0.07)	0.70 (0.07)	0.67 (0.07)
50	20	2	0.73 (0.08)	0.78 (0.05)	0.75 (0.07)
100	20	1	0.71 (0.03)	0.83 (0.03)	0.84 (0.03)
100	20	2	0.78 (0.04)	0.87 (0.03)	0.88 (0.03)
50	100	1	0.66 (0.07)	0.64 (0.09)	0.63 (0.08)
50	100	2	0.69 (0.11)	0.71 (0.10)	0.73 (0.09)
75	100	1	0.67 (0.04)	0.75 (0.06)	0.77 (0.04)
75	100	2	0.76 (0.05)	0.78 (0.06)	0.79 (0.06)
100	100	1	0.71 (0.06)	0.79 (0.04)	0.82 (0.03)
100	100	2	0.74 (0.03)	0.81 (0.05)	0.84 (0.04)
200	100	1	0.72 (0.03)	0.84 (0.02)	0.84 (0.02)
200	100	2	0.77 (0.03)	0.88 (0.02)	0.88 (0.02)

TABLE 4.5: *Mean (standard deviation) of the C-index (computed over all test sets) for the prediction of subgroup* $s = 1, 2$.

In summary, MCMC mixing and convergence of all models is poor when $n \leq 50$, but improves rapidly with increasing sample size. The combined model only identifies joint effects but fails to detect subgroup-specific effects, and thus is inferior to the other models. When $n > p$ the subgroup and CoxBVSSL model perform similarly well with regard to selection and prediction accuracy, except for $p = 20$, $n = 50$ where the subgroup model performs better. However, when $n \leq p$ the CoxBVSSL model has superior predictive ability and larger power to identify relevant genes compared to standard approaches. This suggests that incorporating network information into variable selection can improve detection of true effects. Accuracy of graph structure learning for the proposed model improves for increasing sample size and is quite high for $n \geq 100$.

(A) $n = 50$ and $p = 20$



(B) $n = 100$ and $p = 20$

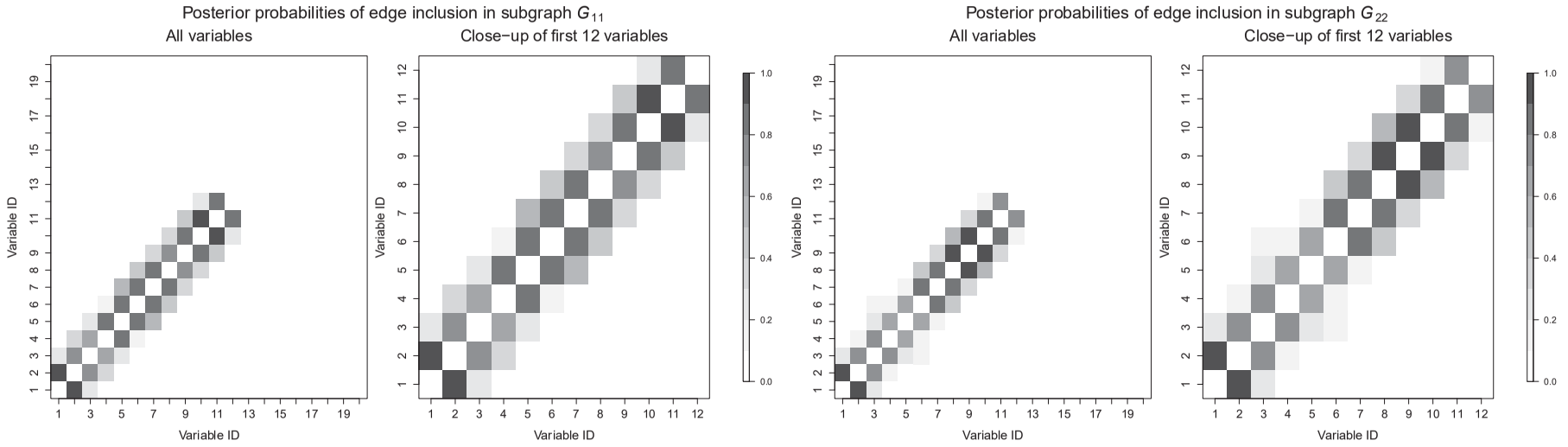


FIGURE 4.19: *Mean posterior probabilities of edge inclusion (average across all simulations) in subgraph $\boldsymbol{G}_{ss}$ for subgroup $s = 1$ (left) and $s = 2$ (right). Edges in the graph are assumed independent Bernoulli a priori with parameter $\pi = 2/(p-1)$.*

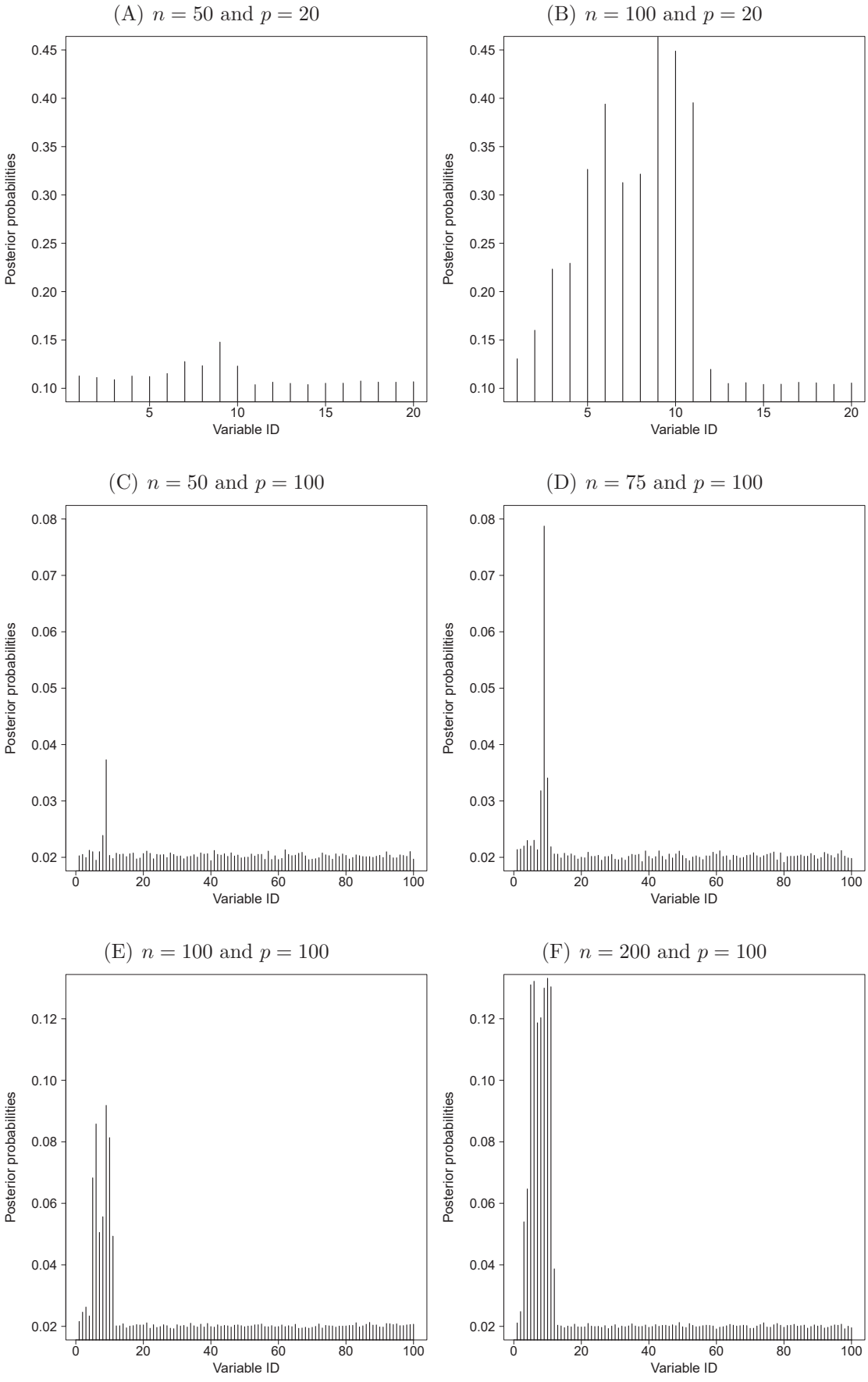


FIGURE 4.20: *Posterior probabilities of edge inclusion for diagonal elements in subgraph* $\boldsymbol{G}_{12}$ *(average across all simulations). Edges in the graph are assumed independent Bernoulli a priori with parameter* $\pi = 2/(p-1)$.

4.2.5 Application to lung cancer studies

For the application example the standard combined, subgroup and CoxBVSSL model are applied to four lung cancer studies. Only genomic covariates are included using two different gene filters: 30 Kratz genes and top-100-variance genes. The former comprise 30 probe sets belonging to 13 prognostic genes from Kratz et al. (2012) with matches on the Affymetrix HG-U133 Plus 2.0 array. The top-100-variance genes are defined by the 100 probe sets with the highest variability in gene expression values across all four studies. The complete data are randomly divided into a training and a test data set, drawing 63.2% observations without replacement and stratified according to study and censoring indicator for training and using all remaining observations as test data for validation. This procedure is repeated ten times similarly as before in simulations.

Convergence is assessed in terms of trace plots, running mean plots and autocorrelation plots of regression coefficients. Results of the subgroup and CoxBVSSL model are relatively similar and mostly the same variables are selected regardless of the gene filter. In most situations, autocorrelation values decrease rapidly and suggest good convergence. However, for some selected genes in the corresponding subgroups autocorrelations, trace plots and running means indicate slow convergence. Trace plots and running mean plots of the 30 prognostic Kratz genes are shown representatively for the first training data set in Figures B.47 and B.48 in the Appendix. Cox models are first compared with regard to posterior variable selection probabilities and posterior estimates of regression coefficients, and afterward in terms of prediction performance. Results are averaged across all training sets unless stated otherwise.

Similar variable selection results are obtained for the CoxBVSSL and subgroup model. From the 30 prognostic Kratz genes, the following genes are included in both models: BRCA1 in subgroups GSE37745 and GSE50081, and two different probe sets of CDC6, one each in GSE29013 and GSE31210, respectively. These two genes are also selected by the combined model of all subgroups. The subgroup model additionally identifies gene ERBB3 in GSE29013 and gene IL11 in GSE31210, whereas the CoxBVSSL model selects one further probe set of BRCA1 in GSE37745. Estimated regression coefficients of all selected genes are positive, except for IL11 and one probe set of BRCA1 (see Figures 4.21 and 4.22; Figure B.49 in the Appendix).

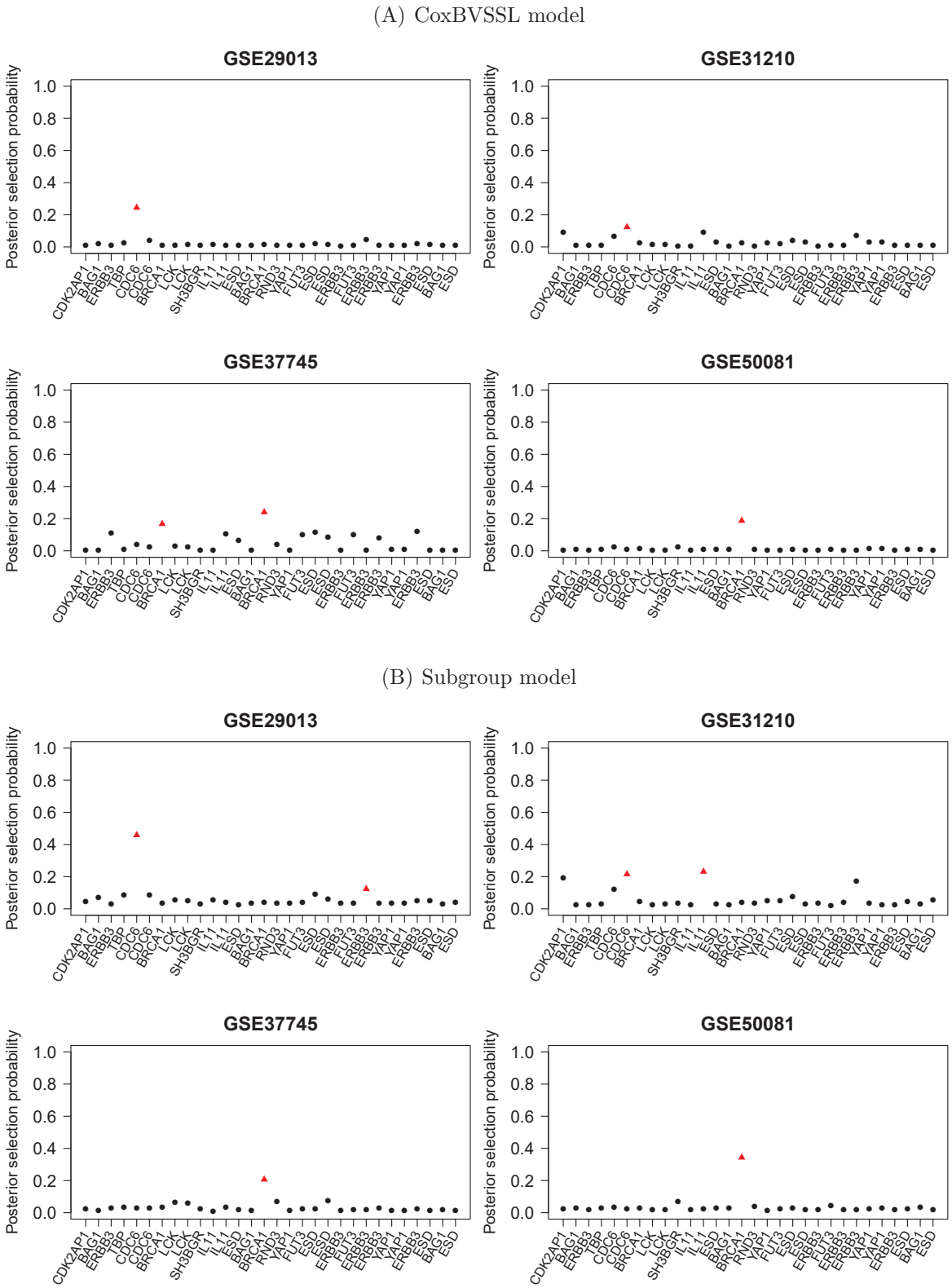


FIGURE 4.21: *Mean posterior inclusion probabilities of the 30 Kratz genes (average across all training data sets) for (A) the CoxBVSSL model, and (B) the subgroup model. Selected variables are highlighted as red triangles.*

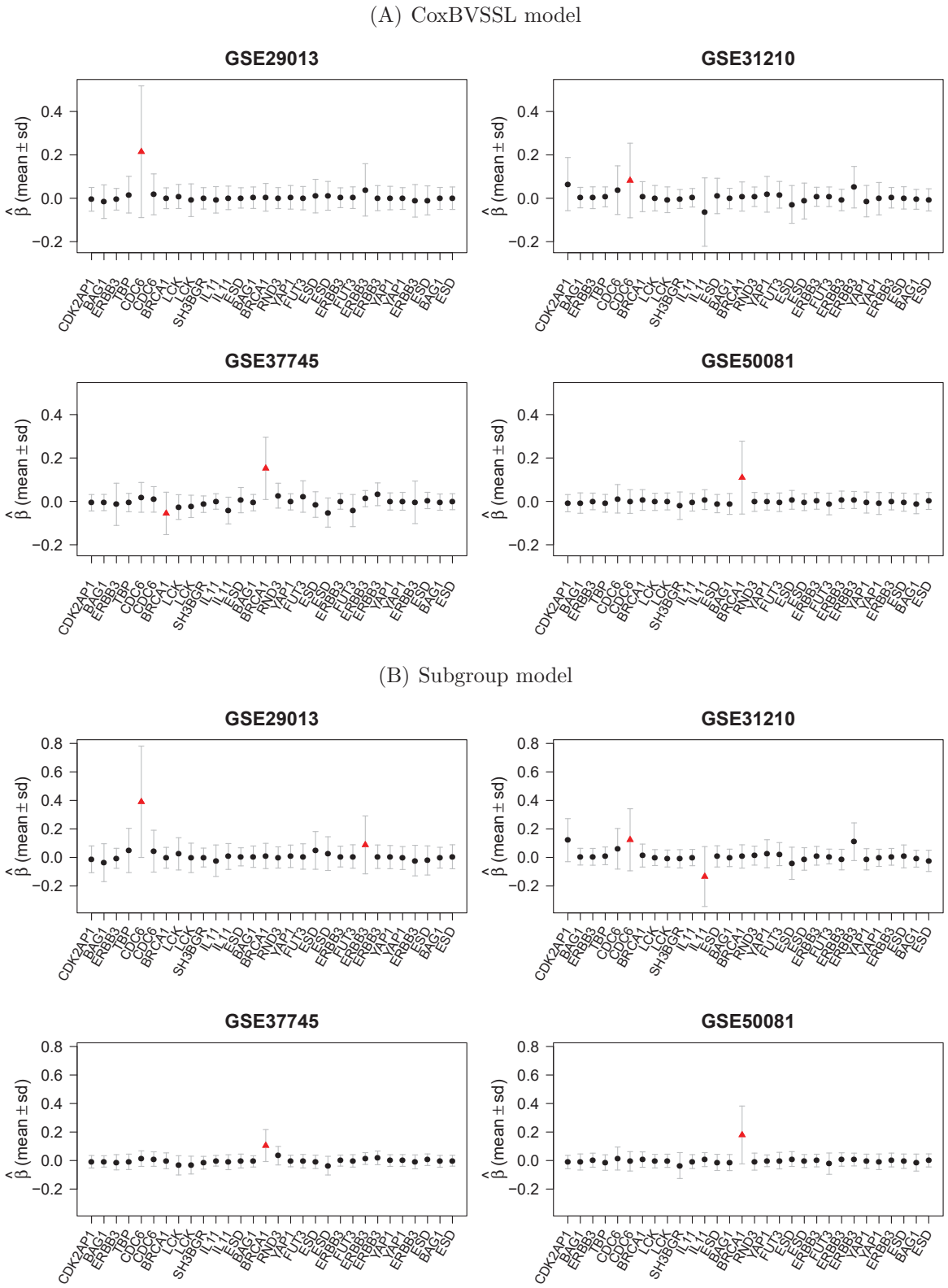
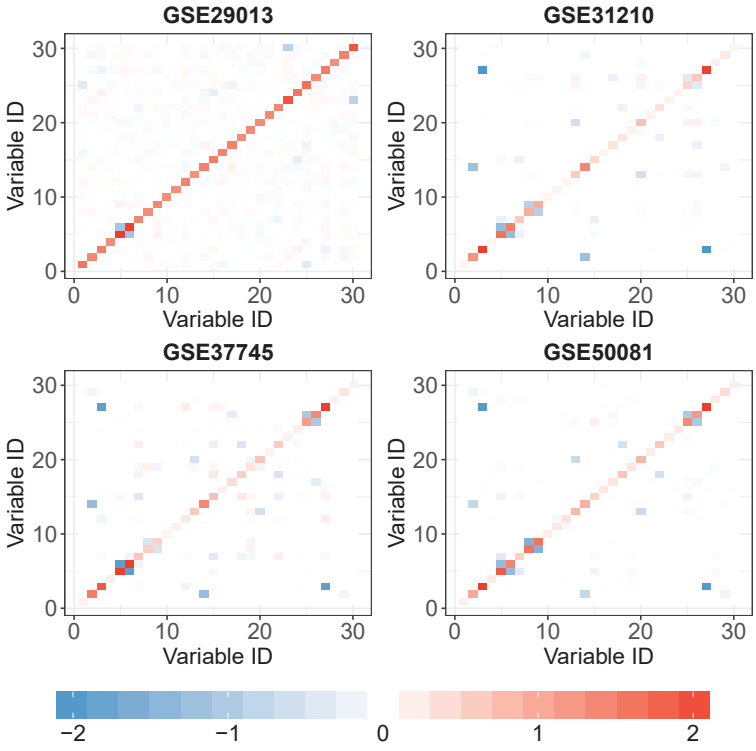


FIGURE 4.22: *Posterior mean and standard deviation (sd) of the regression coefficients of the 30 Kratz genes (average across all training data sets) for (A) the CoxBVSSL model, and (B) the subgroup model. Selected variables are highlighted as red triangles.*

Posterior probabilities of edge inclusion in subgraphs $\boldsymbol{G}_{ss}$ and posterior means of elements in the corresponding precision matrices $\boldsymbol{\Omega}_s$ for all subgroups s are displayed in Figure 4.23. They indicate a similar conditional dependence structure among the 30 Kratz genes in subgroups GSE31210, GSE37745 and GSE50081. High edge inclusion probabilities accompanied by relatively large negative mean values of the precision matrix are present between two different probe sets belonging to the same gene, such as variable IDs 5 and 6 (representing gene CDC6), IDs 8 and 9 (gene LCK), IDs 25 and 26 (gene YAP1), IDs 3 and 27 (gene ERBB3), as well as IDs 2 and 14 (gene BAG1). Since the precision matrix $\boldsymbol{\Omega} = (\omega_{ij})$ is the inverse of the covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$, its entries ω_{ij} are negatively proportional to the partial correlations $\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$. Under the assumption of Gaussianity, the partial correlation ρ_{ij} is zero if and only if variables i and j are conditionally independent given all remaining variables. A large absolute partial correlation value close to one indicates a strong positive or negative association between the two variables, with the effect of all remaining variables removed. Thus, negative values of the precision matrix correspond to positive partial correlations, which is exactly what we would expect for different probe sets belonging to the same gene. Posterior probabilities of edge inclusion in subgraphs $\boldsymbol{G}_{rs,r<s}$ linking subgroups r and s are much smaller compared to subgraphs $\boldsymbol{G}_{ss}$ (< 0.1). Genes that are jointly selected in two different subgroups have the largest edge inclusion probabilities in the corresponding subgraph. This refers to gene CDC6 in subgroups GSE29013 and GSE31210, as well as gene BRCA1 in subgroups GSE37745 and GSE50081 (see Figure 4.24).

(A) Posterior mean of elements in precision matrix $\boldsymbol{\Omega}_s$



(B) Mean posterior probabilities of edge inclusion in subgraph $\boldsymbol{G}_{ss}$

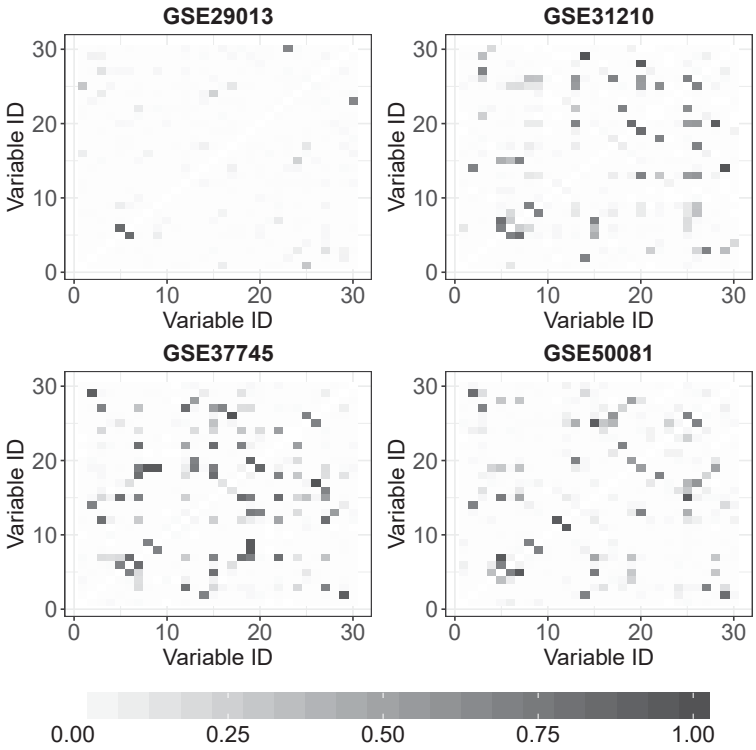


FIGURE 4.23: *Mean posterior estimates of precision matrix $\boldsymbol{\Omega}_s$ and subgraph $\boldsymbol{G}_{ss}$ for all subgroups s and the 30 Kratz genes (average across all training data sets). Edges in the graph are assumed independent Bernoulli a priori with parameter $\pi = 2/(p-1)$, $p = 30$. The prior of the diagonal entries of the precision matrix is exponential with parameter $\frac{1}{2}$, and the prior of the off-diagonal entries is a mixture of two normal distributions with zero mean and variance $\nu_0^2 = 0.1^2$ for non-selected edges and variance $\nu_1^2 = 5^2$ for selected edges.*

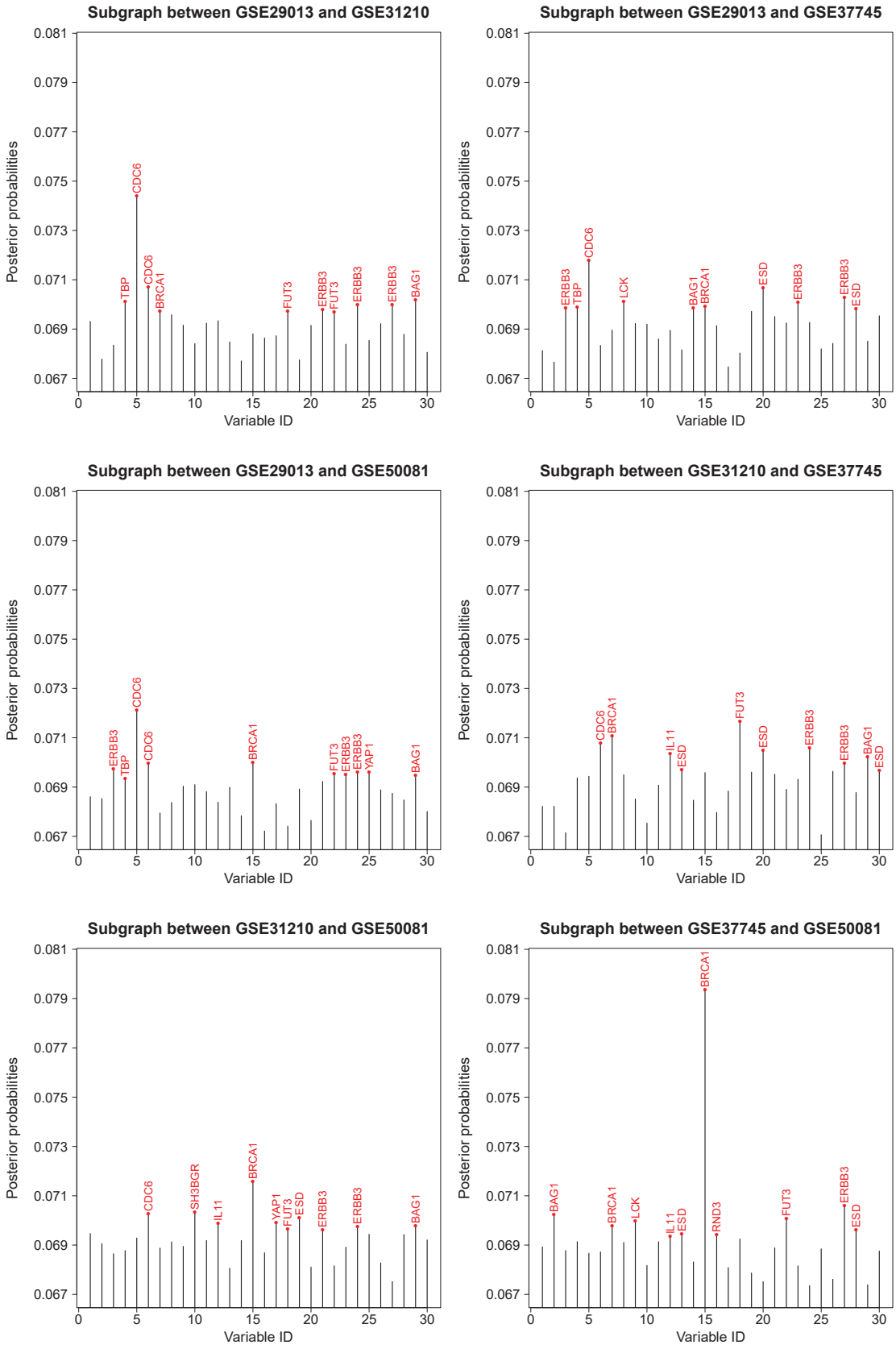


FIGURE 4.24: *Posterior probabilities of edge inclusion (PPI) for diagonal elements in subgraphs $\boldsymbol{G}_{rs,r<s}$ between subgroups r and s for the 30 Kratz genes (average across all training data sets). Edges in the graph are assumed independent Bernoulli a priori with parameter $\pi = 2/(p-1)$, $p = 30$. The ten genes with highest PPI are highlighted in red.*

Posterior probabilities of variable inclusion and posterior estimates of regression coefficients of the top-100-variance genes are shown in Figures B.50, B.51 and B.52 in the Appendix. For each subgroup, the genes selected on average in any of the three models are presented in Figure 4.25 and Table C.22 (Appendix). In the following, gene-specific information on encoded proteins, related pathways, associated diseases and related articles in PubMed is retrieved from the NCBI Gene (Brown et al., 2015) and GeneCards (*GeneCards®: The Human Gene Database*) databases. The subgroup and CoxBVSSL model including the top-100-variance genes provide different sets of selected genes for each subgroup, without any overlapping genes between subgroups. Both models identify two immunoglobulins related to immune response (IGHM and IGKC) that have been shown to be involved in the recognition and elimination of precancerous and cancerous lesions and are associated with better prognosis. Furthermore, two cancer-related genes (MMP12 and SCGB3A2) are detected by the subgroup and CoxBVSSL model. MMP12 is involved in the degradation of extracellular matrix in normal as well as in disease processes such as metastasis. Mutations are related to lung function and chronic obstructive pulmonary disease. SCGB3A2 encodes a secreted lung surfactant protein and is highly expressed in lung and trachea. Diseases associated with SCGB3A2 include Asthma and lung cancer. Additional cancer-related genes are identified by either the subgroup model only (215125_s_at/UGT1A and NTS) or by the CoxBVSSL model (XIST and SCGB3A1). NTS is distributed in central nervous and digestive systems and has been reported to promote tumor metastasis. SCGB3A1 is highly expressed in lung and encodes a cytokine-like protein that regulates cell proliferation and inhibits cell growth. XIST is a long non-coding RNA associated with tumorigenesis of different cancers. The subgroup and CoxBVSSL model provide very similar posterior estimates of regression coefficients and variable inclusion indicators for all selected genes, with the exception of XIST. Probe sets of the XIST gene have mean posterior inclusion probabilities around 0.4 in CoxBVSSL and 0 in the subgroup model. However, corresponding posterior mean values of regression coefficients are close to 0 in both models and variances are large in CoxBVSSL. The combined model only identifies the immune-related gene IGKC and the ribosomal protein encoding gene 200869_at/RPL18A that is involved in viral replication and associated with viral diseases such as Hepatitis C but not with cancer.

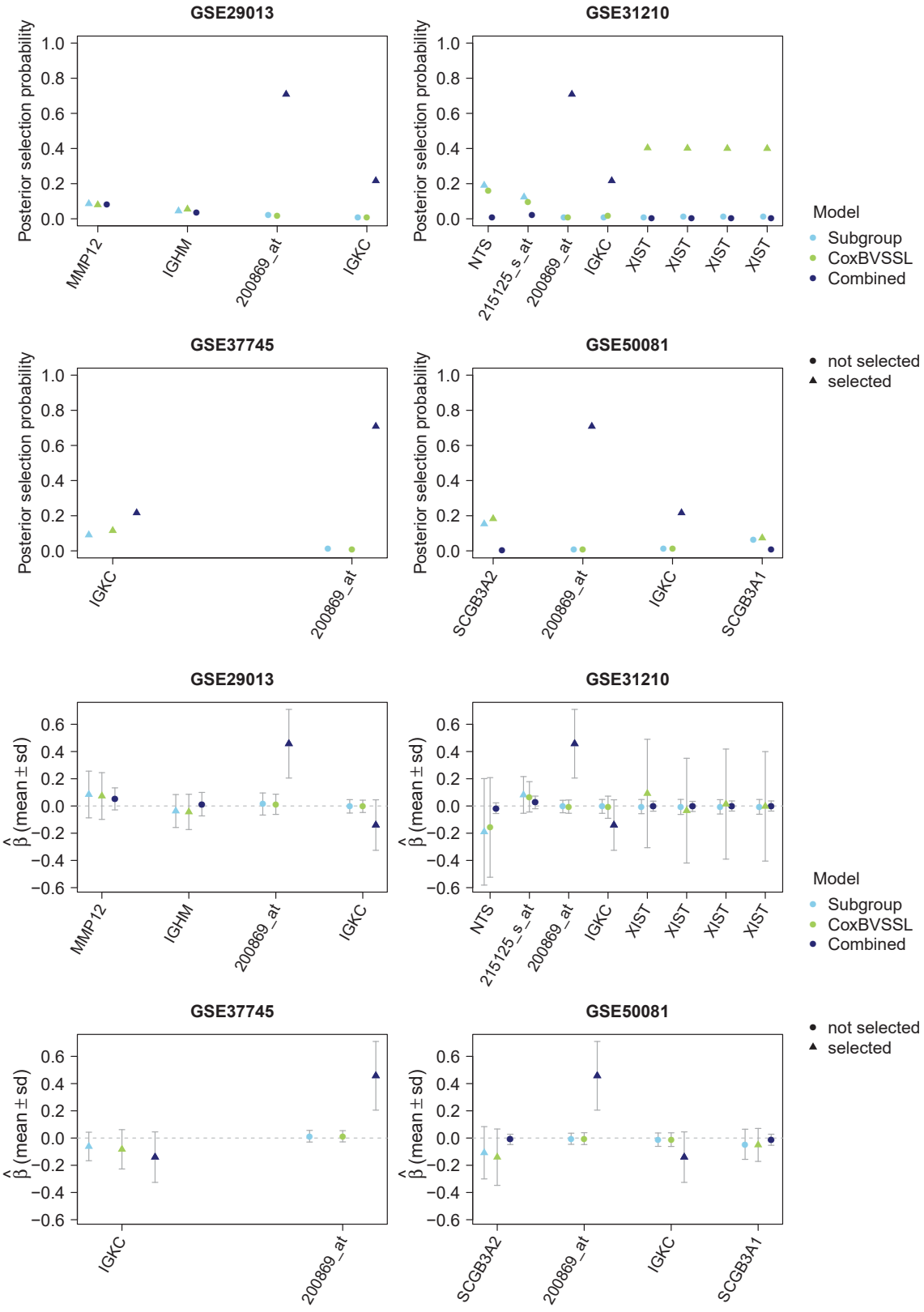


FIGURE 4.25: *Mean posterior inclusion probabilities (top) and posterior mean/standard deviation (bottom) (average across all training data sets) of the regression coefficients of the top-100-variance genes selected in any model.*

The inferred networks and conditional dependencies among the top-100-variance genes within each subgroup are displayed in Figure B.53 in the Appendix. Interdependence structures look quite similar, in particular for GSE31210, GSE37745 and GSE50081. Therefore, all pairs of genes with absolute posterior mean values in the precision matrix larger than 10 in all three subgroups are extracted (posterior means of elements in precision matrix averaged across all training data sets). The majority are pairs of two different probe sets belonging to the same gene (XIST, SFTPC, TOX3, SFTPB, CLCA2, ADH1B, KRT6A and MAGEA6), with negative precision matrix entries corresponding to positive partial correlations. This is not surprising since probe sets of the same gene can be expected to be highly correlated. Apart from that, two pairs of genes are identified encoding for related proteins associated with immune response. HLA-DQA1 and HLA-DQB1 encode for membrane proteins and are related to the G-protein signaling N-RAS regulation pathway and Cytokine Signaling in Immune system pathway. 215176_x_at/IGKV1-39 and 217378_x_at/IGKV1OR2-108 are immunoglobulins and paralogs that participate in the antigen recognition. Different pairs of genes are obtained when choosing all edges in $\boldsymbol{G}_{ss}$ with posterior inclusion probabilities (averaged across all training data sets) larger than 0.5 in all three subgroups. Among these pairs of genes are two different probe sets of the XIST gene, one pair of immunoglobulins both related to immune response (211644_x_at/IGKV3-20 and IGHM), and three pairs of unrelated genes (AKR1B10 and 215125_s_at/UGT1A, SFTA2 and FOLR1, AKR1B10 and GPX2). The protein encoded by FOLR1 is a member of the folate receptor family and involved in folic acid binding and transmembrane transport. It is required for normal cell proliferation. SFTA2 encodes a surfactant associated protein predominantly expressed in lung. No information on related pathways and interacting proteins are available. Proteins encoded by AKR1B10 and GPX2 are not related in the same network but both involved in the "SuperPathway" Metabolism. Three of the above-mentioned genes (XIST, GPX2, IGHM) are also selected in the Cox models.

Results of the inferred graphs $\boldsymbol{G}_{rs,r<s}$ between different subgroups r and s are shown in Figure B.54 in the Appendix. Posterior inclusion probabilities (PPIs) are very small and, as expected from simulations, smaller than in the low-dimensional setting with 30 Kratz genes. Among the ten genes with highest PPIs in each subgroup are some genes selected in the Cox models (XIST, MMP12, SCGB3A1, IGHM) and most of the above described genes linked in networks within subgroups (XIST, TOX3, HLA-DQA1, HLA-DQB1, SFTPB, 215125_s_at/UGT1A, ADH1B, 217378_x_at/IGKV1OR2-108, 215176_x_at/IGKV1-39, SFTPC, AKR1B10, MAGEA6, IGHM, CLCA2). However, there are no striking genes as it is the case for the 30 Kratz genes (Figure 4.24 in the Appendix), which may be due to the fact that Cox models of all subgroups have no overlapping jointly selected genes.

Prediction performance is assessed in terms of mean C-index and mean integrated Brier score (IBS) averaged across all test data sets (Table 4.6 and Table C.23 in the Appendix), as well as by prediction error curves exemplary for the first test data set (Figure B.55 in the Appendix). Prediction accuracy of all three models is very similar regardless of the gene filter. All models perform not much better than random prediction (C-index=0.5) and not better than reference models without covariates (Kaplan-Meier estimators).

Gene filter	s	Combined	Subgroup	CoxBVSSL
30 Kratz	GSE29013	0.64 (0.09)	0.67 (0.12)	0.68 (0.09)
30 Kratz	GSE31210	0.64 (0.06)	0.65 (0.05)	0.65 (0.04)
30 Kratz	GSE37745	0.55 (0.05)	0.51 (0.05)	0.52 (0.05)
30 Kratz	GSE50081	0.57 (0.05)	0.54 (0.03)	0.55 (0.03)
Top-100-v.	GSE29013	0.63 (0.09)	0.50 (0.09)	0.50 (0.08)
Top-100-v.	GSE31210	0.60 (0.08)	0.67 (0.06)	0.65 (0.07)
Top-100-v.	GSE37745	0.54 (0.04)	0.51 (0.03)	0.51 (0.03)
Top-100-v.	GSE50081	0.53 (0.05)	0.55 (0.05)	0.55 (0.05)

TABLE 4.6: *Mean (standard deviation) of C-index (computed over all test sets) for the prediction of subgroup s.*

Chapter 5

Summary and Discussion

This thesis focuses on three major objectives: the prediction of a patient's survival function, selection of important covariates, and consideration of heterogeneity in data due to known subgroups of patients. Specifically, we aim at providing a separate prediction model for each subgroup that allows the identification of common as well as subgroup-specific effects and has improved prediction accuracy over standard approaches. The latter refer to standard subgroup analysis, including only patients of the subgroup of interest, and standard combined analysis that pools patients of all subgroups. Small sample size is a problem of standard subgroup analysis potentially leading to unstable results and reduced power to detect prognostic effects, while standard combined analysis may suffer from biased results and averaging of subgroup-specific effects. Therefore, we seek an alternative approach that allows sharing information between subgroups to increase power when this is supported by data, meaning that subgroups are similar in their relationship between covariates and survival outcome. To accomplish this, a novel classical frequentist and a novel Bayesian Cox proportional hazards model are proposed.

The frequentist model uses a lasso penalty for variable selection and a weighted version of the Cox partial likelihood that includes patients of all subgroups but assigns them individual weights based on their subgroup affiliation. Weights for a specific subgroup are estimated from the training data by classification and cross-validation such that they represent the probability of belonging to that subgroup given the observed covariates and survival outcome. These predicted conditional probabilities are divided by the a priori probability of the respective subgroup to obtain the subgroup-specific weights for each patient. Patients who fit well into the subgroup of interest receive higher weights in the subgroup-specific model.

The proposed Bayesian Cox model uses a stochastic search variable selection prior with latent indicators of variable inclusion. We assume a sparse graphical model that links covariates within subgroups and the same covariates across different subgroups. This graph structure is not known a priori and inferred simultaneously with the important variables of each subgroup. It favors the selection of related covariates in the graph and represents conditional dependencies among covariates within subgroups and joint prognostic covariates shared by different subgroups. Thus, the proposed model allows identification of predictors that are both relevant to survival outcome and linked to each other in a conditional dependence network. Both approaches are evaluated through simulations and applied to four lung cancer studies. Main findings, discussion of limitations and outlook for further analyses are summarized in the following, beginning with the frequentist approach.

In simulation studies in the frequentist setting, we considered a varying number of genomic covariates and sample size with a focus on high-dimensional settings where sample size is small compared to the number of variables, a typical characteristic of gene expression data. Subgroups differed in their simulated survival times, in particular different true effects, and partly in their covariate values. Genomic covariates were simulated considering various correlation structures and increasing differences between mean values of prognostic genes in distinct subgroups. The latter served to control the degree of similarity between subgroups. We found that sample size and dissimilarity between subgroups had the strongest impact on the proposed weighted model, with larger values greatly improving the performance of weights estimation and Cox model fitting. In contrast, there were no recognizable differences for varying correlation structures.

We considered three different classification methods (multinomial logistic regression with lasso or ridge penalty and random forest) for weights estimation in combination with inclusion or exclusion of interactions between genomic covariates and survival time, as well as replacement of the survival time by the Nelson–Aalen estimator of the cumulative hazard rate (HR) in the set of covariates. The latter was proposed by White and Royston (2009) in the context of multiple imputation. We compared these parameters for weights estimation with regard to prediction performance and found that inclusion of interactions and cumulative HR in the classification model only improved prediction accuracy of random forest when all subgroups were similar. In all other situations predictive quality of these parameters hardly differed with a tendency towards better results for multinomial logistic regression without cumulative HR. Random forest outperformed multinomial logistic regression when all subgroups were similar or when sample size was low.

In a small simulation study, we considered the case of unbalanced subgroup sizes with one small subgroup and three equally large subgroups. We compared standard classification without sampling techniques with two oversampling techniques (random oversampling and synthetic minority oversampling technique). Oversampling increases sample size of the small subgroup so that it is balanced with respect to the other subgroups. As expected, the predicted probabilities estimated by classification were always smaller for the minority subgroup compared to the other subgroups. However, this effect was compensated for when predicted probabilities were divided by the relative frequencies of each subgroup to obtain the estimated weights. Thus, estimated weights did not differ due to unbalanced sample size and oversampling techniques had no effect on classification performance in the present simulations.

The proposed weighted Cox model was compared to a standard combined and subgroup Cox model, as well as a weighted Cox model with different fixed weights as proposed by Weyer and Binder (2015). Observations belonging to a certain subgroup were assigned a weight of 1 in the subgroup-specific likelihood, while all other observations were down-weighted with a constant weight $w \in \{0.1, 0.2, ..., 0.9\}$. When subgroups were hardly distinguishable from each other with respect to their covariate values and differed mainly in their relationship between prognostic covariates and survival outcome, classification methods failed to discriminate between distinct subgroups and all observations were assigned a weight around one similarly to the standard combined model. In these situations, results of the combined model and the proposed weighted model were similar. Both models had better prediction performance and larger power

to detect joint effects than the standard subgroup model when sample size was small ($n \leq p$). However, they tended to average subgroup-specific effects resulting in biased estimates. For increasing sample size, the standard subgroup model outperformed the other models regarding prediction and selection accuracy, in particular in terms of correct estimation of subgroup-specific effects. When differences between subgroups became larger, classification succeeded in discriminating between different subgroups and the proposed weighted model improved over the combined model in identifying subgroup-specific effects and in prediction accuracy. It was competitive with the subgroup model and even better for small sample sizes. In the case of unbalanced subgroup sizes, the prediction performance of the subgroup model was much worse regarding the small subgroup, whereas results of the standard combined model and weighted model remained almost unchanged. Results of the weighted Cox model with fixed weights lay between the standard subgroup model and the combined model depending on the weight size. Fixed weights tended to perform better than estimated weights when subgroups were similar and classification performance was bad.

In one simulation study, we compared the different types of Cox models with lasso penalty to componentwise likelihood-based boosting for Cox models. The latter is implemented in the R package `CoxBoost` and was used by Weyer and Binder (2015) in weighted Cox regression with fixed weights. Our findings suggested that both algorithms had similar prediction performance, except for situations with large differences between subgroups. Here the Cox model with lasso penalty outperformed the boosting algorithm when including weights (particularly estimated weights).

In the application example, we considered multiple lung cancer studies as subgroups comprising overall survival outcome, and gene expression data and clinical information (age, sex, pTNM stage, histology, and smoking status) as covariates. One further objective was to examine the additional predictive value of genomic covariates over established clinical covariates. To accomplish this, all frequentist Cox models were fitted including only genomic covariates, a combination of genomic and clinical covariates, and only clinical covariates. Three different gene filters were used: all available genes, top-1000-variance genes, and a literature-based selection of prognostic genes.

Different classification methods, parameters for weights estimation (interactions, cumulative HR) and oversampling techniques provided very similar prediction performance and no clear distinctions could be determined. Estimated weights based on genomic covariates suggested large differences between all subgroups. Observations belonging to the subgroup of interest received a high weight in the subgroup-specific model, while observations of all remaining subgroups obtained weights close to zero and hardly contributed to the subgroup-specific prediction model. Subgroups appeared to be more similar when weights were estimated based on clinical covariates only.

Prediction performance of all Cox models was mainly moderate and not much better than random prediction or reference models without any covariates. The combined model and the weighted model with fixed weights of increasing size showed the highest predictive accuracy when genomic covariates were included. The proposed weighted model and the subgroup model performed similarly bad. When only clinical covariates were used all weighted models and the combined model had similar performance and were better than the subgroup model. Fixed weights had in most cases the highest variable selection stability, followed by the combined model. These models mainly identified genes with joint effects in all subgroups whereof some are known to be associated with

prognosis in various cancers. However, corresponding estimated regression coefficients were often relatively small suggesting weak effects on survival outcome. Few candidate genes with reported cancer relation and relatively strong subgroup-specific effects were selected most frequently by either the subgroup model or the proposed weighted model. The additional predictive value of genomic covariates over clinical covariates remains unclear.

Weyer and Binder (2015) propose a weighted and stratified Cox regression model with separate baseline hazard rates in different subgroups (strata) to account for heterogeneity in data. Stratification supports the assumption that hazard rates of distinct subgroups may not be proportional to one another which contradicts the main assumption of the Cox proportional hazards model. In our proposed frequentist model the extent to which each observation contributes to parameter estimation and variable selection in the subgroup-specific model is controlled by individual subgroup-specific weights. However, the same baseline hazard rate across all subgroups is assumed. Our weighted version of the Cox partial likelihood could be extended to allow for varying baseline hazard rates in different subgroups.

The approach by Weyer and Binder (2015) gives observations in the subgroup of interest a weight of 1, while all other observations are assigned a constant weight in $(0, 1)$. Alternatively to our proposed approach with estimated individual weights, weights can be considered as a tuning parameter and optimized by model-based optimization (MBO) to improve prediction performance. This approach is more flexible than the one by Weyer and Binder (2015) since it allows different fixed weights for different subgroups in each subgroup model. MBO helps to identify the best combination of fixed weights with regard to prediction accuracy. Richter, Madjar, and Rahnenführer (2018) introduce MBO of subgroup weights in the Cox model and evaluate this approach on various lung cancer studies.

In the Bayesian setting, we also examined a varying number of genomic covariates, sample size and true effects on survival outcome through simulation studies, with a focus on small sample sizes. The proposed model with Markov random field prior for the variable selection indicators was compared to a standard subgroup and a combined Bayesian Cox model with Bernoulli prior for the variable selection indicators following Treppmann, Ickstadt, and Zucknick (2017). Through simulations, we have demonstrated that our approach can achieve improved variable selection and prediction accuracy over competing standard approaches. The combined model was inferior to the other models in that it could only identify joint effects and failed to detect subgroup-specific effects. The subgroup model and the proposed model performed similarly well when sample size was large. However, when sample size was small compared to the number of covariates ($n \leq p$), the proposed model outperformed standard Bayesian variable selection in terms of both selection and prediction accuracy. This suggests that incorporating network information into variable selection can increase power to identify true associations between covariates and survival outcome. Inference of the graph showed relatively high accuracy for learning the conditional dependence structure among genes within subgroups and for detecting joint effects across different subgroups.

Bayesian approaches were further validated in application to lung cancer studies including only genomic covariates with two different gene filters: 30 prognostic genes from literature, and top-100-variance genes. The standard subgroup model and the proposed model performed very similar in terms of variable selection and prediction

accuracy. Prediction performance of all models was relatively bad as for the frequentist models.

The main reason for the overall moderate prediction accuracy in the application example may be that the present lung cancer studies are heterogeneous. On the one hand, they comprise different histological subtypes that are known to be associated with a different prognosis. One could think of using only patients belonging to the same histological subtype such as adenocarcinoma. On the other hand, tissue processing and RNA extraction for generating gene expression data as well as patient inclusion criteria vary between studies. In GSE29013 genome-wide expression profiling was based on formalin-fixed paraffin-embedded (FFPE) tissues rather than fresh frozen tissues like in GSE37745 and GSE50081, which might influence expression levels. GSE31210 and GSE50081 include only patients with stage I and II, and GSE31210 is additionally restricted to lung adenocarcinomas. Further confounding variables in gene expression data can be variation in array data generation and reporting between different laboratories. Recommendations for experimental design and data generation of Microarray experiments are discussed in The Tumor Analysis Best Practices Working Group (2004). The proposed models should be validated on further cancer studies, using for example data of our collection of breast cancer, ovarian cancer or colon cancer studies.

Both proposed models have demonstrated in simulations that they can achieve improved prediction and variable selection accuracy over standard subgroup models when the sample size is low. However, one drawback of the frequentist weighted model is that it may average subgroup-specific effects resulting in biased estimates, in particular when discrimination between differing subgroups is challenging. Advantages of the frequentist Cox model, implemented in the R package `glmnet`, over the Bayesian model are the much shorter computation time and that it is well suited to include all kinds of covariates (categorical or continuous) and tens of thousands of covariates which is characteristic of genomic data. Furthermore, it can accommodate not only time-to-event endpoints but also continuous, binary and multinomial outcomes and thus, is very flexible with respect to covariates and response.

The code for the MCMC samplers of the Bayesian Cox models was implemented entirely in R, in contrast to the R packages used for the frequentist models, and can be optimized for computational speed. Due to relatively slow computation, only up to 100 preselected covariates have been considered so far in the proposed Bayesian model. The analysis of many thousands of genes is not yet feasible, but could be enabled by a computationally more efficient implementation. The current implementation of the Bayesian models allows exclusively for survival outcome and would need some adaptions to accommodate other outcomes. Our proposed Bayesian model relies on a Gaussian graphical model that assumes multivariate normal covariates being suitable for our application to gene expression data. However, for other types of covariates, this assumption can be violated. In further simulations, it might be interesting to examine how much deviation from Gaussianity is acceptable, for example by drawing covariates from a multivariate t-distribution.

Advantages of Bayesian approaches in general are the modeling of uncertainty, the possibility to incorporate prior information, for example derived from literature, historical data or additionally available data sources, and the more flexible modeling of complex data and dependencies such as heterogeneous subgroups. One specific advantage

of the proposed Bayesian Cox model is that it allows accounting for uncertainty over both variable and graph selection. Prior knowledge of the network among the covariates is not required. Instead, inference of a sparse graph is performed that reveals relationships among the covariates. This allows the identification of relevant genes and pathways, which in turn may lead to better understanding of molecular mechanisms and tumorigenesis, and may improve targeted therapies. In situations where pathway information is available and the network structure is known, it may be desirable to incorporate this structural information in the Markov random field prior for variable selection via a fixed graph.

In the frequentist implementation, there is the possibility to include mandatory clinical covariates in an unpenalized manner. This is not yet implemented in the present Bayesian models where a variable selection prior is imposed on all covariates simultaneously. Bayesian Cox models can be extended to have separate prior distributions for mandatory clinical covariates and penalized genomic covariates. For mandatory covariates, a weakly informative normal prior can be used as in Zucknick, Saadati, and Benner (2015), whereas for penalized covariates a spike-and-slab prior as in this thesis can be applied or a shrinkage Laplace prior (corresponding to the frequentist lasso penalty).

Bibliography

Aalen, O. (1978): Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics* 6 (4), pp. 701–726. DOI: 10.1214/aos/1176344247.

Affymetrix (2003a): *GeneChip® Expression Analysis Technical Manual*. rev 4.0 edition.

— (2003b): *Technical Note. Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays*. rev 2.0 edition.

Almendro, V., Marusyk, A., and Polyak, K. (2013): Cellular heterogeneity and molecular evolution in cancer. *Annual Review of Pathology* 8, pp. 277–302. DOI: 10.1146/annurev-pathol-020712-163923.

Anandappa, G. and Popat, S. (2016): Management of lung cancer. *Medicine*. Respiratory Disorders (Part 1 of 3) 44 (4), pp. 244–248. DOI: 10.1016/j.mpmed.2016.02.002.

Atay-Kayis, A. and Massam, H. (2005): A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* 92 (2), pp. 317–335. DOI: 10.1093/biomet/92.2.317.

Bair, E. and Tibshirani, R. (2004): Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. *PLoS Biology* 2 (4). DOI: 10.1371/journal.pbio.0020108.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006): Prediction by Supervised Principal Components. *Journal of the American Statistical Association* 101 (473), pp. 119–137. DOI: 10.1198/016214505000000628.

Banerjee, S. and Ghosal, S. (2015): Bayesian Structure Learning in Graphical Models. *Journal of Multivariate Analysis* 136 (C), pp. 147–162. DOI: 10.1016/j.jmva.2015.01.015.

Beath, K. J. (2014): A finite mixture method for outlier detection and robustness in meta-analysis. *Research Synthesis Methods* 5 (4), pp. 285–293. DOI: 10.1002/jrsm.1114.

Bedard, P. L., Hansen, A. R., Ratain, M. J., and Siu, L. L. (2013): Tumour heterogeneity in the clinic. *Nature* 501 (7467), pp. 355–364. DOI: 10.1038/nature12627.

Bender, R., Augustin, T., and Blettner, M. (2005): Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 24 (11), pp. 1713–1723. DOI: 10.1002/sim.2059.

Bergersen, L. C., Glad, I. K., and Lyng, H. (2011): Weighted lasso with data integration. *Statistical Applications in Genetics and Molecular Biology* 10 (1). DOI: 10.2202/1544-6115.1703.

Bernardo, J. M. and Smith, A. F. M. (1994): *Bayesian theory*. 1st ed. Wiley Series in Probability and Mathematical Statistics. Chichester [u.a.]: Wiley.

Bhattacharjee, S., Rajaraman, P., Jacobs, K., Wheeler, W., Melin, B., Hartge, P., Yeager, M., Chung, C., Chanock, S., and Chatterjee, N. (2012):

A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. *American Journal of Human Genetics* 90 (5), pp. 821–835. DOI: 10.1016/j.ajhg.2012.03.015.

Bühlmann, P. and Geer, S. van de (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Berlin [u.a.]: Springer.

Bühlmann, P. and Hothorn, T. (2007): Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science* 22 (4), pp. 477–505. DOI: 10.1214/07-STS242.

Böhning, D., Dietz, E., and Schlattmann, P. (1998): Recent developments in computer-assisted analysis of mixtures. *Biometrics* 54 (2), pp. 525–536.

Bickel, S., Bogojeska, J., Lengauer, T., and Scheffer, T. (2008): "Multi-task Learning for HIV Therapy Screening". *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. New York, USA: ACM, pp. 56–63. DOI: 10.1145/1390156.1390164.

Binder, H., Porzelius, C., and Schumacher, M. (2011): An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. *Biometrical Journal* 53 (2), pp. 170–189. DOI: 10.1002/bimj.201000152.

Binder, H. and Schumacher, M. (2008): Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* 9, p. 14. DOI: 10.1186/1471-2105-9-14.

Binder, H., Allignol, A., Schumacher, M., and Beyersmann, J. (2009): Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* 25 (7), pp. 890–896. DOI: 10.1093/bioinformatics/btp088.

Binder, H., Müller, T., Schwender, H., Golka, K., Steffens, M., Hengstler, J. G., Ickstadt, K., and Schumacher, M. (2012): Cluster-localized sparse logistic regression for SNP data. *Statistical Applications in Genetics and Molecular Biology* 11 (4). DOI: 10.1515/1544-6115.1694.

Bogojeska, J. and Lengauer, T. (2012): Hierarchical Bayes Model for Predicting Effectiveness of HIV Combination Therapies. *Statistical Applications in Genetics and Molecular Biology* 11 (3). DOI: 10.1515/1544-6115.1769.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003): A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19 (2), pp. 185–193.

Bolstad, B. M. (2004): *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. Dissertation. University of California, Berkeley.

Bommert, A., Rahnenführer, J., and Lang, M. (2017): A Multicriteria Approach to Find Predictive and Sparse Models with Stable Feature Selection for High-Dimensional Data. *Computational and Mathematical Methods in Medicine* vol. 2017, Article ID 7907163, p. 18. DOI: 10.1155/2017/7907163.

Boulesteix, A.-L. and Sauerbrei, W. (2011): Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics* 12 (3), pp. 215–229. DOI: 10.1093/bib/bbq085.

Boulesteix, A.-L. and Slawski, M. (2009): Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics* 10 (5), pp. 556–568. DOI: 10.1093/bib/bbp034.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017): IPF-LASSO: Integrative L_1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and Mathematical Methods in Medicine* vol. 2017, Article ID 7691937, p. 14. DOI: 10.1155/2017/7691937.

Breiman, L. (2001): Random Forests. *Machine Learning* 45 (1), pp. 5–32. DOI: 10.1023/A:1010933404324.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984): *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Wadsworth: Belmont, CA: Taylor & Francis.

Breslow, N. (1974): Covariance analysis of censored survival data. *Biometrics* 30 (1), pp. 89–99.

Brier, G. W. (1950): Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78 (1), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brooks, S. P. and Gelman, A. (1998): General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* 7 (4), pp. 434–455. DOI: 10.1080/10618600.1998.10474787.

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K. D., Maglott, D. R., and Murphy, T. D. (2015): Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research* 43 (D1), pp. D36–D42. DOI: 10.1093/nar/gku1055.

Bumgarner, R. (2013): DNA microarrays: Types, Applications and their future. *Current Protocols in Molecular Biology* 101 (1), pp. 22.1.1–22.1.11. DOI: 10.1002/0471142727.mb2201s101.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. C. (2007): Predicting survival from microarray data - a comparative study. *Bioinformatics* 23 (16), pp. 2080–2087. DOI: 10.1093/bioinformatics/btm305.

Bøvelstad, H. M., Nygård, S., and Borgan, Ø. (2009): Survival prediction from clinico-genomic models - a comparative study. *BMC Bioinformatics* 10 (1), p. 413. DOI: 10.1186/1471-2105-10-413.

Chang, L.-C., Lin, H.-M., Sibille, E., and Tseng, G. C. (2013): Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics* 14 (1), p. 368. DOI: 10.1186/1471-2105-14-368.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2011): SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.

Chen, B. and Ye, K. (2015): Componentwise variable selection in finite mixture regression. *Statistics and Its Interface* 8 (2), pp. 239–254. DOI: 10.4310/SII.2015.v8.n2.a11.

Chen, G., Zhong, H., Belousov, A., and Devanarayan, V. (2015): A PRIM approach to predictive-signature development for patient stratification. *Statistics in Medicine* 34 (2), pp. 317–342. DOI: 10.1002/sim.6343.

Chen, M.-H., Shao, Q., and Ibrahim, J. G. (2000): *Monte Carlo methods in Bayesian computation*. Springer Series in Statistics. New York [u.a.]: Springer.

Congdon, P. (2006): *Bayesian statistical modelling*. 2nd ed. Wiley Series in Probability and Statistics. Chichester: Wiley.

Cox, D. R. (1972): Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (2), pp. 187–220.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., METABRIC Group, Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C., and Aparicio, S. (2012): The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486 (7403), pp. 346–352. DOI: 10.1038/nature10983.

Danaher, P., Wang, P., and Witten, D. M. (2014): The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 76 (2), pp. 373–397. DOI: 10.1111/rssb.12033.

De Bin, R., Sauerbrei, W., and Boulesteix, A.-L. (2014): Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in Medicine* 33 (30), pp. 5310–5329. DOI: 10.1002/sim.6246.

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011): Bayesian Inference for General Gaussian Graphical Models With Application to Multivariate Lattice Data. *Journal of the American Statistical Association* 106 (496), pp. 1418–1433. DOI: 10.1198/jasa.2011.tm10465.

Doove, L. L., Dusseldorp, E., Deun, K. V., and Mechelen, I. V. (2014): A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Advances in Data Analysis and Classification* 8 (4), pp. 403–425. DOI: 10.1007/s11634-013-0159-x.

Drton, M. and Maathuis, M. H. (2017): Structure Learning in Graphical Modeling. *Annual Review of Statistics and Its Application* 4 (1), pp. 365–393. DOI: 10.1146/annurev-statistics-060116-053803.

Dyson, G. and Sing, C. F. (2014): Efficient identification of context dependent subgroups of risk from genome wide association studies. *Statistical Applications in Genetics and Molecular Biology* 13 (2), pp. 217–226. DOI: 10.1515/sagmb-2013-0062.

Edgar, R., Domrachev, M., and Lash, A. E. (2002): Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* 30 (1), pp. 207–210.

Efron, B. (1977): The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association* 72 (359), pp. 557–565. DOI: 10.2307/2286217.

Efron, B. and Tibshirani, R. (1997): Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association* 92 (438), pp. 548–560. DOI: 10.2307/2965703.

Eicher, T. S., Papageorgiou, C., and Raftery, A. E. (2011): Default priors and predictive performance in Bayesian model averaging, with application to growth

determinants. *Journal of Applied Econometrics* 26 (1), pp. 30–55. DOI: 10.1002/jae.1112.

Ein-Dor, L., Zuk, O., and Domany, E. (2006): Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences* 103 (15), pp. 5923–5928. DOI: 10.1073/pnas.0601231103.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005): Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21 (2), pp. 171–178. DOI: 10.1093/bioinformatics/bth469.

Evangelou, E. and Ioannidis, J. P. A. (2013): Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics* 14 (6), pp. 379–389. DOI: 10.1038/nrg3472.

Fawcett, T. (2006): An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

Fehringer, G., Kraft, P., Pharoah, P. D., Eeles, R. A., Chatterjee, N., Schumacher, F. R., and 126 other authors (2016): Cross-Cancer Genome-Wide Analysis of Lung, Ovary, Breast, Prostate, and Colorectal Cancer Reveals Novel Pleiotropic Associations. *Cancer Research* 76 (17), pp. 5103–5114. DOI: 10.1158/0008-5472.CAN-15-2980.

Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011): Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30 (24), pp. 2867–2880. DOI: 10.1002/sim.4322.

Fraley, C. and Raftery, A. E. (2002): Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* 97 (458), pp. 611–631. DOI: 10.1198/016214502760047131.

Friedman, J., Hastie, T., and Tibshirani, R. (2010): Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33 (1), pp. 1–22.

Friedman, J. H. and Fisher, N. I. (1999): Bump hunting in high-dimensional data. *Statistics and Computing* 9 (2), pp. 123–143. DOI: 10.1023/A:1008894516817.

Gao, C., Zhu, Y., Shen, X., and Pan, W. (2016): Estimation of multiple networks in Gaussian mixture models. *Electronic Journal of Statistics* 10 (1), pp. 1133–1154. DOI: 10.1214/16-EJS1135.

Gelman, A. (2004): *Bayesian data analysis*. 2nd ed. Texts in Statistical Science. Chapman & Hall/CRC.

Gelman, A. and Rubin, D. B. (1992): Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7 (4), pp. 457–472. DOI: 10.1214/ss/1177011136.

GeneCards®: The Human Gene Database. https://www.genecards.org. Accessed: June 2018.

George, E. I. and McCulloch, R. E. (1993): Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association* 88 (423), pp. 881–889. DOI: 10.1080/01621459.1993.10476353.

Gerds, T. A. and Schumacher, M. (2006): Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal* 48 (6), pp. 1029–1040. DOI: 10.1002/bimj.200610301.

Geyer, C. J. (1992): Practical Markov Chain Monte Carlo. *Statistical Science* 7 (4), pp. 473–483. DOI: 10.1214/ss/1177011137.

Göhlmann, H. and Talloen, W. (2009): *Gene Expression Studies Using Affymetrix Microarray*. Mathematical and Computational Biology Series. Chapman & Hall/CRC.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995): Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 44 (4), pp. 455–472. DOI: 10.2307/2986138.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286 (5439), pp. 531–537.

Gordon, L. and Olshen, R. A. (1985): Tree-structured survival analysis. *Cancer Treatment Reports* 69 (10), pp. 1065–1069.

Gower, J. C. (1971): A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27 (4), pp. 857–871. DOI: 10.2307/2528823.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999): Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 18 (17-18), pp. 2529–2545.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011): Joint estimation of multiple graphical models. *Biometrika* 98 (1), pp. 1–15. DOI: 10.1093/biomet/asq060.

Han, B. and Eskin, E. (2011): Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *American Journal of Human Genetics* 88 (5), pp. 586–598. DOI: 10.1016/j.ajhg.2011.04.014.

— (2012): Interpreting Meta-Analyses of Genome-Wide Association Studies. *PLoS Genetics* 8 (3). DOI: 10.1371/journal.pgen.1002555.

Hanahan, D. and Weinberg, R. A. (2011): Hallmarks of cancer: the next generation. *Cell* 144 (5), pp. 646–674. DOI: 10.1016/j.cell.2011.02.013.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996): Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine* 15 (4), pp. 361–387. DOI: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

Hastie, T., Tibshirani, R., and Friedman, J. (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics.

He, H. and Garcia, E. A. (2009): Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21 (9), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

He, Z. and Yu, W. (2010): Stable feature selection for biomarker discovery. *Computational Biology and Chemistry* 34 (4), pp. 215–225. DOI: 10.1016/j.compbiolchem.2010.07.002.

Heagerty, P. J. and Zheng, Y. (2005): Survival Model Predictive Accuracy and ROC Curves. *Biometrics* 61 (1), pp. 92–105. DOI: 10.1111/j.0006-341X.2005.030814.x.

Heimes, A.-S., Madjar, K., Edlund, K., Battista, M. J., Almstedt, K., Elger, T., Krajnak, S., Rahnenführer, J., Brenner, W., Hasenburg, A., Hengstler, J. G., and Schmidt, M. (2017): Subtype-specific prognostic impact of different immune signatures in node-negative breast cancer. *Breast Cancer Research and Treatment* 165 (2), pp. 293–300. DOI: 10.1007/s10549-017-4327-0.

Held, L., Gravestock, I., and Sabanés Bové, D. (2016): Objective Bayesian model selection for Cox regression. *Statistics in Medicine* 35 (29), pp. 5376–5390. DOI: 10.1002/sim.7089.

Hellwig, B., Madjar, K., Edlund, K., Marchan, R., Cadenas, C., Heimes, A.-S., Almstedt, K., Lebrecht, A., Sicking, I., Battista, M. J., Micke, P., Schmidt, M., Hengstler, J. G., and Rahnenführer, J. (2016): Epsin Family Member 3 and Ribosome-Related Genes Are Associated with Late Metastasis in Estrogen Receptor-Positive Breast Cancer and Long-Term Survival in Non-Small Cell Lung Cancer Using a Genome-Wide Identification and Validation Strategy. *PLoS ONE* 11 (12). DOI: 10.1371/journal.pone.0167585.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003): Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal* 327 (7414), pp. 557–560.

Hoerl, A. E. and Kennard, R. W. (1970): Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12 (1), pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

Hothorn, T. and Bühlmann, P. (2006): Model-based boosting in high dimensions. *Bioinformatics* 22 (22), pp. 2828–2829. DOI: 10.1093/bioinformatics/btl462.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006): Survival ensembles. *Biostatistics* 7 (3), pp. 355–373. DOI: 10.1093/biostatistics/kxj011.

Houwelingen, H. C. van, Bruinsma, T., Hart, A. A. M., Veer, L. J. van't, and Wessels, L. F. A. (2006): Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine* 25 (18), pp. 3201–3216. DOI: 10.1002/sim.2353.

Huang, Y., Fong, Y., Wei, J., and Feng, Z. (2011): Borrowing Information across Populations in Estimating Positive and Negative Predictive Values. *Journal of the Royal Statistical Society. Series C, Applied statistics* 60 (5), pp. 633–653. DOI: 10.1111/j.1467-9876.2011.00761.x.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2005): *Bayesian survival analysis*. Corr. 2nd print. Springer Series in Statistics. New York [u.a.]: Springer.

Ickstadt, K., Schäfer, M., and Zucknick, M. (2018): Integrative Bayesian Approaches for Molecular Biology. *Annual Review of Statistics and Its Application* 5 (1), pp. 141–167. DOI: 10.1146/annurev-statistics-031017-100438.

Ioannidis, J. P., Patsopoulos, N. A., and Evangelou, E. (2007): Heterogeneity in Meta-Analyses of Genome-Wide Association Investigations. *PLoS ONE* 2 (9). DOI: 10.1371/journal.pone.0000841.

Irizarry, R. A., Hobbs, B., Collin, F., BeazerBarclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003): Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4 (2), pp. 249–264. DOI: 10.1093/biostatistics/4.2.249.

Ishwaran, H. and Rao, J. S. (2005): Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics* 33 (2), pp. 730–773. DOI: 10.1214/009053604000001147.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005): Experiments in Stochastic Computation for High-Dimensional Graphical Models. *Statistical Science* 20 (4), pp. 388–400. DOI: 10.1214/088342305000000304.

Junttila, M. R. and Sauvage, F. J. de (2013): Influence of tumour microenvironment heterogeneity on therapeutic response. *Nature* 501 (7467), pp. 346–354. DOI: 10.1038/nature12626.

Kammers, K., Lang, M., Hengstler, J. G., Schmidt, M., and Rahnenführer, J. (2011): Survival models with preclustered gene groups as covariates. *BMC Bioinformatics* 12 (478). DOI: 10.1186/1471-2105-12-478.

Kaplan, E. L. and Meier, P. (1958): Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 53 (282), pp. 457–481. DOI: 10.2307/2281868.

Kar, S. P., Beesley, J., Al Olama, A. A., Michailidou, K., Tyrer, J., Kote-Jarai, Z. A., and 217 other authors (2016): Genome-Wide Meta-Analyses of Breast, Ovarian, and Prostate Cancer Association Studies Identify Multiple New Susceptibility Loci Shared by at Least Two Cancer Types. *Cancer Discovery* 6 (9), pp. 1052–1067. DOI: 10.1158/2159-8290.CD-15-1227.

Kass, R. E. and Wasserman, L. (1996): The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association* 91 (435), pp. 1343–1370. DOI: 10.2307/2291752.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012): Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 28 (24), pp. 3290–3297. DOI: 10.1093/bioinformatics/bts595.

Klein, J. P. and Moeschberger, M. L. (2003): *Survival analysis*. 2nd ed. Statistics for Biology and Health. New York [u.a.]: Springer.

Kratz, J. R., He, J., Van Den Eeden, S. K., Zhu, Z.-H., Gao, W., Pham, P. T., Mulvihill, M. S., Ziaei, F., Zhang, H., Su, B., Zhi, X., Quesenberry, C. P., Habel, L. A., Deng, Q., Wang, Z., Zhou, J., Li, H., Huang, M.-C., Yeh, C.-C., Segal, M. R., Ray, M. R., Jones, K. D., Raz, D. J., Xu, Z., Jahan, T. M., Berryman, D., He, B., Mann, M. J., and Jablons, D. M. (2012): A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. *The Lancet* 379 (9818), pp. 823–832. DOI: 10.1016/S0140-6736(11)61941-7.

Lausser, L., Müssel, C., Maucher, M., and Kestler, H. A. (2013): Measuring and visualizing the stability of biomarker selection techniques. *Computational Statistics* 28 (1), pp. 51–65. DOI: 10.1007/s00180-011-0284-y.

LeBlanc, M. and Crowley, J. (1993): Survival Trees by Goodness of Split. *Journal of the American Statistical Association* 88 (422), pp. 457–467. DOI: 10.2307/2290325.

Lee, K.-J., Chen, R.-B., and Wu, Y. N. (2016): Bayesian variable selection for finite mixture model of linear regressions. *Computational Statistics & Data Analysis* 95, pp. 1–16. DOI: 10.1016/j.csda.2015.09.005.

Lee, K. H., Chakraborty, S., and Sun, J. (2011): Bayesian Variable Selection in Semiparametric Proportional Hazards Model for High Dimensional Survival Data.

The International Journal of Biostatistics 7 (1), pp. 1–32. DOI: 10.2202/1557-4679.1301.

Ley, E. and Steel, M. F. (2009): On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics* 24 (4), pp. 651–674. DOI: 10.1002/jae.1057.

Li, F. and Zhang, N. R. (2010): Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics. *Journal of the American Statistical Association* 105 (491), pp. 1202–1214. DOI: 10.1198/jasa.2010.tm08177.

Li, J. and Tseng, G. C. (2011): An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* 5 (2A), pp. 994–1019. DOI: 10.1214/10-AOAS393.

Li, Q., Wang, S., Huang, C.-C., Yu, M., and Shao, J. (2014): Meta-Analysis Based Variable Selection for Gene Expression Data. *Biometrics* 70 (4), pp. 872–880. DOI: 10.1111/biom.12213.

Lipkovich, I., Dmitrienko, A., and D'Agostino, R. B. (2017): Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine* 36 (1), pp. 136–196. DOI: 10.1002/sim.7064.

Liu, J., Huang, J., and Ma, S. (2014): Integrative Analysis of Cancer Diagnosis Studies with Composite Penalization. *Scandinavian Journal of Statistics, Theory and Applications* 41 (1), pp. 87–103. DOI: 10.1111/j.1467-9469.2012.00816.x.

Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., and Ma, S. (2014): Integrative Analysis of Prognosis Data on Multiple Cancer Subtypes. *Biometrics* 70 (3), pp. 480–488. DOI: 10.1111/biom.12177.

Loh, W.-Y. (2014): Fifty Years of Classification and Regression Trees. *International Statistical Review* 82 (3), pp. 329–348.

Ma, S., Huang, J., and Moran, M. S. (2009): Identification of genes associated with multiple cancers via integrative analysis. *BMC Genomics* 10 (1), p. 535. DOI: 10.1186/1471-2164-10-535.

Marchan, R., Büttner, B., Lambert, J., Edlund, K., Glaeser, I., Blaszkewicz, M., Leonhardt, G., Marienhoff, L., Kaszta, D., Anft, M., Watzl, C., Madjar, K., Grinberg, M., Rempel, E., Hergenröder, R., Selinski, S., Rahnenführer, J., Lesjak, M. S., Stewart, J. D., Cadenas, C., and Hengstler, J. G. (2017): Glycerol-3-phosphate Acyltransferase 1 Promotes Tumor Cell Migration and Poor Survival in Ovarian Carcinoma. *Cancer Research* 77 (17), pp. 4589–4601. DOI: 10.1158/0008-5472.CAN-16-2065.

Matsui, S., Buyse, M., and Simon, R. (2015): *Design and Analysis of Clinical Trials for Predictive Medicine*. Chapman & Hall/CRC Biostatistics Series.

Matsui, S., Noma, H., Qu, P., Sakai, Y., Matsui, K., Heuck, C., and Crowley, J. (2017): Multi-subgroup gene screening using semi-parametric hierarchical mixture models and the optimal discovery procedure: Application to a randomized clinical trial in multiple myeloma. *Biometrics* 74 (1), pp. 313–320. DOI: 10.1111/biom.12716.

Mauguen, A., Zabor, E. C., Thomas, N. E., Berwick, M., Seshan, V. E., and Begg, C. B. (2017): Defining Cancer Subtypes With Distinctive Etiologic Profiles: An Application to the Epidemiology of Melanoma. *Journal of the American Statistical Association* 112 (517), pp. 54–63. DOI: 10.1080/01621459.2016.1191499.

McCall, M. N., Bolstad, B. M., and Irizarry, R. A. (2010): Frozen robust multiarray analysis (fRMA). *Biostatistics* 11 (2), pp. 242–253. DOI: 10.1093/biostatistics/kxp059.

McLachlan, G. and Peel, D. (2000): *Finite Mixture Models*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. DOI: 10.1002/0471721182.

Meinshausen, N. and Bühlmann, P. (2010): Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4), pp. 417–473. DOI: 10.1111/j.1467-9868.2010.00740.x.

Melzer, D., Pilling, L. C., Fellows, A. D., Holly, A. C., Harries, L. W., and Ferrucci, L. (2013): Gene Expression Biomarkers and Longevity. *Annual Review of Gerontology and Geriatrics* 33 (1), pp. 233–258. DOI: 10.1891/0198-8794.33.233.

Michiels, S., Koscielny, S., and Hill, C. (2005): Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet* 365 (9458), pp. 488–492. DOI: 10.1016/S0140-6736(05)17866-0.

Mitra, R., Müller, P., and Ji, Y. (2016): Bayesian Graphical Models for Differential Pathways. *Bayesian Analysis* 11 (1), pp. 99–124. DOI: 10.1214/14-BA931.

Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012): Evaluating Random Forests for Survival Analysis using Prediction Error Curves. *Journal of Statistical Software* 50 (11), pp. 1–23.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005): Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 21 (15), pp. 3301–3307. DOI: 10.1093/bioinformatics/bti499.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., and Boivin, J.-F. (2005): Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Statistics and Computing* 15 (3), pp. 231–239. DOI: 10.1007/s11222-005-1311-z.

Nelson, W. (1972): Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics* 14 (4), pp. 945–966. DOI: 10.2307/1267144.

Netzer, C. (2013): *Vorhersage der Überlebenswahrscheinlichkeit für Patientenuntergruppen mit hochdimensionalen Daten am Beispiel zweier Lungenkrebskohorten*. Dissertation. TU Dortmund Universtiy. DOI: http://dx.doi.org/10.17877/DE290R-5760.

Neupane, B., Loeb, M., Anand, S. S., and Beyene, J. (2012): Meta-analysis of genetic association studies under heterogeneity. *European Journal of Human Genetics* 20 (11), pp. 1174–1181. DOI: 10.1038/ejhg.2012.75.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., and Posch, M. (2016): Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics* 26 (1), pp. 99–119. DOI: 10.1080/10543406.2015.1092034.

Park, T. and Casella, G. (2008): The Bayesian Lasso. *Journal of the American Statistical Association* 103 (482), pp. 681–686. DOI: 10.1198/016214508000000337.

PDQ® Adult Treatment Editorial Board (2018): *PDQ® Non-Small Cell Lung Cancer Treatment-Health Professional Version*. Bethesda, MD: National Cancer Institute. Updated 2018-03-20. Available at: https://www.cancer.gov/

types/lung/hp/non-small-cell-lung-treatment-pdq. Accessed 2018-05-04.

Pei, Y.-F., Tian, Q., Zhang, L., and Deng, H.-W. (2016): Exploring the major sources and extent of heterogeneity in a genome-wide association meta-analysis. *Annals of Human Genetics* 80 (2), pp. 113–122. DOI: 10.1111/ahg.12143.

Perou, C. M., Sørlie, T., Eisen, M. B., Rijn, M. van de, Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000): Molecular portraits of human breast tumours. *Nature* 406 (6797), pp. 747–752. DOI: 10.1038/35021093.

Peterson, C., Stingo, F. C., and Vannucci, M. (2015): Bayesian Inference of Multiple Gaussian Graphical Models. *Journal of the American Statistical Association* 110 (509), pp. 159–174. DOI: 10.1080/01621459.2014.896806.

Peterson, C. B., Stingo, F. C., and Vannucci, M. (2016): Joint Bayesian variable and graph selection for regression models with network-structured predictors. *Statistics in Medicine* 35 (7), pp. 1017–1031. DOI: 10.1002/sim.6792.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002): Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* 21 (19), pp. 2917–2930. DOI: 10.1002/sim.1296.

Raim, A. M., Neerchal, N. K., and Morel, J. G. (2014): "Large Cluster Approximation to the Finite Mixture Information Matrix with an Application to Meta-Analysis". *JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association*, pp. 4025–4037.

Richter, J., Madjar, K., and Rahnenführer, J. (2018): *Model-Based Optimization of Subgroup Weights for Survival Analysis*. Tech. rep. 3. Faculty of Statistics, TU Dortmund University.

Rothwell, P. M. (2005): Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet* 365 (9454), pp. 176–186. DOI: 10.1016/S0140-6736(05)17709-5.

Roverato, A. (2002): Hyper Inverse Wishart Distribution for Non-decomposable Graphs and its Application to Bayesian Inference for Gaussian Graphical Models. *Scandinavian Journal of Statistics* 29 (3), pp. 391–411. DOI: 10.1111/1467-9469.00297.

Saegusa, T. and Shojaie, A. (2016): Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics* 10 (1), pp. 1341–1392. DOI: 10.1214/16-EJS1137.

Sauerbrei, W., Boulesteix, A.-L., and Binder, H. (2011): Stability investigations of multivariable regression models derived from low- and high-dimensional data. *Journal of Biopharmaceutical Statistics* 21 (6), pp. 1206–1231. DOI: 10.1080/10543406.2011.629890.

Scarbrough, P. M., Weber, R. P., Iversen, E. S., Brhane, Y., Amos, C. I., Kraft, P., Hung, R. J., Sellers, T. A., Witte, J. S., Pharoah, P., Henderson, B. E., Gruber, S. B., Hunter, D. J., Garber, J. E., Joshi, A. D., McDonnell, K., Easton, D. F., Eeles, R., Kote-Jarai, Z., Muir, K., Doherty, J. A., and Schildkraut, J. M. (2016): A Cross-Cancer Genetic Association Analysis of the DNA Repair and DNA Damage Signaling Pathways for Lung,

Ovary, Prostate, Breast, and Colorectal Cancer. *Cancer Epidemiology Biomarkers & Prevention* 25 (1), pp. 193–200. DOI: 10.1158/1055-9965.EPI-15-0649.

Schäfer, J. and Strimmer, K. (2005): A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* 4 (1). DOI: 10.2202/1544-6115.1175.

Schmoor, C., Ulm, K., and Schumacher, M. (1993): Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial. *Statistics in Medicine* 12 (24), pp. 2351–2366.

Schumacher, M., Binder, H., and Gerds, T. (2007): Assessment of survival prediction models based on microarray data. *Bioinformatics* 23 (14), pp. 1768–1774. DOI: 10.1093/bioinformatics/btm232.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009): Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25 (22), pp. 2906–2912. DOI: 10.1093/bioinformatics/btp543.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011): Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* 39 (5), pp. 1–13. DOI: 10.18637/jss.v039.i05.

Simon, R. (2002): Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in Medicine* 21 (19), pp. 2909–2916. DOI: 10.1002/sim.1295.

Song, C. and Tseng, G. C. (2014): Hypothesis setting and order statistic for robust genomic meta-analysis. *The Annals of Applied Statistics* 8 (2), pp. 777–800.

Stingo, F. C. and Vannucci, M. (2011): Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics* 27 (4), pp. 495–501. DOI: 10.1093/bioinformatics/btq690.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011): Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics* 5 (3), pp. 1978–2002. DOI: 10.1214/11-AOAS463.

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008): Interaction Trees with Censored Survival Data. *The International Journal of Biostatistics* 4 (1). DOI: 10.2202/1557-4679.1071.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009): Subgroup Analysis via Recursive Partitioning. *Journal of Machine Learning Research* 10, pp. 141–158.

Su, X., Meneses, K., McNees, P., and Johnson, W. O. (2011): Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60 (3), pp. 457–474. DOI: 10.1111/j.1467-9876.2010.00754.x.

Subramanian, J. and Simon, R. (2010): Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *Journal of the National Cancer Institute* 102 (7), pp. 464–474. DOI: 10.1093/jnci/djq025.

Tang, H., Wang, S., Xiao, G., Schiller, J., Papadimitrakopoulou, V., Minna, J., Wistuba, I. I., and Xie, Y. (2017): Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies. *Annals of Oncology* 28 (4), pp. 733–740. DOI: 10.1093/annonc/mdw683.

The Tumor Analysis Best Practices Working Group (2004): Expression profiling — best practices for data generation and interpretation in clinical trials. *Nature Reviews Genetics* 5 (3), pp. 229–237. DOI: 10.1038/nrg1297.

Thompson, J. R., Attia, J., and Minelli, C. (2011): The meta-analysis of genome-wide association studies. *Briefings in Bioinformatics* 12 (3), pp. 259–269. DOI: 10.1093/bib/bbr020.

Tibshirani, R. (1996): Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1), pp. 267–288.

— (1997): The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine* 16 (4), pp. 385–395. DOI: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

Treppmann, T., Ickstadt, K., and Zucknick, M. (2017): Integration of Multiple Genomic Data Sources in a Bayesian Cox Model for Variable Selection and Prediction. *Computational and Mathematical Methods in Medicine* vol. 2017, Article ID 7340565, p. 19. DOI: 10.1155/2017/7340565.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012): Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research* 40 (9), pp. 3785–3799. DOI: 10.1093/nar/gkr1265.

Tutz, G. and Binder, H. (2005): Localized classification. *Statistics and Computing* 15 (3), pp. 155–166. DOI: 10.1007/s11222-005-1305-x.

Tutz, G. and Binder, H. (2006): Generalized Additive Modeling with Implicit Variable Selection by Likelihood-Based Boosting. *Biometrics* 62 (4), pp. 961–971. DOI: 10.1111/j.1541-0420.2006.00578.x.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011): On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Statistics in Medicine* 30 (10), pp. 1105–1117. DOI: 10.1002/sim.4154.

Verweij, P. J. and Van Houwelingen, H. C. (1994): Penalized likelihood in Cox regression. *Statistics in Medicine* 13 (23-24), pp. 2427–2436.

Wang, H. (2012): Bayesian Graphical Lasso Models and Efficient Posterior Computation. *Bayesian Analysis* 7 (4), pp. 867–886. DOI: 10.1214/12-BA729.

— (2015): Scaling It Up: Stochastic Search Structure Learning in Graphical Models. *Bayesian Analysis* 10 (2), pp. 351–377. DOI: 10.1214/14-BA916.

Wang, M., Spiegelman, D., Kuchiba, A., Lochhead, P., Kim, S., Chan, A. T., Poole, E. M., Tamimi, R., Tworoger, S. S., Giovannucci, E., Rosner, B., and Ogino, S. (2016): Statistical methods for studying disease subtype heterogeneity. *Statistics in Medicine* 35 (5), pp. 782–800. DOI: 10.1002/sim.6793.

Wen, X. and Stephens, M. (2014): Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene-environment interactions. *The Annals of Applied Statistics* 8 (1), pp. 176–203. DOI: 10.1214/13-AOAS695.

Weyer, V. and Binder, H. (2015): A weighting approach for judging the effect of patient strata on high-dimensional risk prediction signatures. *BMC Bioinformatics* 16, p. 294. DOI: 10.1186/s12859-015-0716-8.

White, I. R. and Royston, P. (2009): Imputing missing covariate values for the Cox model. *Statistics in Medicine* 28 (15), pp. 1982–1998. DOI: 10.1002/sim.3618.

Witten, D. M. and Tibshirani, R. (2010a): A framework for feature selection in clustering. *Journal of the American Statistical Association* 105 (490), pp. 713–726. DOI: 10.1198/jasa.2010.tm09415.

Witten, D. M. and Tibshirani, R. (2010b): Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research* 19 (1), pp. 29–51. DOI: 10.1177/0962280209105024.

World Health Organization (2014): *World Cancer Report 2014*. Ed. by Bernard W Stewart and Christopher P Wild. International Agency for Research on Cancer, Lyon, France.

Wright, M. and Ziegler, A. (2017): ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77 (1), pp. 1–17. DOI: 10.18637/jss.v077.i01.

Xu, Y., Yu, M., Zhao, Y.-Q., Li, Q., Wang, S., and Shao, J. (2015): Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics* 71 (3), pp. 645–653. DOI: 10.1111/biom.12322.

Yajima, M., Telesca, D., Ji, Y., and Muller, P. (2012): Differential Patterns of Interaction and Gaussian Graphical Models. *Collection of Biostatistics Research Archive, COBRA Preprint Series* 91.

Yang, A., Jiang, X., Liu, P., and Lin, J. (2016): Sparse Bayesian multinomial probit regression model with correlation prior for high-dimensional data classification. *Statistics & Probability Letters* 119, pp. 241–247. DOI: 10.1016/j.spl.2016.08.008.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., and 51 other authors (2018): Ensembl 2018. *Nucleic Acids Research* 46 (D1), pp. D754–D761. DOI: 10.1093/nar/gkx1098.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012): Estimating Optimal Treatment Regimes from a Classification Perspective. *Stat* 1 (1), pp. 103–114. DOI: 10.1002/sta.411.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012): Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association* 107 (449), pp. 1106–1118. DOI: 10.1080/01621459.2012.695674.

Zhu, J. and Hastie, T. (2004): Classification of gene microarrays by penalized logistic regression. *Biostatistics* 5 (3), pp. 427–443. DOI: 10.1093/biostatistics/kxg046.

Zou, H. and Hastie, T. (2005): Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2), pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

Zucknick, M., Saadati, M., and Benner, A. (2015): Nonidentical twins: Comparison of frequentist and Bayesian lasso for Cox models. *Biometrical Journal* 57 (6), pp. 959–981. DOI: 10.1002/bimj.201400160.

Appendix A

Algorithms

A.1 Regularization path for the Cox model via cyclical coordinate descent

The regression coefficients in the Cox model are estimated via cyclical coordinate descent from a regularized partial log-likelihood. The algorithm proposed by Simon et al. (2011) is implemented in the R package `glmnet` (version 2.0-13) and described in the following.

The scaled partial log-likelihood with elastic net penalty is given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \frac{2}{n} l(\boldsymbol{\beta}) - \lambda_P \cdot P_\alpha(\boldsymbol{\beta}) \right\}, \; P_\alpha(\boldsymbol{\beta}) = \left(\alpha \|\boldsymbol{\beta}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}\|_2^2 \right), \; \alpha \in [0,1].$$

This maximization problem is similar to the standard Newton-Raphson algorithm but instead of solving a general least squares problem, a penalized reweighted least squares problem is solved here.

Let $\boldsymbol{x} \in \mathbb{R}^{n \times p}$ be the design matrix and $\boldsymbol{\beta}$ the p-dimensional vector of regression coefficients. Let $\dot{l}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ denote the gradient and $\ddot{l}(\boldsymbol{\beta}) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$ the Hessian of the partial log-likelihood with respect to $\boldsymbol{\beta}$. In the case of tied events Breslow's modification of the partial likelihood (section 3.1.2.1) is used. A two term Taylor series expansion of the partial log-likelihood centered at $\tilde{\boldsymbol{\beta}}$ is given by

$$\begin{aligned} l(\boldsymbol{\beta}) &\approx l(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})' \dot{l}(\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})' \ddot{l}(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \\ &= l(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{x}\boldsymbol{\beta} - \boldsymbol{x}\tilde{\boldsymbol{\beta}})' \dot{l}(\boldsymbol{x}\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{x}\boldsymbol{\beta} - \boldsymbol{x}\tilde{\boldsymbol{\beta}})' \ddot{l}(\boldsymbol{x}\tilde{\boldsymbol{\beta}})(\boldsymbol{x}\boldsymbol{\beta} - \boldsymbol{x}\tilde{\boldsymbol{\beta}}) \\ &\approx \frac{1}{2} \left(\zeta(\boldsymbol{x}\tilde{\boldsymbol{\beta}}) - \boldsymbol{x}\boldsymbol{\beta} \right)' \ddot{l}(\boldsymbol{x}\tilde{\boldsymbol{\beta}}) \left(\zeta(\boldsymbol{x}\tilde{\boldsymbol{\beta}}) - \boldsymbol{x}\boldsymbol{\beta} \right) + C(\boldsymbol{x}\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}), \end{aligned} \tag{A.1}$$

where $\zeta(\boldsymbol{x}\tilde{\boldsymbol{\beta}}) = \boldsymbol{x}\tilde{\boldsymbol{\beta}} - \ddot{l}(\boldsymbol{x}\tilde{\boldsymbol{\beta}})^{-1} \dot{l}(\boldsymbol{x}\tilde{\boldsymbol{\beta}})$ and $C(\boldsymbol{x}\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}})$ is independent of $\boldsymbol{\beta}$. To speed up computation, the full matrix $\ddot{l}(\boldsymbol{x}\tilde{\boldsymbol{\beta}})$ is replaced by a diagonal matrix with the diagonal entries of $\ddot{l}(\boldsymbol{x}\tilde{\boldsymbol{\beta}})$. Equation (A.1) simplifies to

$$l(\boldsymbol{\beta}) \approx \frac{1}{2} \sum_{m=1}^{n} l_m(\boldsymbol{x}\tilde{\boldsymbol{\beta}}) \left(\zeta_m(\boldsymbol{x}\tilde{\boldsymbol{\beta}}) - \boldsymbol{\beta}' \boldsymbol{x}_m \right)^2 + C(\boldsymbol{x}\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}), \tag{A.2}$$

where $l_m(\boldsymbol{x}\tilde{\boldsymbol{\beta}})$ is the m-th diagonal element of $\ddot{l}(\boldsymbol{x}\tilde{\boldsymbol{\beta}})$.

The algorithm works as follows:

1. Initialize $\tilde{\boldsymbol{\beta}}$
2. Calculate $\ddot{l}(\boldsymbol{x}\tilde{\boldsymbol{\beta}})$ and $\zeta(\boldsymbol{x}\tilde{\boldsymbol{\beta}})$
3. Solve the penalized weighted least squares problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{m=1}^{n} l_m(\boldsymbol{x}\tilde{\boldsymbol{\beta}}) \left(\zeta_m(\boldsymbol{x}\tilde{\boldsymbol{\beta}}) - \boldsymbol{\beta}'\boldsymbol{x}_m \right)^2 + \lambda_P P_\alpha(\boldsymbol{\beta}) \right\} \tag{A.3}$$

   via cyclical coordinate descent. In each coordinate descent step the objective function in (A.3) is partially minimized with respect to β_i by computing the corresponding derivative and solving it for β_i. All other coefficients $\beta_{j,j\neq i}$ are considered to be fixed. This minimization problem is repeated for each element in $\boldsymbol{\beta}$ until convergence minimizes the objective in (A.3).
4. Set $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$
5. Repeat steps 2 to 4 until $\hat{\boldsymbol{\beta}}$ converges.

The optimal $\lambda_P \geq 0$ is defined by K-fold cross-validation, with $K = 10$ as default. Solutions are computed over a grid of n_λ different values for λ_P (default is $n_\lambda = 100$), beginning with $\lambda_{\max}$ (to set $\boldsymbol{\beta} = \mathbf{0}$) and becoming increasingly smaller until $\lambda_{\min}$ (near the unregularized solution, with $\lambda_{\min} = \epsilon\lambda_{\max}$ and $\epsilon = 0.0001$ for $n \geq p$ and $\epsilon = 0.01$ for $n < p$). The cross-validation method proposed by Houwelingen et al. (2006) is applied. All observations are divided into K subsets of approximately the same size. $K-1$ subsets are used to build the model and validation is performed on the k-th subset. For a given λ_P and subset k the estimator of the goodness of fit is defined as

$$\widehat{CV}_k(\lambda_P) = l(\boldsymbol{\beta}_{-k}(\lambda_P)) - l_{-k}(\boldsymbol{\beta}_{-k}(\lambda_P)),$$

where l_{-k} is the partial log-likelihood without subset k and $\boldsymbol{\beta}_{-k}(\lambda_P)$ is the optimal $\boldsymbol{\beta}$ from maximizing $l_{-k} + \lambda_P\|\boldsymbol{\beta}\|_1$. $\widehat{CV}_k(\lambda_P)$ indicates how much the likelihood is improved by adding the k-th subset. The optimal λ_P is received by maximizing $\sum_{k=1}^{K} \widehat{CV}_k(\lambda_P)$ subject to λ_P.

A.2 Regularization path for the multinomial logistic regression via cyclical coordinate descent

For each patient m a categorical response $s_m \in \{1, ..., S\}$ and a q-dimensional vector of covariates $\boldsymbol{z}_m$ are observed, $m = 1, ..., n$. Let $\pi_s(\boldsymbol{z}_m) = P(\mathcal{S} = s|\boldsymbol{z}_m)$ be the probability to be modeled by the multinomial logistic regression and $\boldsymbol{\theta} = (\beta_{01}, \boldsymbol{\beta}_1', ..., \beta_{0S}, \boldsymbol{\beta}_S')' \in \mathbb{R}^{S(q+1)}$ the vector of unknown parameters to be estimated. The log-likelihood is given by

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{m=1}^{n} \ln(\pi_{s_m}(\boldsymbol{z}_m)) \\ &= \sum_{m=1}^{n} \left[\sum_{s=1}^{S} \mathbb{1}(s_m = s)(\beta_{0s} + \boldsymbol{z}_m'\boldsymbol{\beta}_s) - \ln\left(\sum_{s=1}^{S} \exp(\beta_{0s} + \boldsymbol{z}_m'\boldsymbol{\beta}_s) \right) \right]. \end{aligned}$$

The unknown parameters $\boldsymbol{\theta}$ are estimated from the penalized log-likelihood by solving the following maximization problem

$$\hat{\boldsymbol{\theta}}_{\lambda_P} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \frac{1}{n} l(\boldsymbol{\theta}) - \lambda_P \sum_{s=1}^{S} P_\alpha(\boldsymbol{\beta}_s) \right\},$$

where $P_\alpha(\boldsymbol{\beta}_s)$ corresponds to the elatic net penalty in section A.1. $\boldsymbol{\theta}$ is estimated similarly to section A.1 via cyclical coordinate descent. Here, a partial quadratic approximation to the log-likelihood is formed, allowing only $(\beta_{0s}, \boldsymbol{\beta}_s)$ to vary for a single class at a time

$$l_{Qs}(\beta_{0s}, \boldsymbol{\beta}_s) = -\frac{1}{2n} \sum_{m=1}^{n} \tilde{w}_{ms}(\zeta_{ms} - \beta_{0s} - \boldsymbol{z}'_m \boldsymbol{\beta}_s)^2 + C(\tilde{\boldsymbol{\theta}}),$$

with working response $\zeta_{ms} = \tilde{\beta}_{0s} + \boldsymbol{z}'_m \tilde{\boldsymbol{\beta}}_s + \frac{\mathbb{1}(s_m = s) - \tilde{\pi}_s(\boldsymbol{z}_m)}{\tilde{\pi}_s(\boldsymbol{z}_m)(1 - \tilde{\pi}_s(\boldsymbol{z}_m))}$ and weights $\tilde{w}_{ms} = \tilde{\pi}_s(\boldsymbol{z}_m)(1 - \tilde{\pi}_s(\boldsymbol{z}_m))$. $\tilde{\pi}_s(\boldsymbol{z}_m)$ is evaluated at the current parameter estimates and $C(\tilde{\boldsymbol{\theta}})$ is constant. The algorithm works as follows: for each value of λ_P an outer loop is created that cycles over the classes s and computes l_{Qs} about the current parameter estimates $\tilde{\boldsymbol{\theta}}$. In the inner loop coordinate descent is applied to solve the penalized weighted least squares problem

$$\min_{(\beta_{0s}, \boldsymbol{\beta}_s) \in \mathbb{R}^{q+1}} \{ -l_{Qs}(\beta_{0s}, \boldsymbol{\beta}_s) + \lambda_P P_\alpha(\boldsymbol{\beta}_s) \},$$

which results in an update of the parameter estimates $\tilde{\boldsymbol{\theta}}$ and l_{Qs}. This procedure is repeated until convergence (Friedman, Hastie, and Tibshirani, 2010). $\lambda_P \geq 0$ is optimized by cross-validation, as described in section A.1.

A.3 Detailed MCMC algorithm for the Bayesian subgroup model

In the following, steps 1 to 4 of the MCMC sampling scheme in section 3.3.5.3 are explained in more detail.

Step 1: Update of $\boldsymbol{\Omega}_s$

The block Gibbs sampler proposed by Wang (2015) is used to update $\boldsymbol{\Omega}_s$ for subgroups $s = 1, ..., S$. The conditional distribution of $\boldsymbol{\Omega}_s$ is

$$\begin{aligned} p(\boldsymbol{\Omega}_s | \boldsymbol{G}_{ss}, \boldsymbol{X}_s) &\propto p(\boldsymbol{X}_s | \boldsymbol{\Omega}_s) \cdot p(\boldsymbol{\Omega}_s | \boldsymbol{G}_{ss}) \\ &\propto |\boldsymbol{\Omega}_s|^{n_s/2} \exp\{-\frac{1}{2} \operatorname{tr}(\boldsymbol{S}_s \boldsymbol{\Omega}_s)\} \cdot \prod_{i<j} \exp\{-\frac{1}{2} \frac{\omega_{s,ij}^2}{\nu_{g_{ss,ij}}^2}\} \cdot \prod_i \exp\{-\frac{\lambda}{2} \omega_{s,ii}\}. \end{aligned}$$

Consider the following partitions

$$\boldsymbol{\Omega}_s = \begin{pmatrix} \widetilde{\boldsymbol{\Omega}}_{11} & \widetilde{\boldsymbol{\omega}}_{12} \\ \widetilde{\boldsymbol{\omega}}_{12}' & \widetilde{\omega}_{22} \end{pmatrix} = \begin{pmatrix} \omega_{s,11} & \omega_{s,12} & \dots & \omega_{s,1(p-1)} & \omega_{s,1p} \\ \omega_{s,12} & \omega_{s,22} & \dots & \omega_{s,2(p-1)} & \omega_{s,2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \omega_{s,1(p-1)} & \omega_{s,2(p-1)} & \dots & \omega_{s,(p-1)(p-1)} & \omega_{s,(p-1)p} \\ \omega_{s,1p} & \omega_{s,2p} & \dots & \omega_{s,(p-1)p} & \omega_{s,pp} \end{pmatrix}$$

and analogously

$$\boldsymbol{S}_s = \boldsymbol{X}_s'\boldsymbol{X}_s = \begin{pmatrix} \widetilde{\boldsymbol{S}}_{11} & \widetilde{\boldsymbol{s}}_{12} \\ \widetilde{\boldsymbol{s}}_{12}' & \widetilde{s}_{22} \end{pmatrix}, \quad \boldsymbol{V}_s = (\nu^2_{g_{ss,ij}}) = \begin{pmatrix} \widetilde{\boldsymbol{V}}_{11} & \widetilde{\boldsymbol{v}}_{12} \\ \widetilde{\boldsymbol{v}}_{12}' & 0 \end{pmatrix},$$

where $\boldsymbol{V}_s$ is a $(p \times p)$ symmetric matrix with zeros on the diagonal. For the block update of $\boldsymbol{\Omega}_s$ focus on the last column (and row) of $\boldsymbol{\Omega}_s$: $(\widetilde{\boldsymbol{\omega}}_{12}, \widetilde{\omega}_{22})$ with $\widetilde{\boldsymbol{\omega}}_{12} = (\omega_{s,1p}, \omega_{s,2p}, ..., \omega_{s,(p-1)p})'$, $\widetilde{\omega}_{22} = \omega_{s,pp}$. The conditional distribution of the last column of $\boldsymbol{\Omega}_s$ is

$$p(\widetilde{\boldsymbol{\omega}}_{12}, \widetilde{\omega}_{22} | \boldsymbol{X}_s, \boldsymbol{G}_{ss}, \widetilde{\boldsymbol{\Omega}}_{11}) \propto \Big(\widetilde{\omega}_{22} - \widetilde{\boldsymbol{\omega}}_{12}'\widetilde{\boldsymbol{\Omega}}_{11}^{-1}\widetilde{\boldsymbol{\omega}}_{12}\Big)^{n_s/2} \cdot \exp\Big\{ -\frac{1}{2}\Big[\widetilde{\boldsymbol{\omega}}_{12}'\mathrm{diag}(\widetilde{\boldsymbol{v}}_{12}^{-1})\widetilde{\boldsymbol{\omega}}_{12} + 2\widetilde{\boldsymbol{s}}_{12}'\widetilde{\boldsymbol{\omega}}_{12} + (\widetilde{s}_{22}+\lambda)\widetilde{\omega}_{22}\Big]\Big\}.$$

Consider the following transformations

$$\boldsymbol{u} = \widetilde{\boldsymbol{\omega}}_{12}\,, \quad v = \widetilde{\omega}_{22} - \widetilde{\boldsymbol{\omega}}_{12}'\widetilde{\boldsymbol{\Omega}}_{11}^{-1}\widetilde{\boldsymbol{\omega}}_{12}\,.$$

Then the conditional distribution is

$$p(\boldsymbol{u}, v | \boldsymbol{X}_s, \boldsymbol{G}_{ss}, \widetilde{\boldsymbol{\Omega}}_{11}) \propto \underbrace{v^{n_s/2} \exp\Big\{ -\frac{\widetilde{s}_{22}+\lambda}{2} v \Big\}}_{(*_1)} \underbrace{\cdot \exp\Big\{ -\frac{1}{2}\Big[\boldsymbol{u}' \underbrace{\Big(\mathrm{diag}(\widetilde{\boldsymbol{v}}_{12}^{-1}) + (\widetilde{s}_{22}+\lambda)\widetilde{\boldsymbol{\Omega}}_{11}^{-1}\Big)}_{=\boldsymbol{C}^{-1}} \boldsymbol{u} + 2\widetilde{\boldsymbol{s}}_{12}'\boldsymbol{u}\Big]\Big\}}_{(*_2)}$$

$(*_1) \;\propto \mathcal{G}(v | \frac{n_s}{2}+1, \frac{\widetilde{s}_{22}+\lambda}{2})$,
$(*_2) \;\propto \mathcal{N}(\boldsymbol{u} | -\boldsymbol{C}\widetilde{\boldsymbol{s}}_{12}, \boldsymbol{C})$.
Permuting any column in $\boldsymbol{\Omega}_s$ to be updated to the last one leads to a block Gibbs sampler for the update of $\boldsymbol{\Omega}_s$.

Step 2: Update of $\boldsymbol{G}$

Update all elements in $\boldsymbol{G}$ iteratively with Gibbs sampler from their conditional distributions. All elements $g_{rs,ij}$ are assumed independent Bernoulli a priori with $p(g_{rs,ij} = 1) = \pi$ and $p(g_{rs,ij} = 0) = 1 - \pi$.

Update $g_{rs,ii}$ (edges between the same gene in different subgroups), $r, s = 1, ..., S$, $r < s$, $i = 1, ..., p$ from the conditional distribution

$$p(g_{rs,ii} | \boldsymbol{G}_{-rs,ii}, \boldsymbol{\gamma}) = \frac{p(g_{rs,ii}) \cdot p(\boldsymbol{\gamma} | \boldsymbol{G}_{-rs,ii}, g_{rs,ii})}{\sum_{g_{rs,ii} \in \{0,1\}} p(g_{rs,ii}) \cdot p(\boldsymbol{\gamma} | \boldsymbol{G}_{-rs,ii}, g_{rs,ii})}\,,$$

where $\boldsymbol{G}_{-rs,ii}$ denotes all elements in $\boldsymbol{G}$ except for $g_{rs,ii}$. Accept $g_{rs,ii} = 1$ with probability

$$p(g_{rs,ii} = 1|\boldsymbol{G}_{-rs,ii}, \boldsymbol{\gamma}) = \frac{w_a}{w_a + w_b},$$

where

$$\begin{aligned} w_a &= \pi \cdot \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})|_{g_{rs,ii}=1} \\ w_b &= (1-\pi) \cdot \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})|_{g_{rs,ii}=0} \,. \end{aligned}$$

This means, update $g_{rs,ii}$ as follows: $g_{rs,ii} = \begin{cases} 1, & \text{if } u < \frac{w_a}{w_a+w_b},\ u \sim \mathcal{U}[0,1] \\ 0, & \text{else}\,. \end{cases}$

Update $g_{ss,ij}$ (edges between different genes in the same subgroup) , $s = 1, ..., S$, $i, j = 1, ..., p$, $i < j$ from the conditional distribution

$$\begin{aligned} p(g_{ss,ij}|\boldsymbol{G}_{-ss,ij}, \omega_{s,ij}, \boldsymbol{\gamma}) &= \frac{p(g_{ss,ij}) \cdot p(\omega_{s,ij}, \boldsymbol{\gamma}|\boldsymbol{G}_{-ss,ij}, g_{ss,ij})}{\sum_{g_{ss,ij}\in\{0,1\}} p(g_{ss,ij}) \cdot p(\omega_{s,ij}, \boldsymbol{\gamma}|\boldsymbol{G}_{-ss,ij}, g_{ss,ij})} \\ &\propto p(g_{ss,ij}) \cdot p(\omega_{s,ij}|g_{ss,ij}) \cdot p(\boldsymbol{\gamma}|\boldsymbol{G}_{-ss,ij}, g_{ss,ij})\,, \end{aligned}$$

where $\boldsymbol{G}_{-ss,ij}$ denotes all elements in $\boldsymbol{G}$ except for $g_{-ss,ij}$. Accept $g_{ss,ij} = 1$ with probability

$$p(g_{ss,ij} = 1|\boldsymbol{G}_{-ss,ij}, \omega_{s,ij}, \boldsymbol{\gamma}) = \frac{w_a}{w_a + w_b},$$

where

$$\begin{aligned} w_a &= \pi \cdot \mathcal{N}(\omega_{s,ij}|0, \nu_1^2) \cdot \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'G\boldsymbol{\gamma})|_{g_{ss,ij}=1} \\ w_b &= (1-\pi) \cdot \mathcal{N}(\omega_{s,ij}|0, \nu_0^2) \cdot \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'G\boldsymbol{\gamma})|_{g_{ss,ij}=0}. \end{aligned}$$

Step 3: Update of $\boldsymbol{\gamma}$

Update $\gamma_{s,i}$, $s = 1, ..., S$, $i = 1, ..., p$, with Gibbs sampler from the conditional distribution

$$\begin{aligned} p(\gamma_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}, \beta_{s,i}) &= \frac{p(\gamma_{s,i}, \beta_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G})}{\sum_{\gamma_{s,i}\in\{0,1\}} p(\gamma_{s,i}, \beta_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G})} \\ &= \frac{p(\gamma_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}) \cdot p(\beta_{s,i}|\gamma_{s,i}, \boldsymbol{\gamma}_{-s,i}, \boldsymbol{G})}{\sum_{\gamma_{s,i}\in\{0,1\}} p(\gamma_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}) \cdot p(\beta_{s,i}|\gamma_{s,i}, \boldsymbol{\gamma}_{-s,i}, \boldsymbol{G})} \\ &= \frac{p(\gamma_{s,i}, \boldsymbol{\gamma}_{-s,i}|\boldsymbol{G}) \cdot p(\beta_{s,i}|\gamma_{s,i})}{\sum_{\gamma_{s,i}\in\{0,1\}} p(\gamma_{s,i}, \boldsymbol{\gamma}_{-s,i}|\boldsymbol{G}) \cdot p(\beta_{s,i}|\gamma_{s,i})}\,, \end{aligned}$$

where $\boldsymbol{\gamma}_{-s,i}$ denotes all elements in $\boldsymbol{\gamma}$ except for $\gamma_{s,i}$. Accept $\gamma_{s,i} = 1$ with probability

$$p(\gamma_{s,i} = 1|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}, \beta_{s,i}) = \frac{w_a}{w_a + w_b},$$

where

$$\begin{aligned} w_a &= \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})|_{\gamma_{s,i}=1} \cdot \mathcal{N}(\beta_{s,i}|0, c^2\tau^2) \\ w_b &= \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})|_{\gamma_{s,i}=0} \cdot \mathcal{N}(\beta_{s,i}|0, \tau^2). \end{aligned}$$

Step 4: Update of $\boldsymbol{\beta}$

A random walk Metropolis-Hastings algorithm with adaptive jumping rule as proposed by Lee, Chakraborty, and Sun (2011) is used to update $\beta_{s,i}$ for $s = 1, ..., S$ and $i = 1, ..., p$. The full conditional posterior distribution of $\beta_{s,i}$ is

$$\begin{aligned} & p(\beta_{s,i}|\boldsymbol{\beta}_{s,-i}, \boldsymbol{\gamma}_s, \boldsymbol{h}_s, \mathfrak{D}_s) \\ & \quad \propto \; L(\mathfrak{D}_s|\boldsymbol{\beta}_s, \boldsymbol{h}_s) \cdot p(\boldsymbol{\beta}_s|\boldsymbol{\gamma}_s) \\ & \quad \propto \; \prod_{g=1}^{J_s} \left[\exp\Big(- h_{s,g} \sum_{k \in \mathcal{R}_{s,g} - \mathcal{D}_{s,g}} \exp(\boldsymbol{\beta}'_s \boldsymbol{x}_{s,k}) \Big) \prod_{l \in \mathcal{D}_{s,g}} \Big[1 - \exp\Big(- h_{s,g} \exp(\boldsymbol{\beta}'_s \boldsymbol{x}_{s,l}) \Big) \Big] \right] \\ & \qquad \cdot \exp\Big(- \frac{1}{2} \boldsymbol{\beta}'_s \Sigma_{\beta_s}^{-1} \boldsymbol{\beta}_s \Big), \end{aligned}$$

where $\boldsymbol{\beta}_{s,-i}$ denotes the vector $\boldsymbol{\beta}_s$ without the i-th element and $\Sigma_{\beta_s} = \text{diag}(\sigma^2_{\beta_{s,1}}, ..., \sigma^2_{\beta_{s,p}})$ with $\sigma^2_{\beta_{s,i}} = (1 - \gamma_{s,i}) \cdot \tau^2 + \gamma_{s,i} \cdot c^2 \tau^2$.

In MCMC iteration t update $\beta_{s,i}$ as follows:

(i) Sample a proposal $\beta^{(prop)}_{s,i}$ from a proposal distribution $q(\beta^{(prop)}_{s,i}|\beta^{(t-1)}_{s,i}) = \mathcal{N}(\beta^{(prop)}_{s,i}|\mu^{(t-1)}_{\beta_{s,i}}, \nu^{(t-1)}_{\beta_{s,i}})$

(ii) Calculate the ratio of ratios

$$r_{s,i} = \frac{p(\beta^{(prop)}_{s,i}|\boldsymbol{\beta}^{(t-1)}_{s,-i}, \boldsymbol{\gamma}^{(t-1)}_s, \boldsymbol{h}^{(t-1)}_s, \mathfrak{D}_s)/q(\beta^{(prop)}_{s,i}|\beta^{(t-1)}_{s,i})}{p(\beta^{(t-1)}_{s,i}|\boldsymbol{\beta}^{(t-1)}_{s,-i}, \boldsymbol{\gamma}^{(t-1)}_s, \boldsymbol{h}^{(t-1)}_s, \mathfrak{D}_s)/q(\beta^{(t-1)}_{s,i}|\beta^{(prop)}_{s,i})}$$

(iii) Accept the proposal $\beta^{(prop)}_{s,i}$ if $\min\{r_{s,i}, 1\} > u$ with $u \sim \mathcal{U}[0, 1]$.

The mean and variance of the proposal distribution can be approximated based on the first and second derivative of the log conditional posterior distribution with respect to $\beta^{(t-1)}_{s,i}$.

Appendix B

Figures

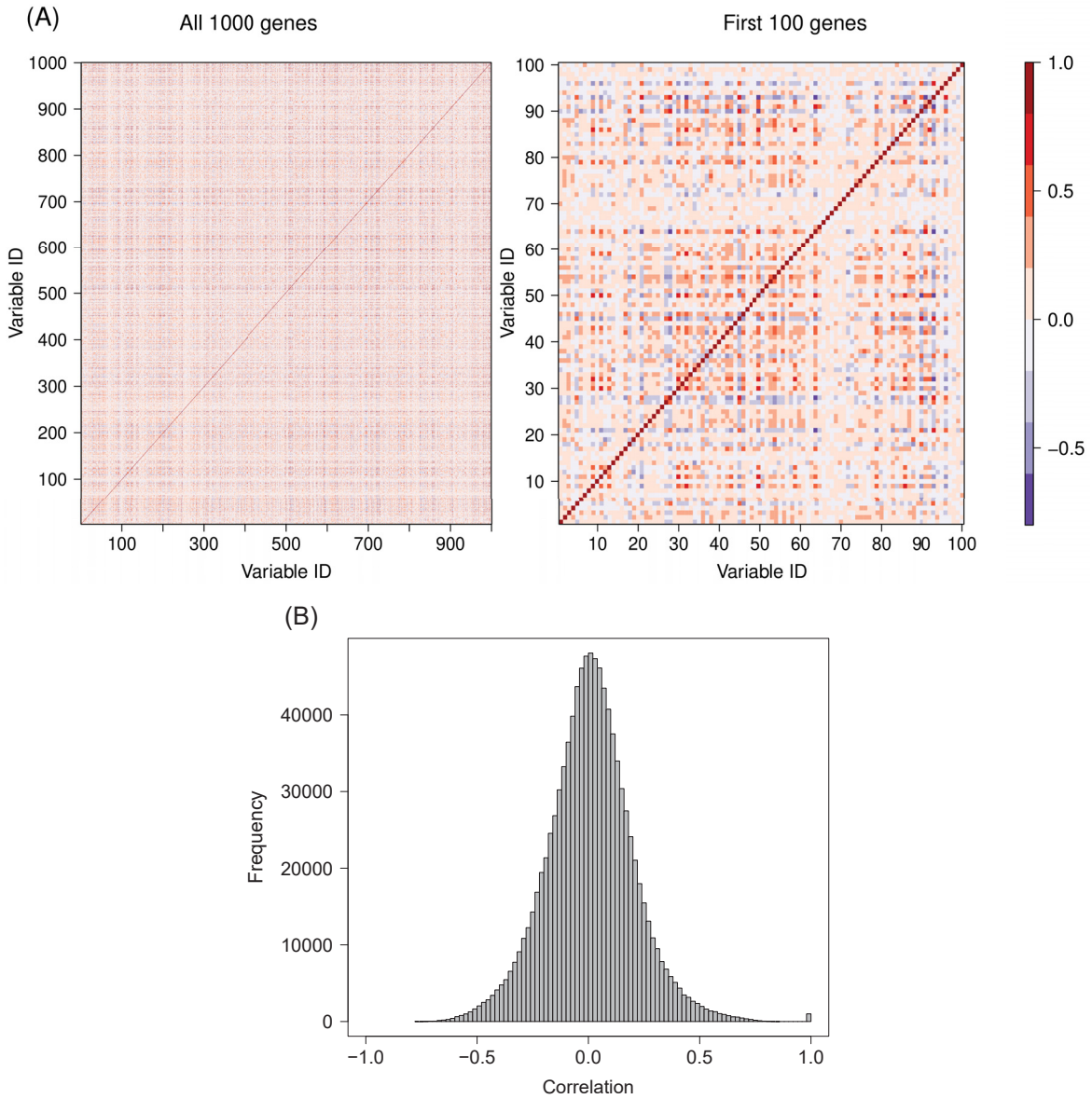


FIGURE B.1: *Shrinkage estimates of the empirical correlation matrix from 1000 randomly selected genes. (A) shrinkage correlation matrix, (B) histogram of the distribution of shrinkage correlation values.*

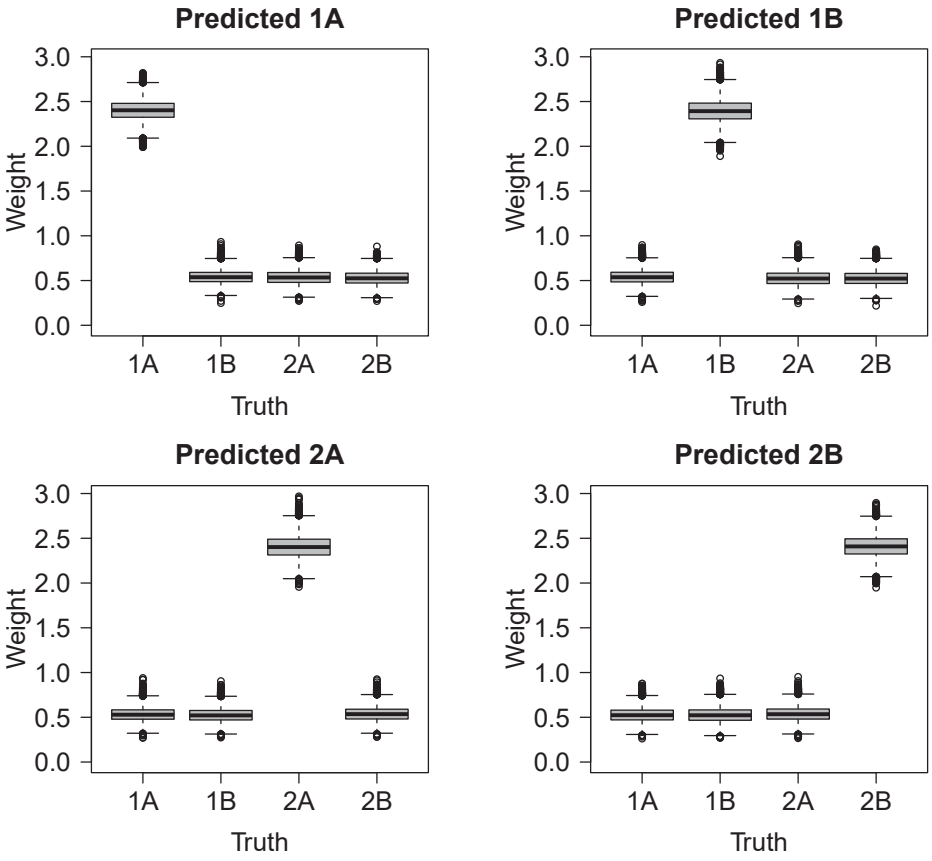


FIGURE B.2: *Weights estimated with random forest without interactions and cumulative HR, for simulated data with $n = 200$, $p = 100$, block correlation and $\epsilon = 0$. Estimation is based on training data without cross-validation which leads to overfitting.*

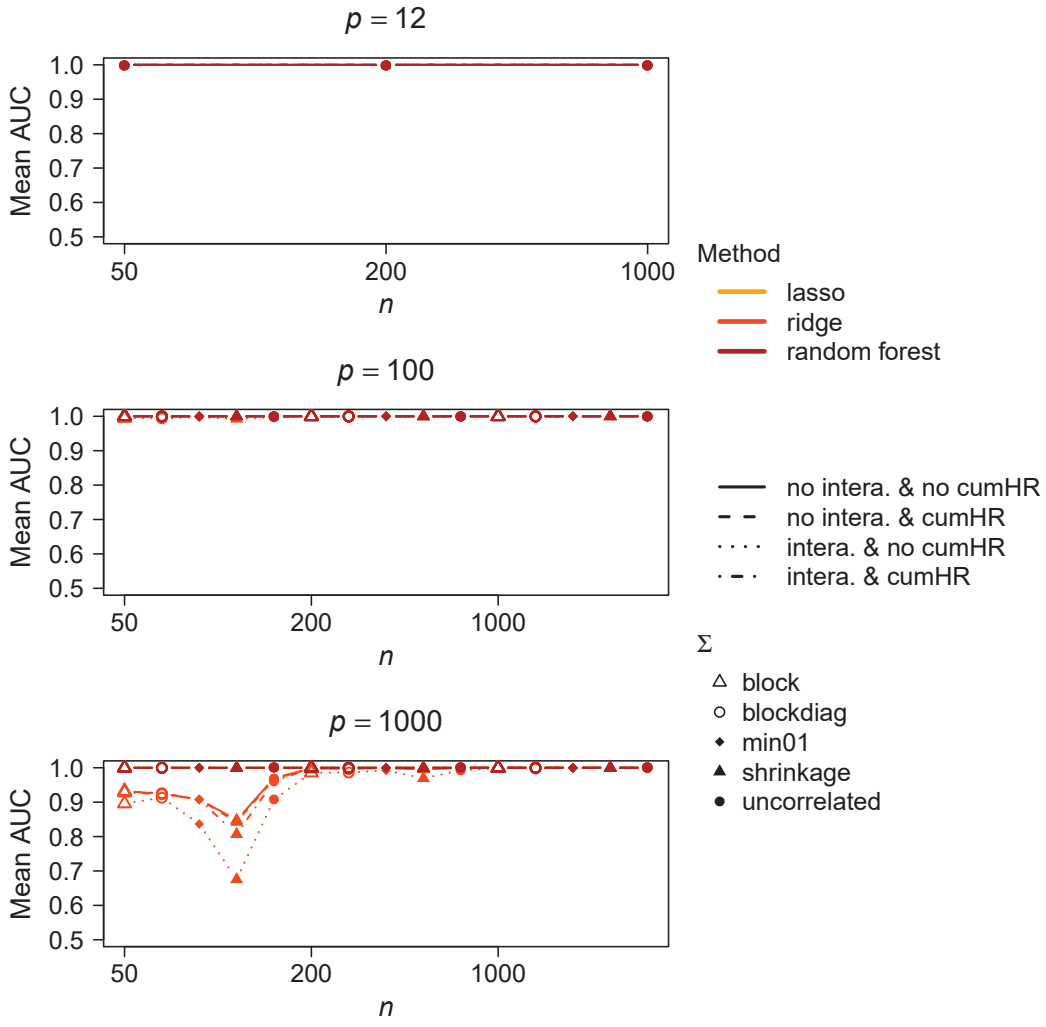


FIGURE B.3: *Mean AUC for weights estimation based on cross-validated training sets. Comparison of different parameters (line type) and methods (color) for weights estimation, and varying parameters for data simulation (p, n, $\mathbf{\Sigma}$) with $\epsilon = 1$.*

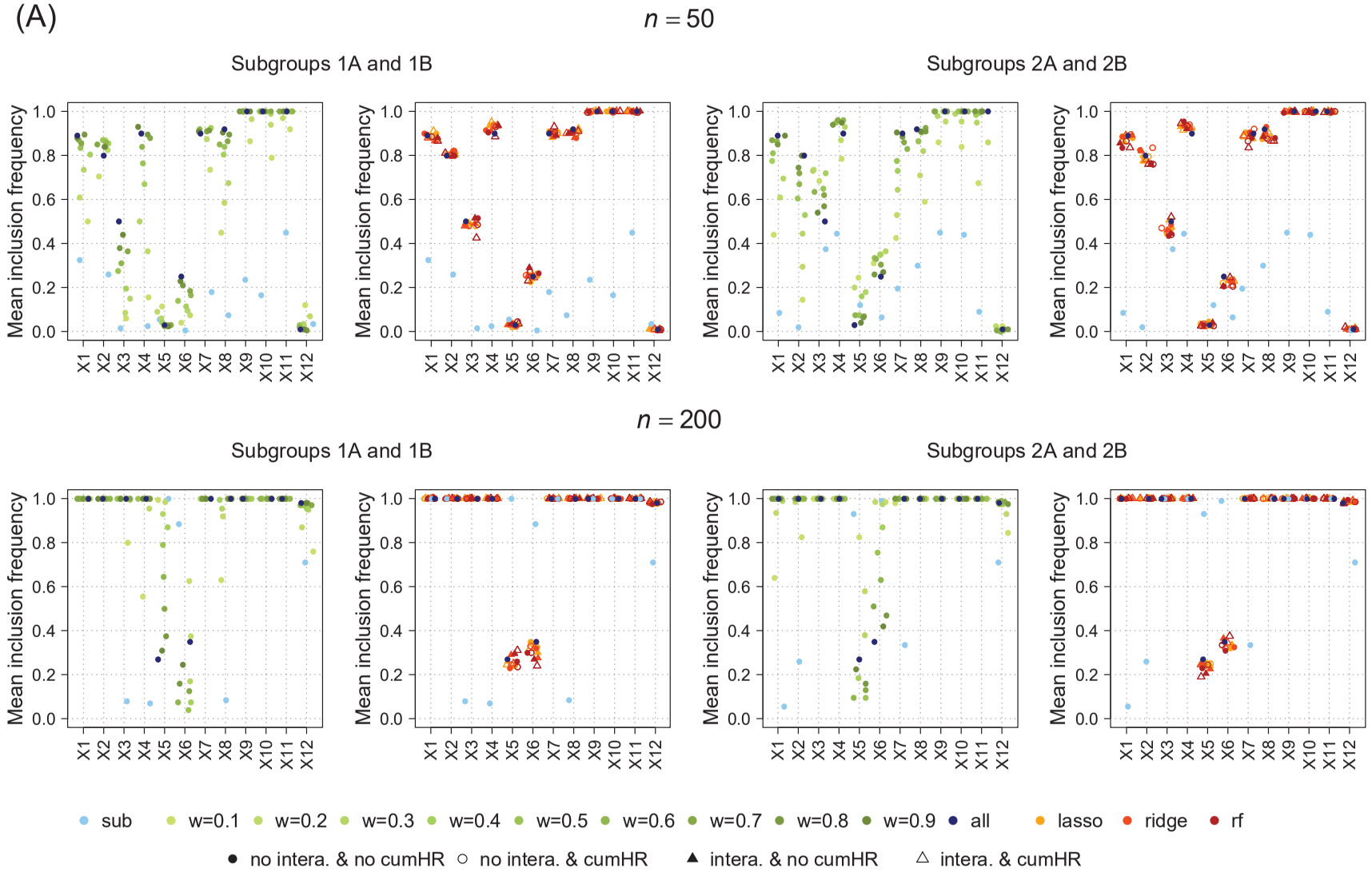


FIGURE B.4: *Mean inclusion frequencies of the Cox model for different model types (colors) and parameter settings for weights estimation (point symbols). Results are based on simulated data with $p = 100$ uncorrelated predictors, $n = 50, 200$, and (A) $\epsilon = 0$, (B) $\epsilon = 1$.*

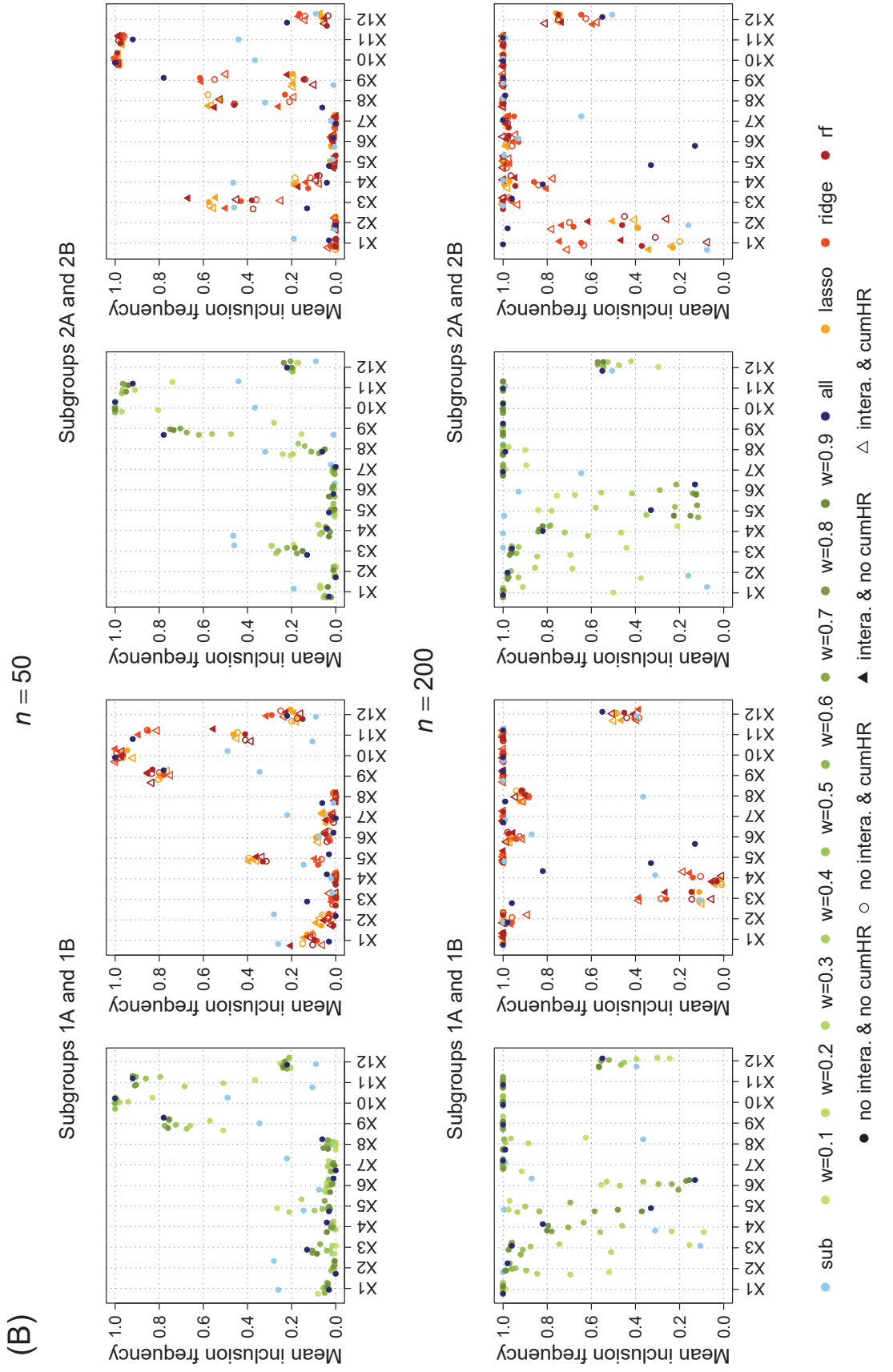


FIGURE B.4: *Mean inclusion frequencies of the Cox model for different model types (colors) and parameter settings for weights estimation (point symbols). Results are based on simulated data with* $p = 100$ *uncorrelated predictors,* $n = 50, 200$*, and (A)* $\epsilon = 0$*, (B)* $\epsilon = 1$ *(cont.).*

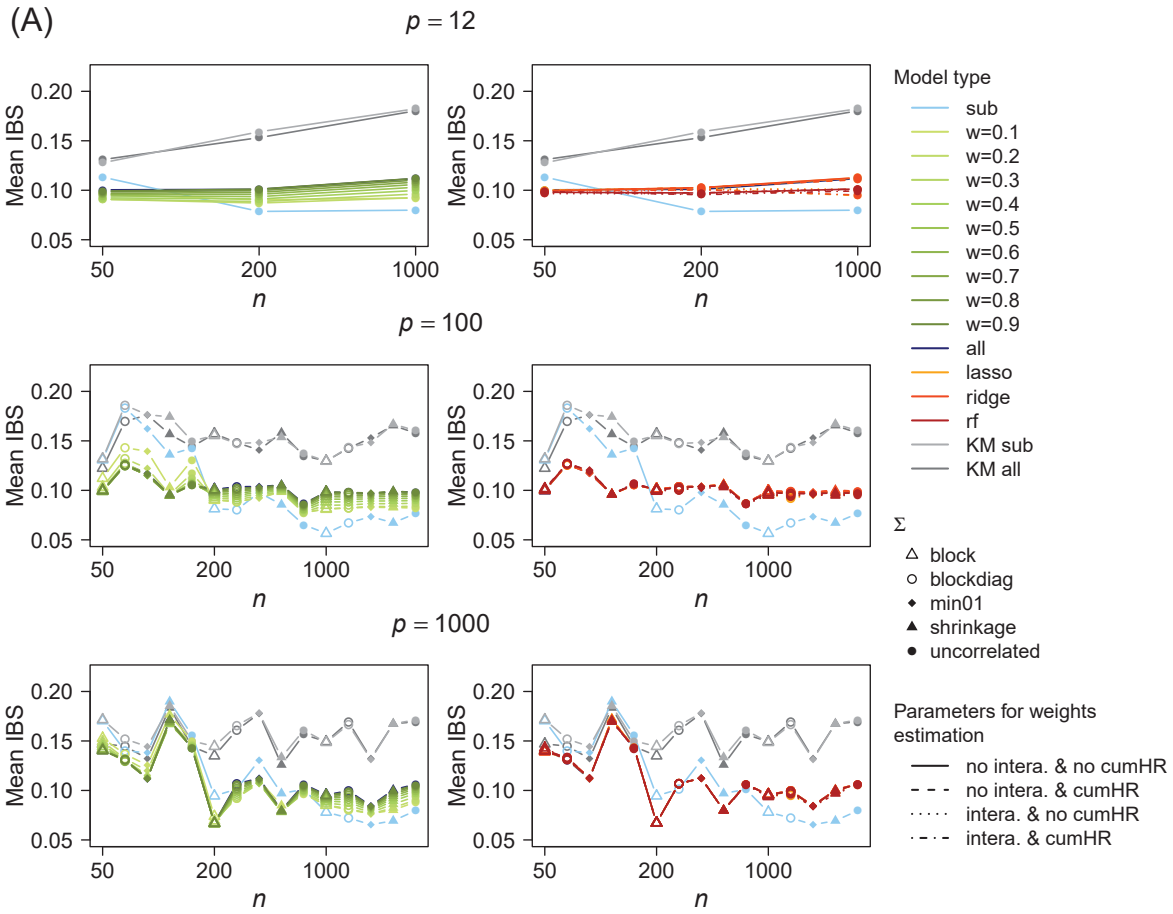


FIGURE B.5: *Mean integrated Brier score (IBS) for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$). Gray lines indicate the Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (sub) or combined (all) training data. (A) $\epsilon = 0$, (B) $\epsilon = 1$.*

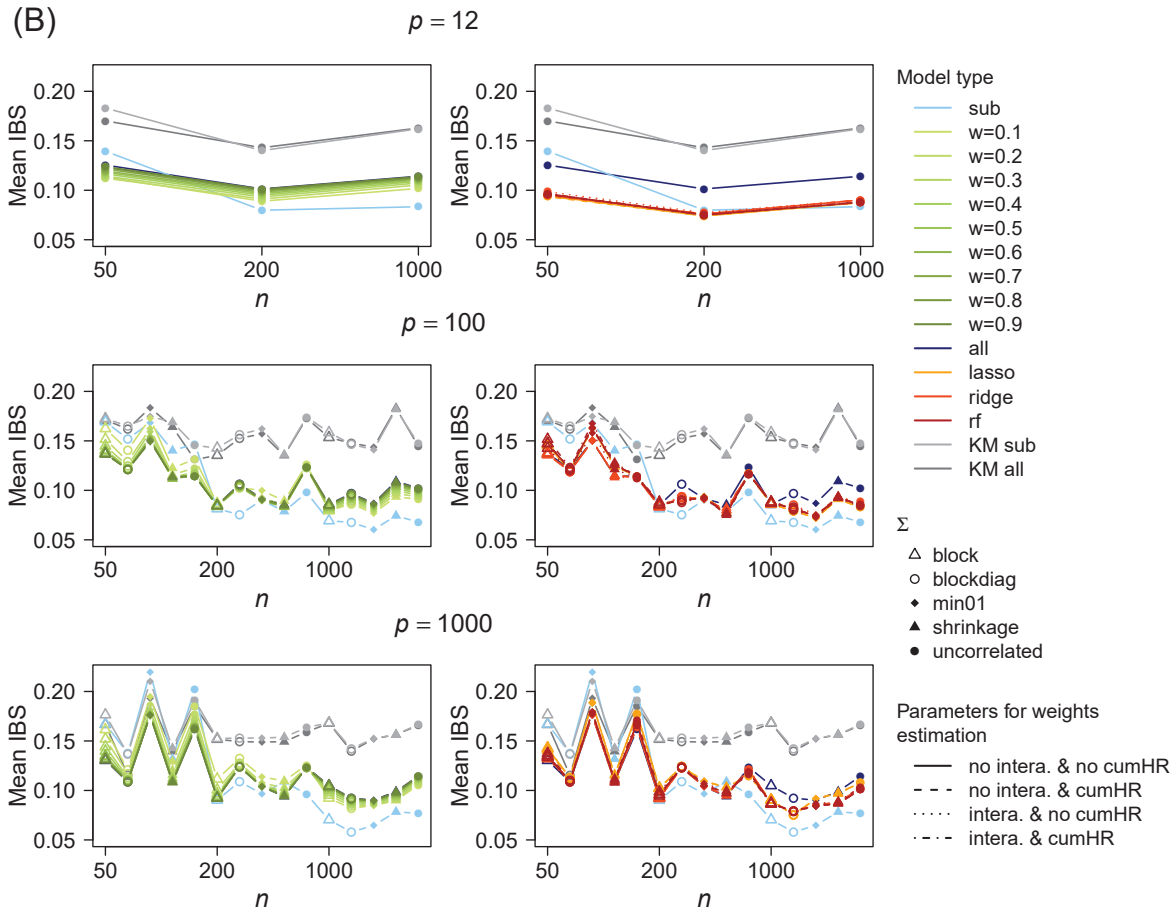


FIGURE B.5: *Mean integrated Brier score (IBS) for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$). Gray lines indicate the Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (sub) or combined (all) training data. (A) $\epsilon = 0$, (B) $\epsilon = 1$ (cont.).*

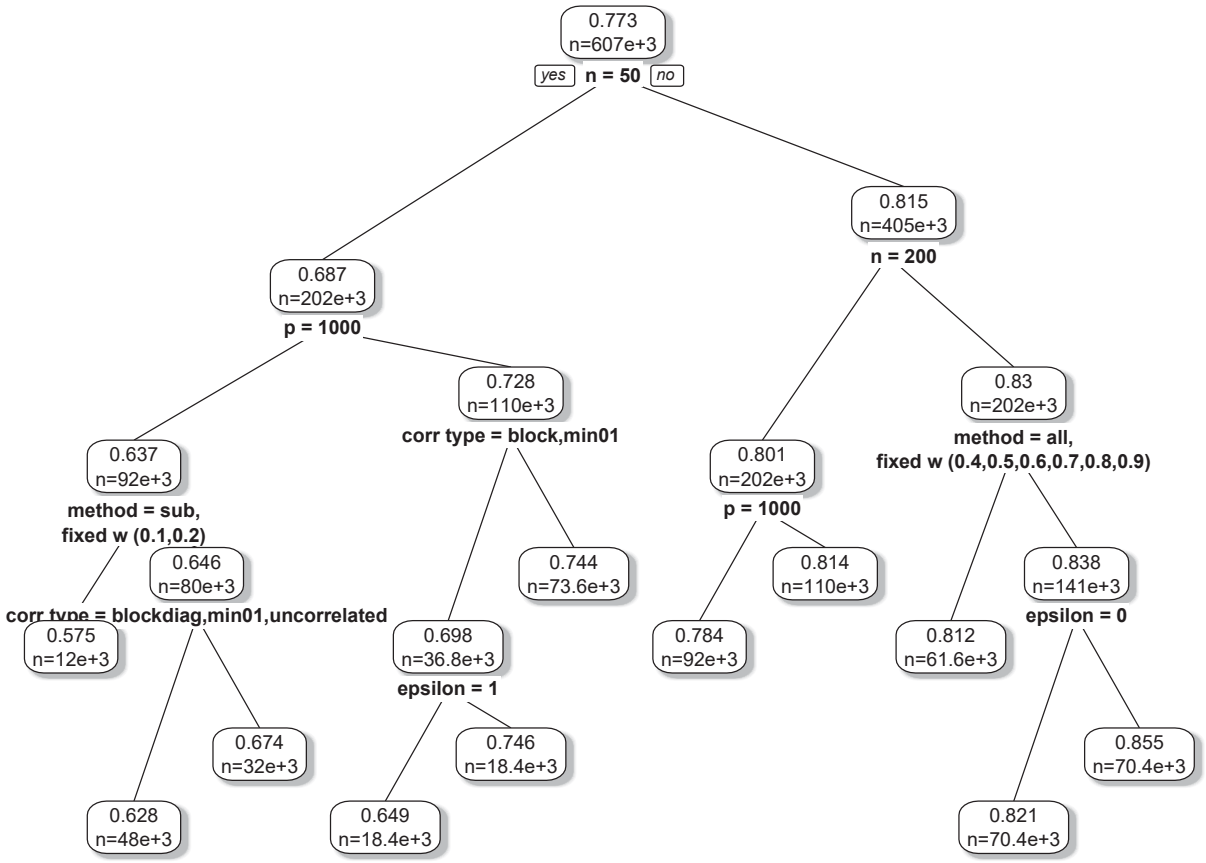


FIGURE B.6: *Regression tree for C-index based on all test sets including all model types and parameter settings for data simulation and weights estimation as predictors. Each box shows mean C-index and sample size in the corresponding node.*

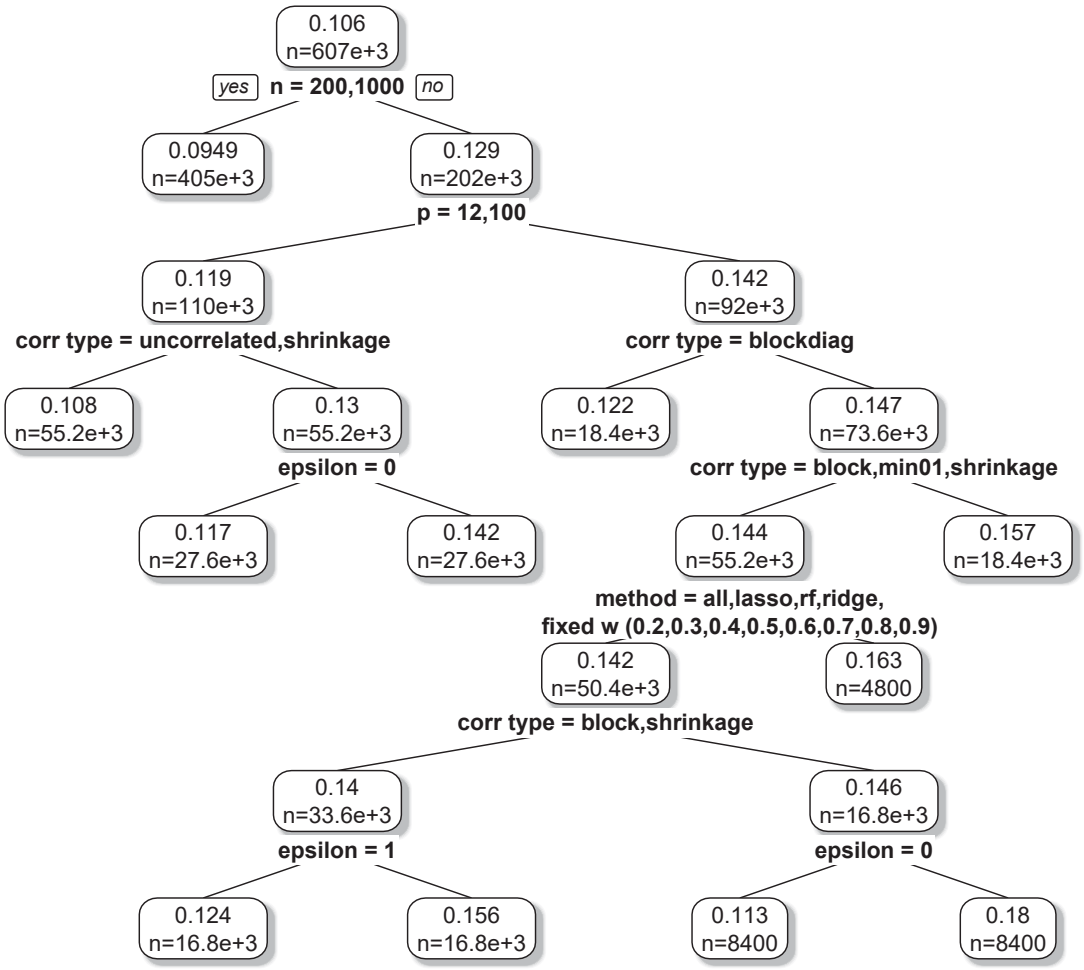


FIGURE B.7: *Regression tree for integrated Brier score (IBS) based on all test sets including all model types and parameter settings for data simulation and weights estimation as predictors. Each box shows mean IBS and sample size in the corresponding node.*

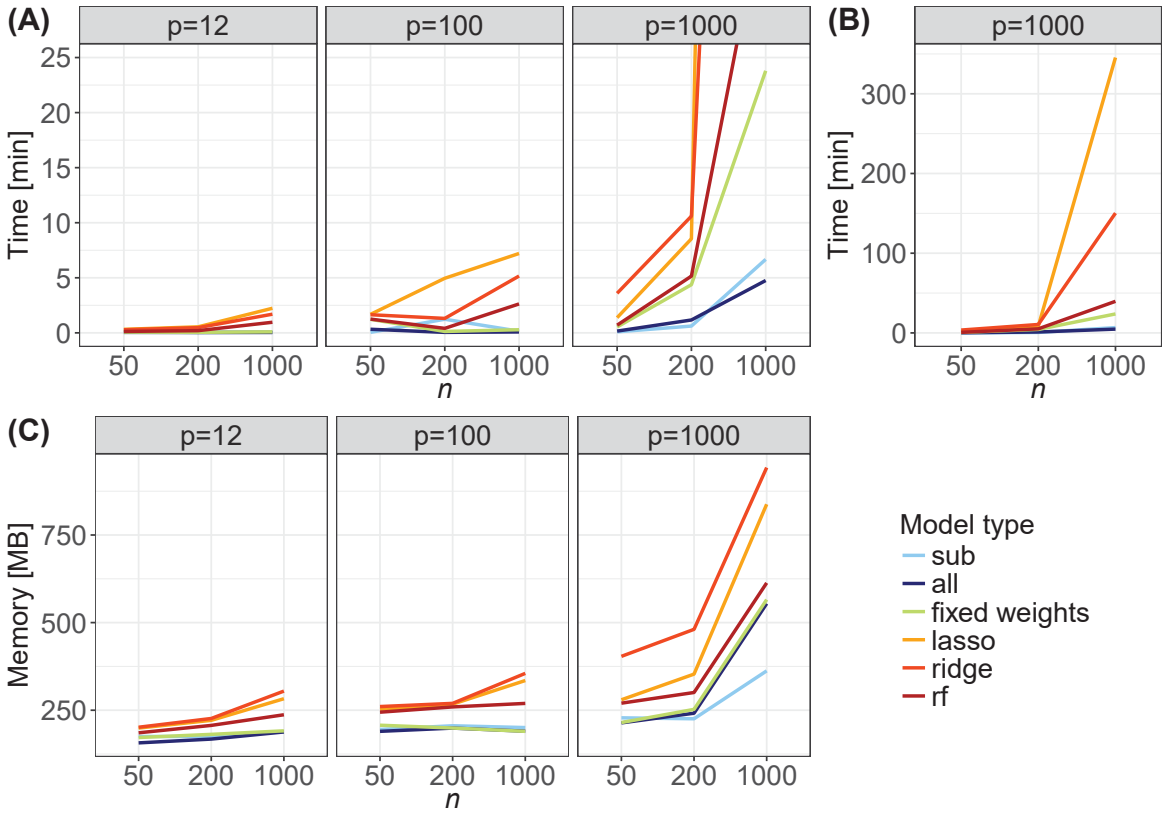


FIGURE B.8: *(A)+(B) Mean computation time, and (C) mean memory for different model types (without interactions and cumulative HR), sample sizes n and number of covariates p. (A) Limitation of y-axis with close-up for $p = 1000$; (B) Complete y-axis range for $p = 1000$. All algorithms were run on the same high performance computing cluster (LiDOng system) at the TU Dortmund University.*

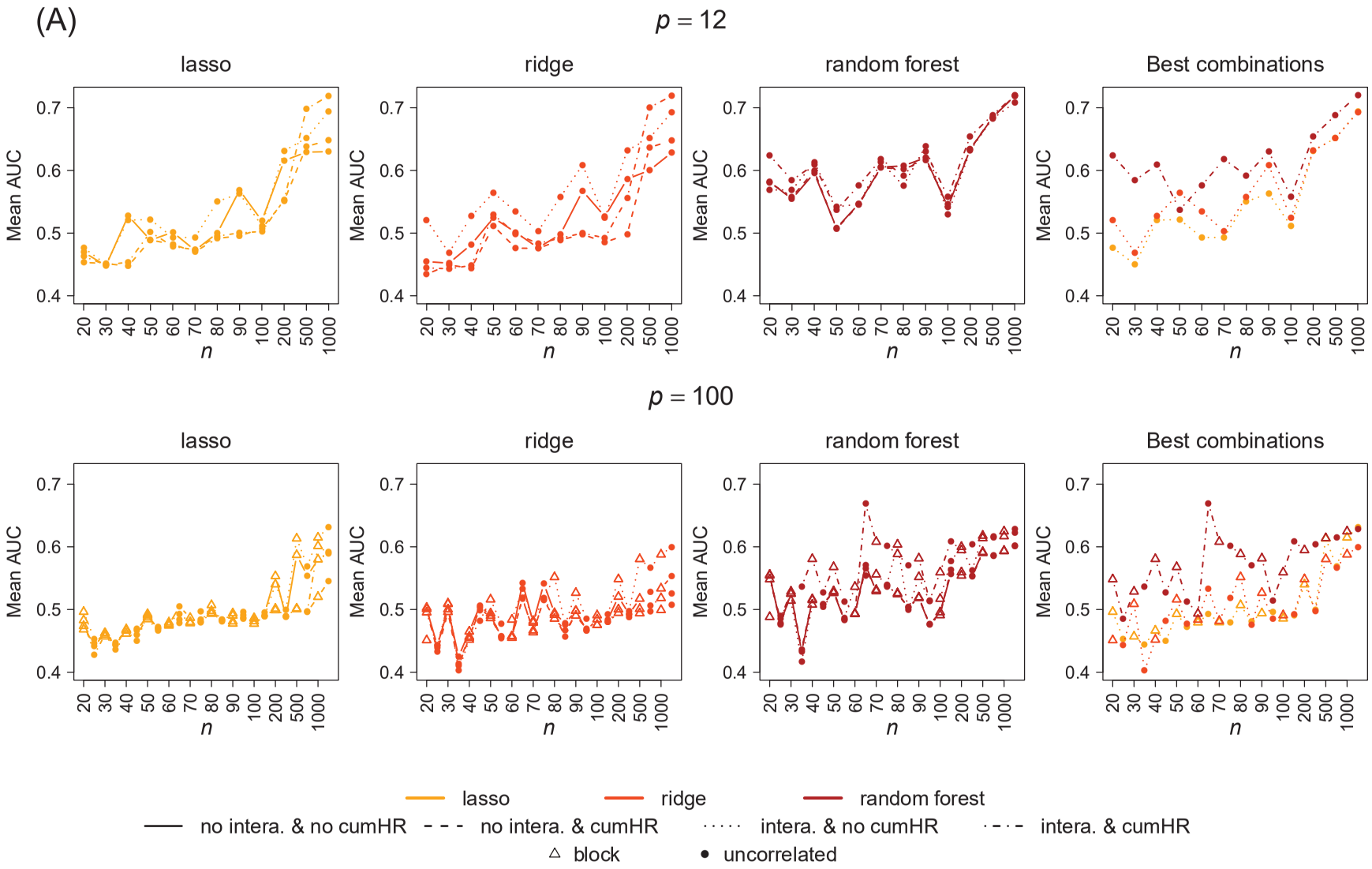


FIGURE B.9: *Mean AUC for weights estimation based on cross-validated training sets. Comparison of different parameters (line type) and methods (color) for weights estimation, and varying parameters for data simulation (p, n, $\mathbf{\Sigma}$). The first three columns show all parameter combinations for the three classification methods, and the fourth column for each method the best combination (mean across different n and $\mathbf{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$.*

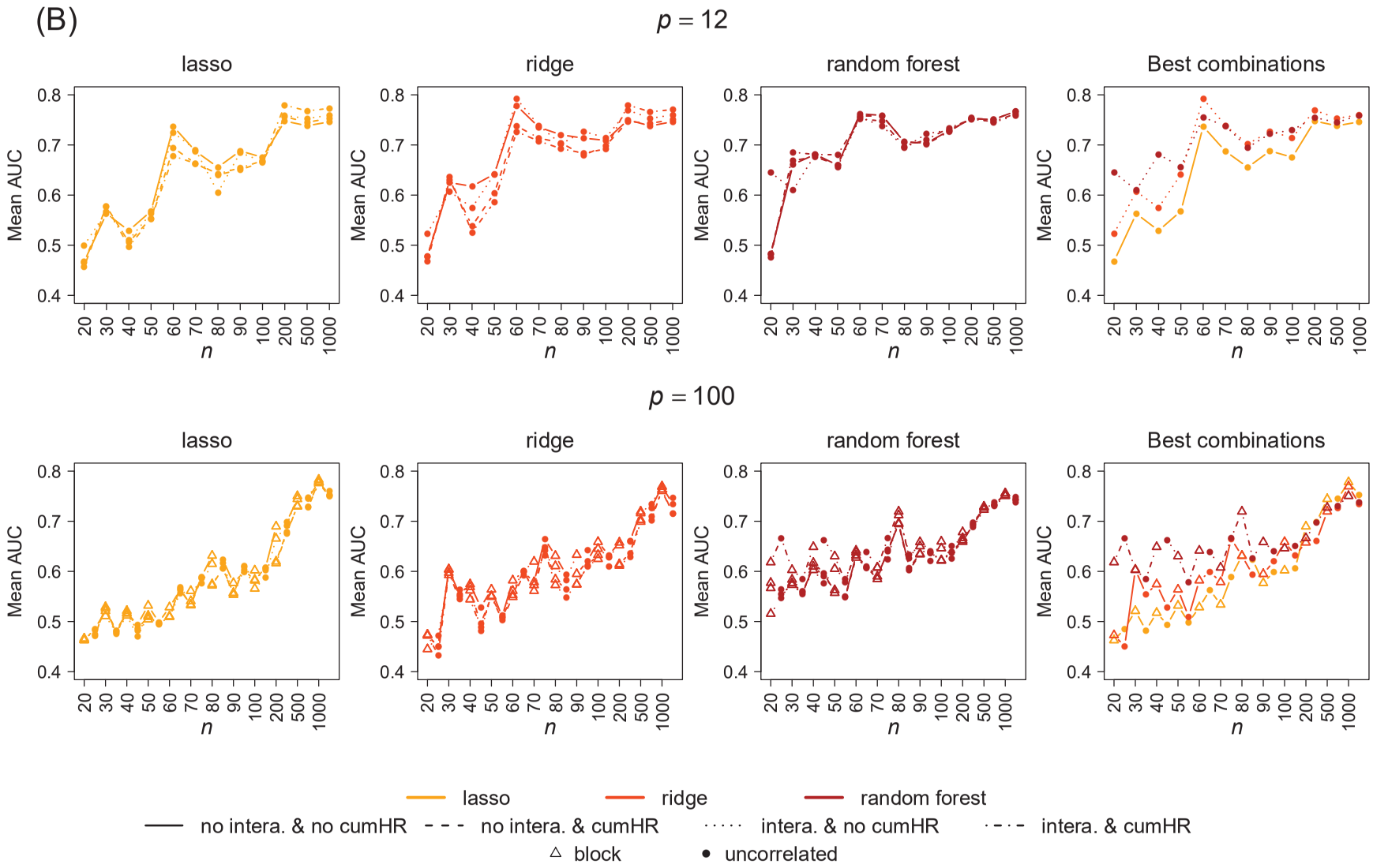


FIGURE B.9: *Mean AUC for weights estimation based on cross-validated training sets. Comparison of different parameters (line type) and methods (color) for weights estimation, and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$). The first three columns show all parameter combinations for the three classification methods, and the fourth column for each method the best combination (mean across different n and $\boldsymbol{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$ (cont.).*

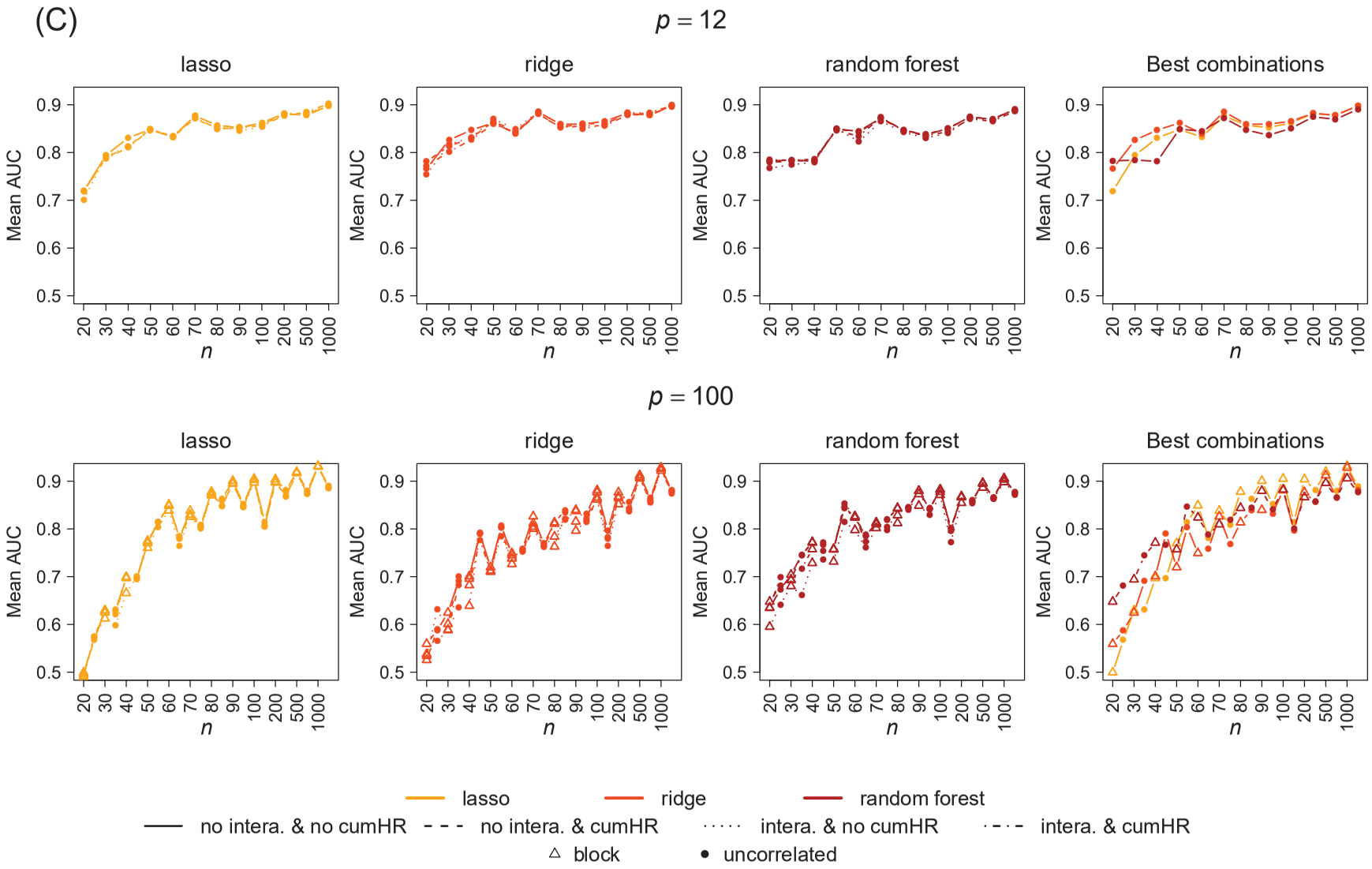


FIGURE B.9: *Mean AUC for weights estimation based on cross-validated training sets. Comparison of different parameters (line type) and methods (color) for weights estimation, and varying parameters for data simulation (p, n, $\mathbf{\Sigma}$). The first three columns show all parameter combinations for the three classification methods, and the fourth column for each method the best combination (mean across different n and $\mathbf{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$ (cont.).*

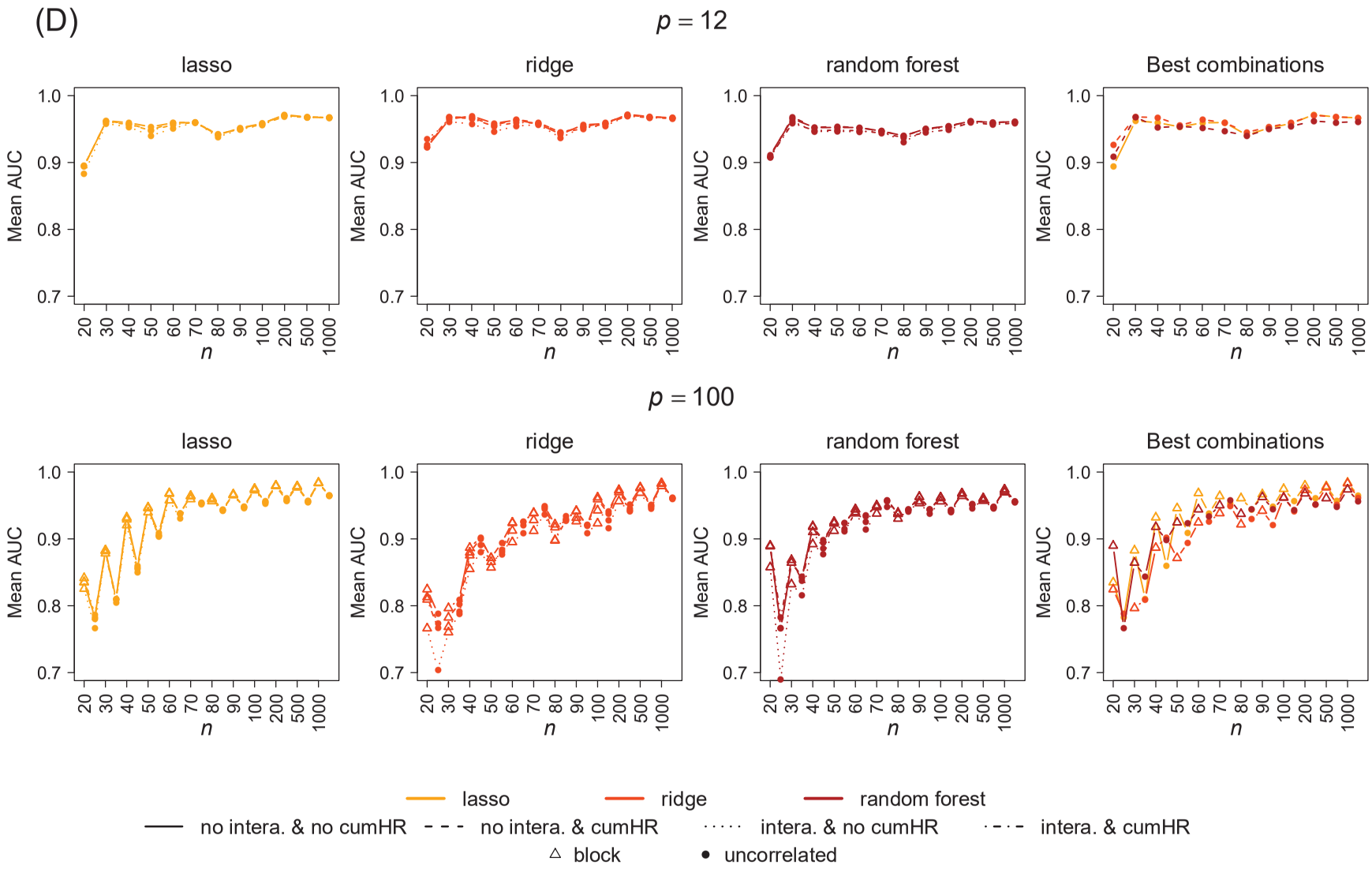


FIGURE B.9: *Mean AUC for weights estimation based on cross-validated training sets. Comparison of different parameters (line type) and methods (color) for weights estimation, and varying parameters for data simulation (p, n, $\mathbf{\Sigma}$). The first three columns show all parameter combinations for the three classification methods, and the fourth column for each method the best combination (mean across different n and $\mathbf{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$ (cont.).*

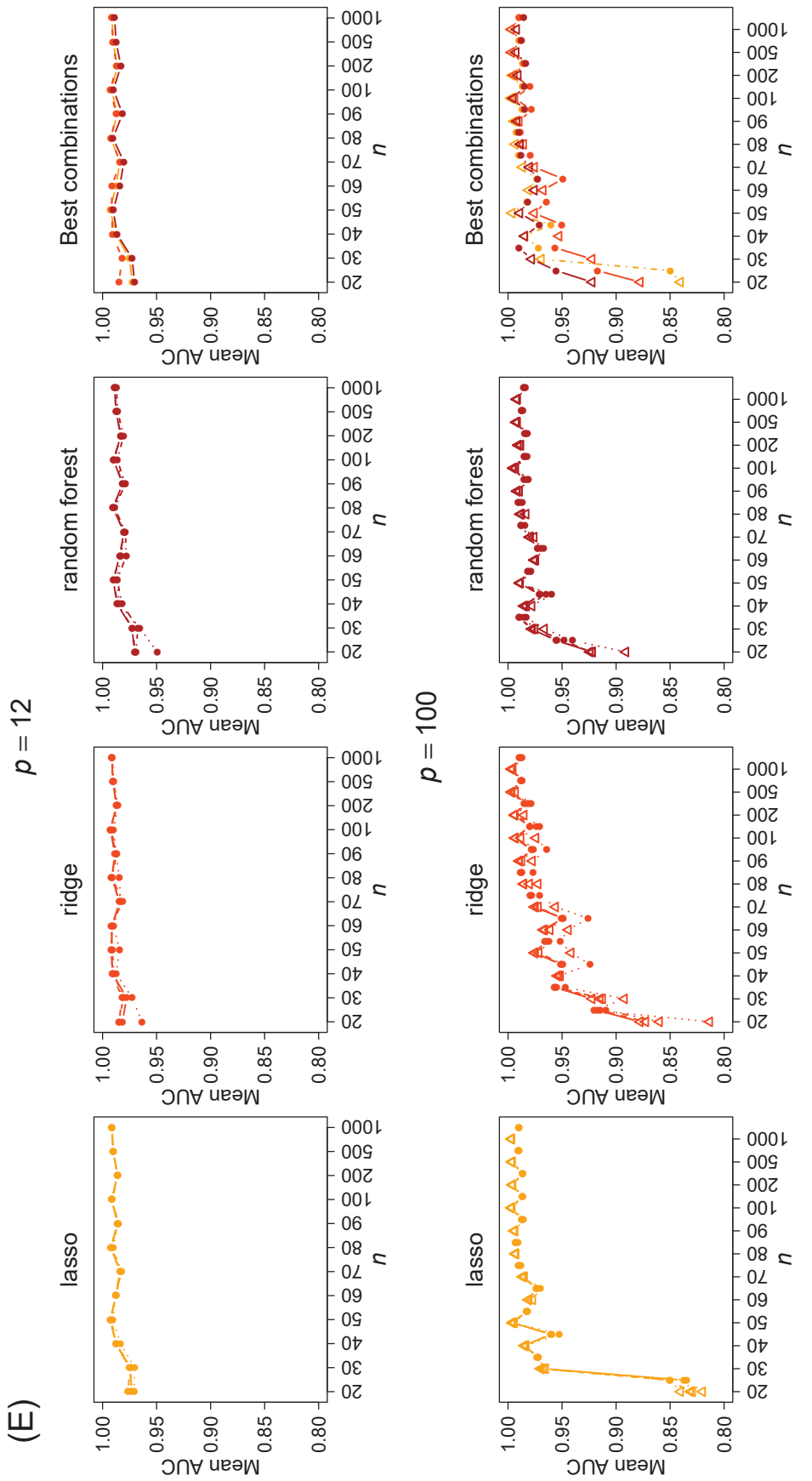


FIGURE B.9: *Mean AUC for weights estimation based on cross-validated training sets. Comparison of different parameters (line type) and methods (color) for weights estimation, and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$). The first three columns show all parameter combinations for the three classification methods, and the fourth column for each method the best combination (mean across different n and $\boldsymbol{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$ (cont.).*

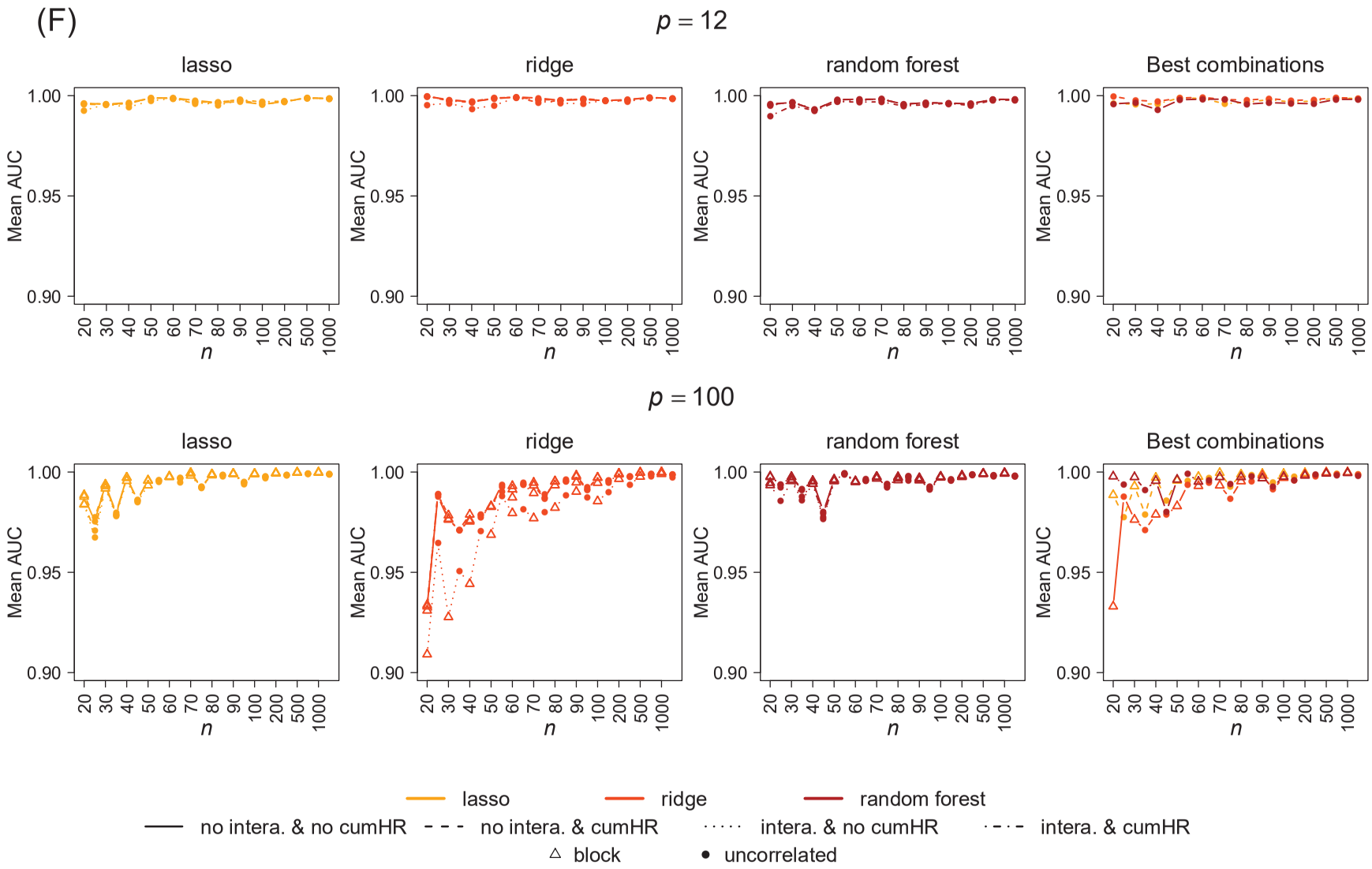


FIGURE B.9: *Mean AUC for weights estimation based on cross-validated training sets. Comparison of different parameters (line type) and methods (color) for weights estimation, and varying parameters for data simulation (p, n, $\mathbf{\Sigma}$). The first three columns show all parameter combinations for the three classification methods, and the fourth column for each method the best combination (mean across different n and $\mathbf{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$ (cont.).*

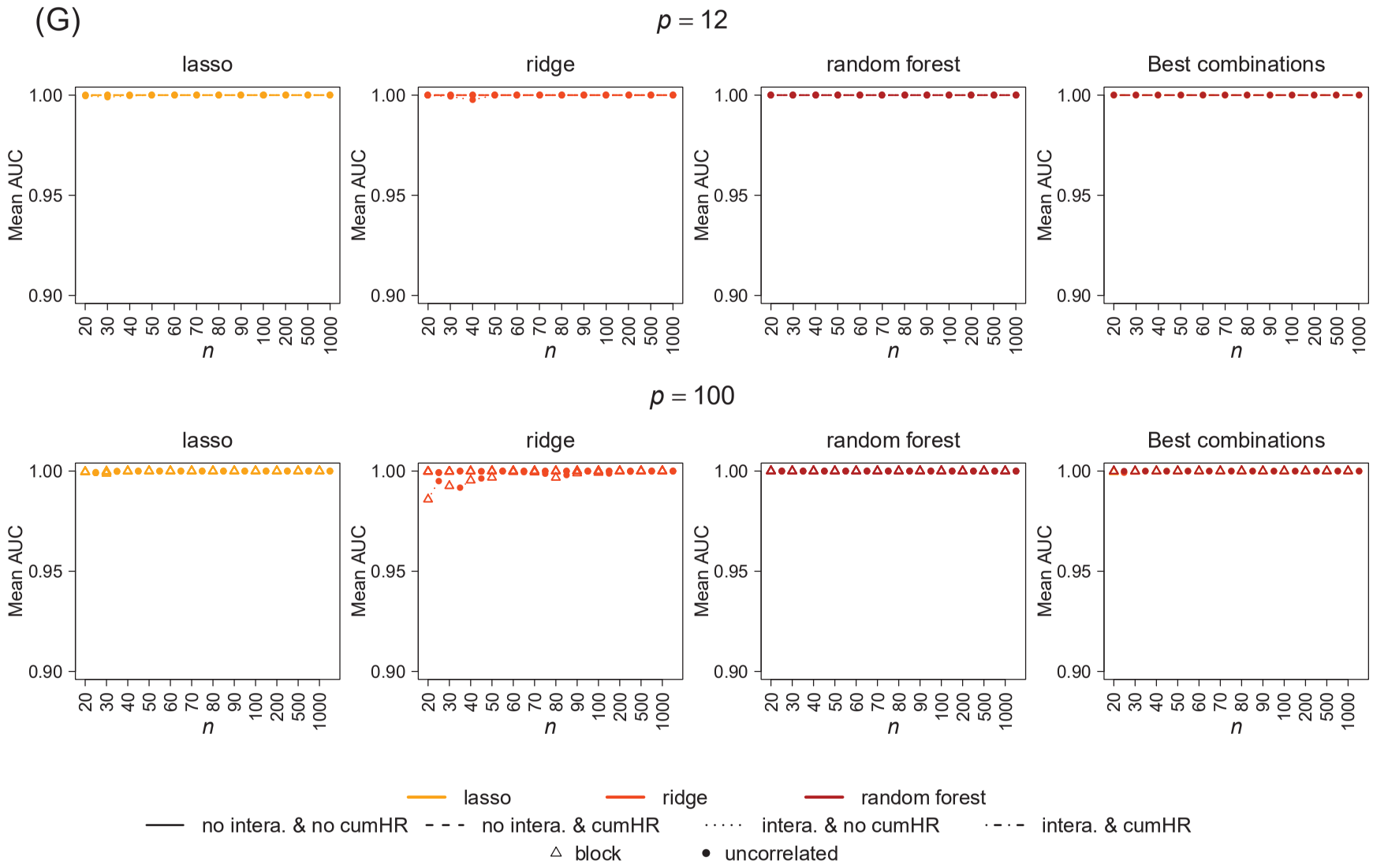


FIGURE B.9: *Mean AUC for weights estimation based on cross-validated training sets. Comparison of different parameters (line type) and methods (color) for weights estimation, and varying parameters for data simulation (p, n, $\mathbf{\Sigma}$). The first three columns show all parameter combinations for the three classification methods, and the fourth column for each method the best combination (mean across different n and $\mathbf{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$ (cont.).*

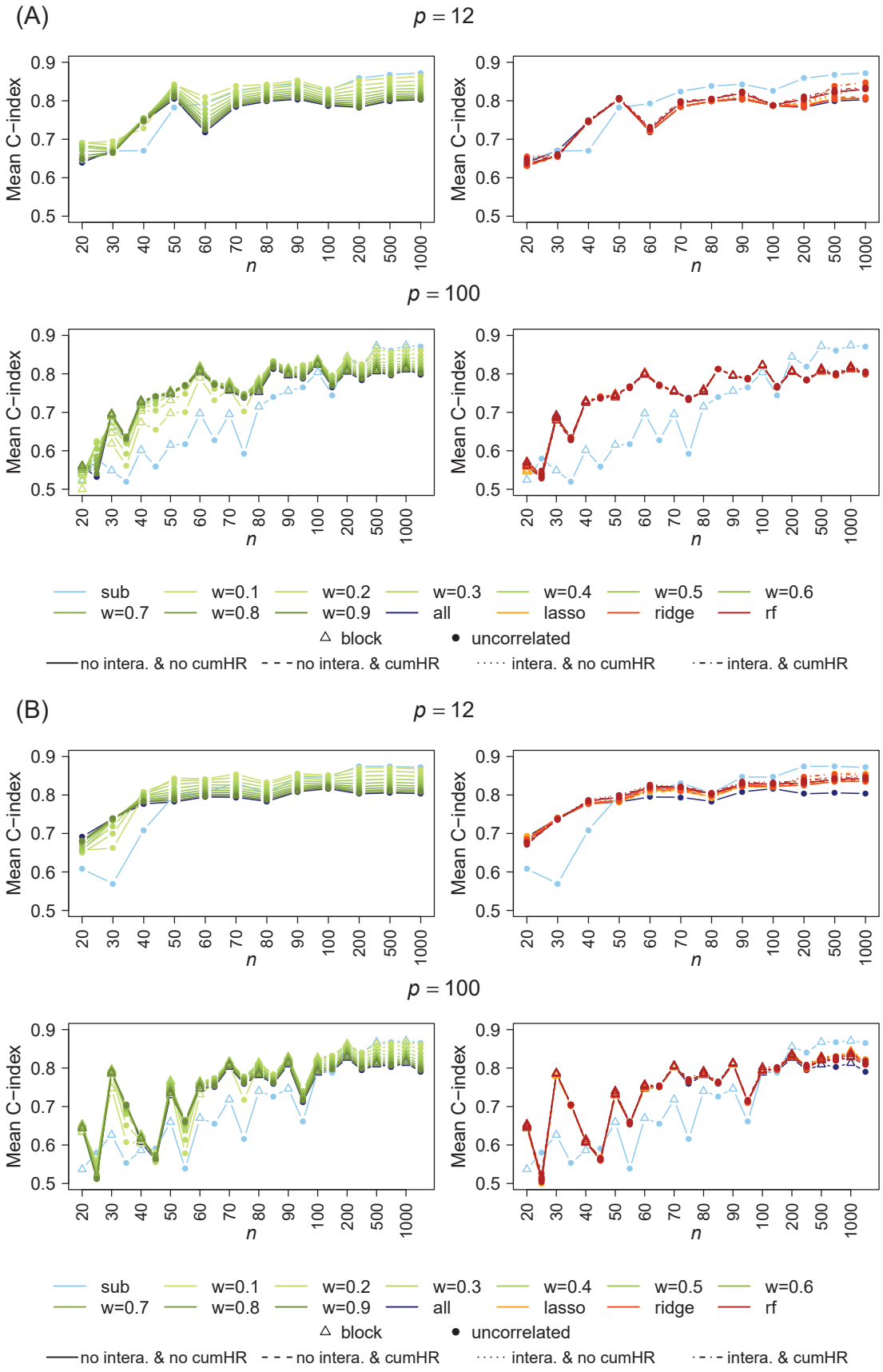


FIGURE B.10: *Mean C-index for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\mathbf{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.3$, (D) $\epsilon = 0.4$, (E) $\epsilon = 0.5$, (F) $\epsilon = 1$.*

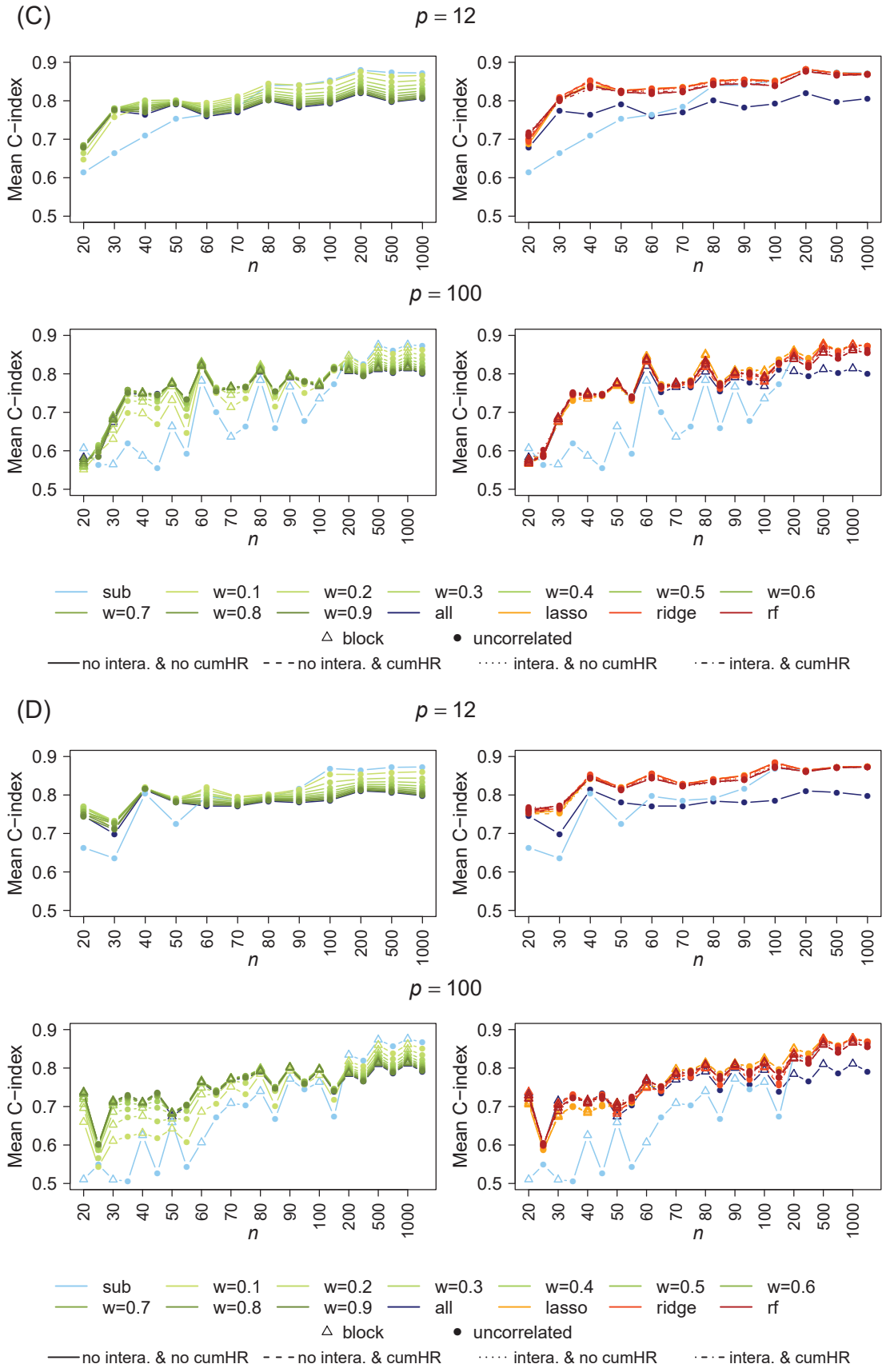


FIGURE B.10: *Mean C-index for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.3$, (D) $\epsilon = 0.4$, (E) $\epsilon = 0.5$, (F) $\epsilon = 1$ (cont.).*

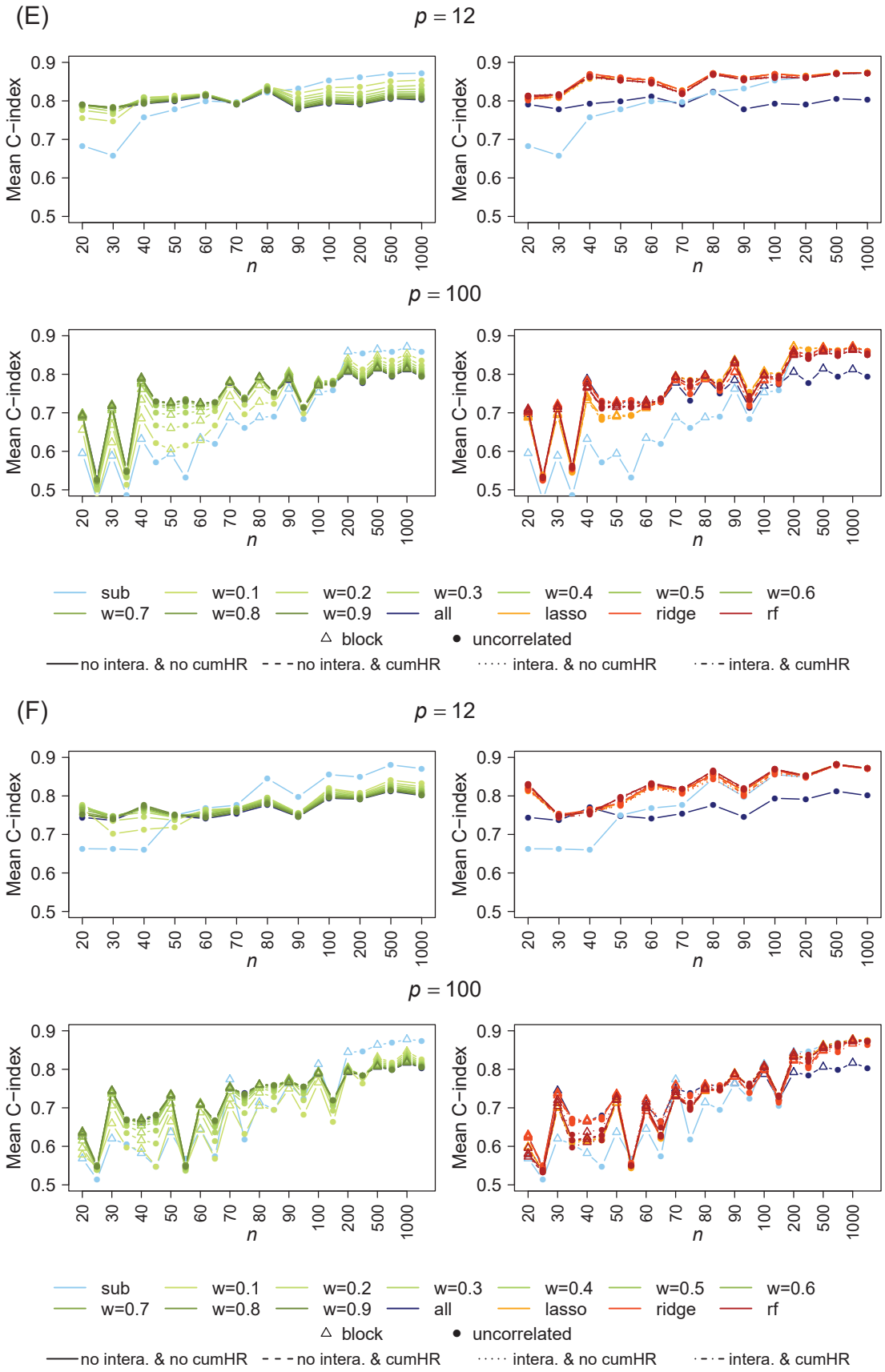


FIGURE B.10: *Mean C-index for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\mathbf{\Sigma}$). (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.3$, (D) $\epsilon = 0.4$, (E) $\epsilon = 0.5$, (F) $\epsilon = 1$ (cont.).*

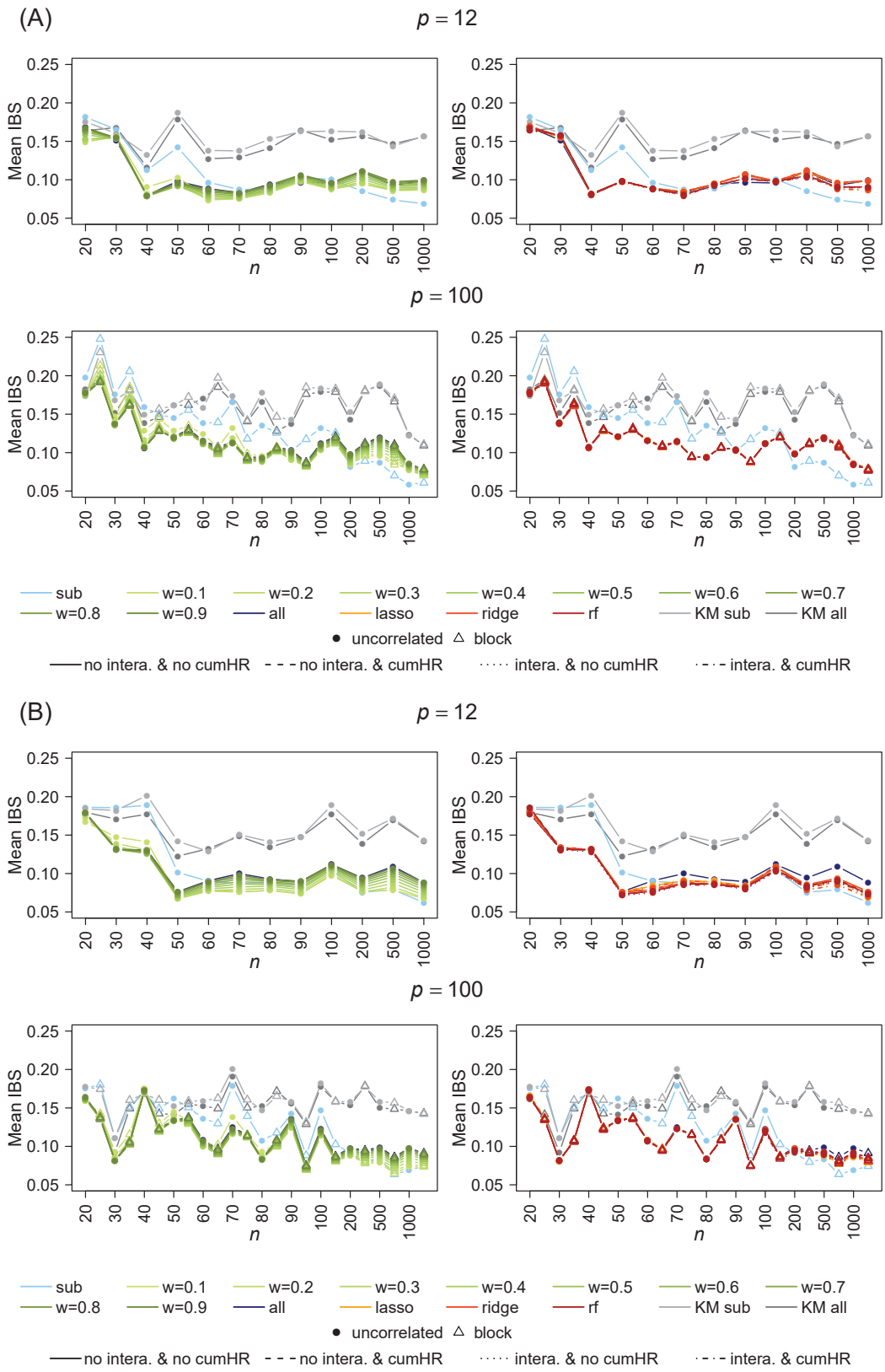


FIGURE B.11: *Mean integrated Brier score (IBS) for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$). Gray lines indicate the Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (sub) or combined (all) training data. (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$.*

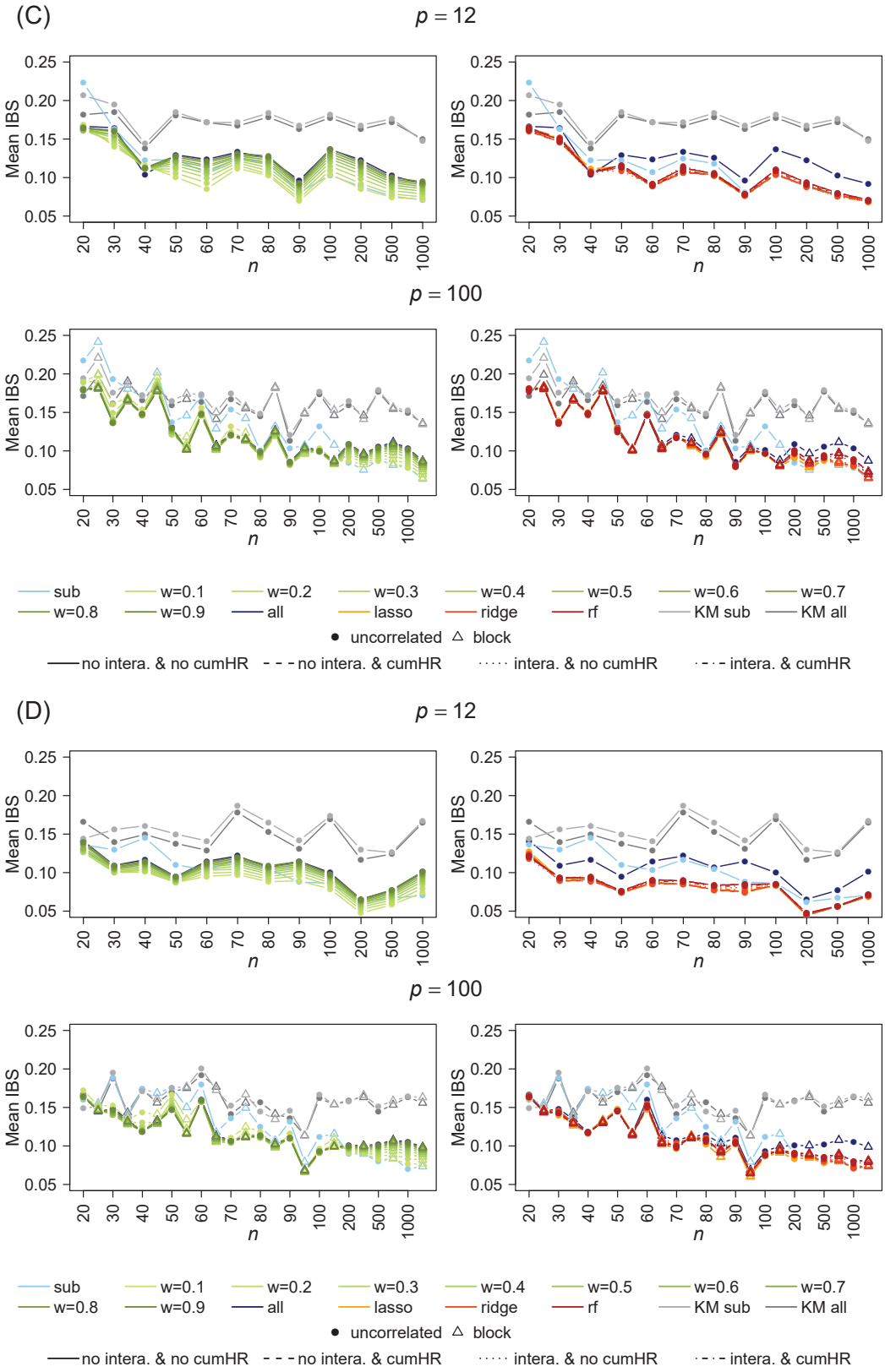


FIGURE B.11: *Mean integrated Brier score (IBS) for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$). Gray lines indicate the Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (sub) or combined (all) training data. (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$ (cont.).*

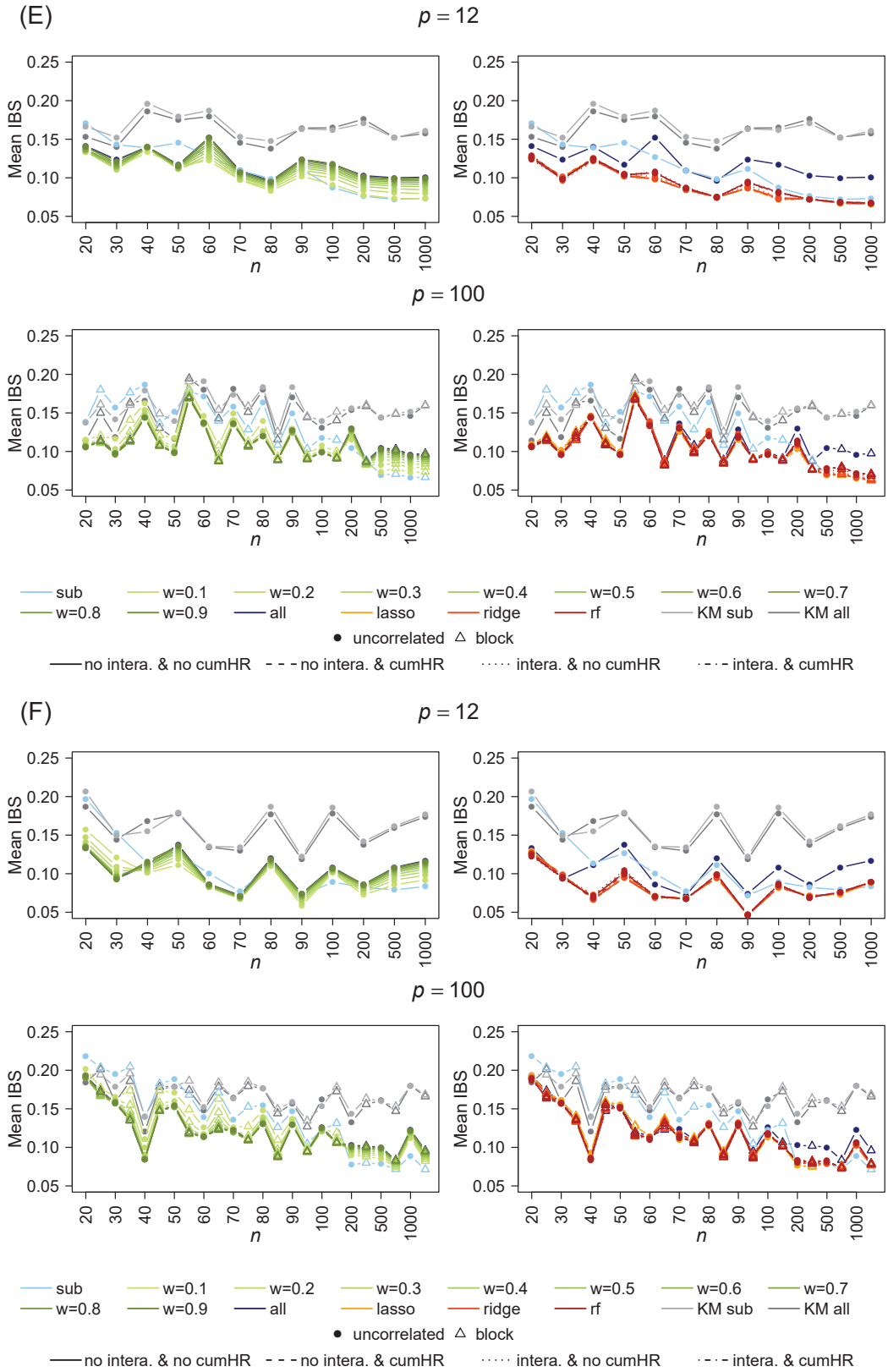


FIGURE B.11: *Mean integrated Brier score (IBS) for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$). Gray lines indicate the Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (sub) or combined (all) training data. (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$ (cont.).*

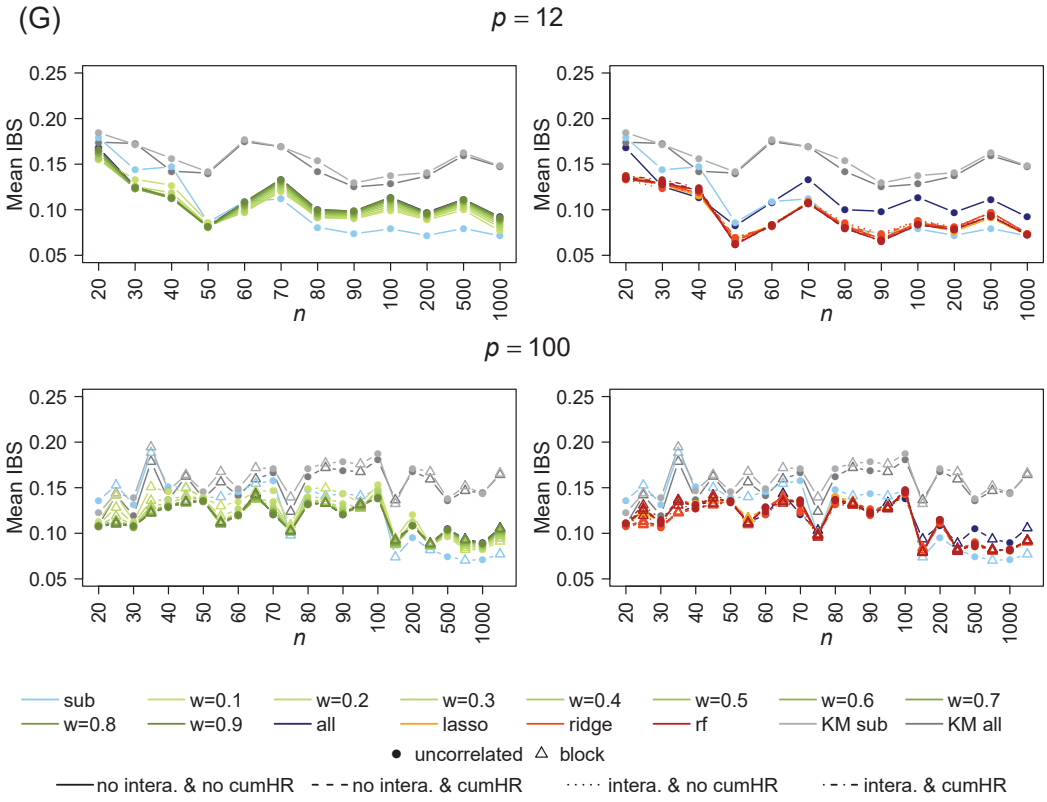


FIGURE B.11: *Mean integrated Brier score (IBS) for the Cox model, averaged across all test sets and subgroups. Comparison of different model types (color), parameter settings for weights estimation (line type), and varying parameters for data simulation (p, n, $\boldsymbol{\Sigma}$). Gray lines indicate the Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (sub) or combined (all) training data. (A) $\epsilon = 0$, (B) $\epsilon = 0.1$, (C) $\epsilon = 0.2$, (D) $\epsilon = 0.3$, (E) $\epsilon = 0.4$, (F) $\epsilon = 0.5$, (G) $\epsilon = 1$ (cont.).*

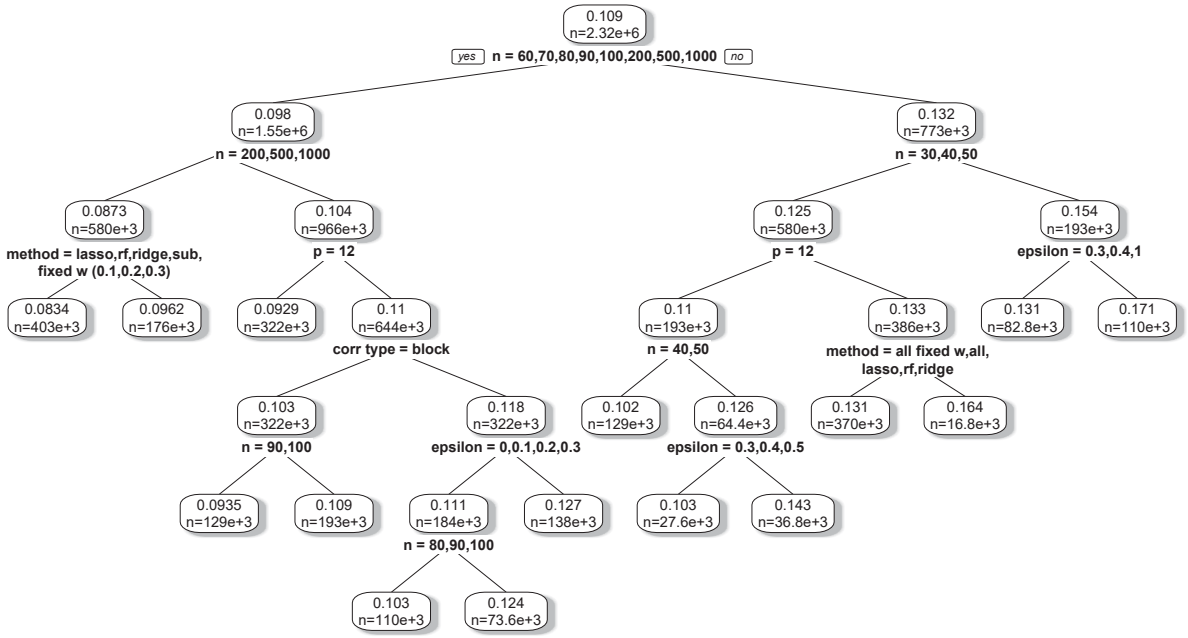


FIGURE B.12: *Regression tree for integrated Brier score (IBS) including all model types and parameter settings for data simulation and weights estimation as predictors. Each box shows mean IBS and sample size in the corresponding node.*

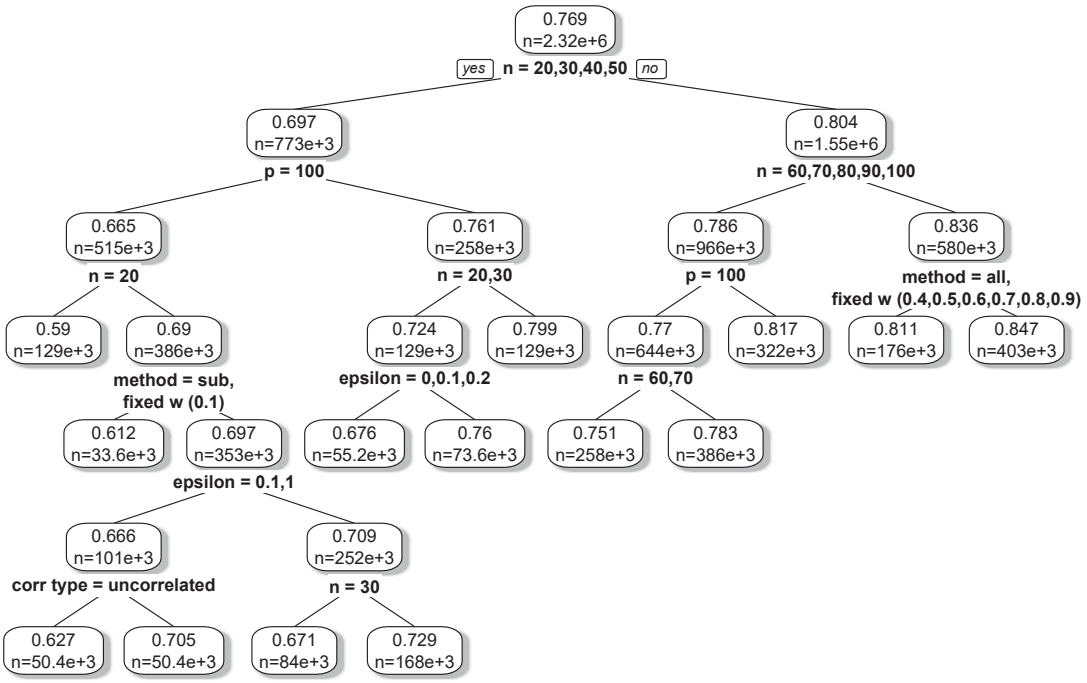


FIGURE B.13: *Regression tree for C-index including all model types and parameter settings for data simulation and weights estimation as predictors. Each box shows mean C-index and sample size in the corresponding node.*

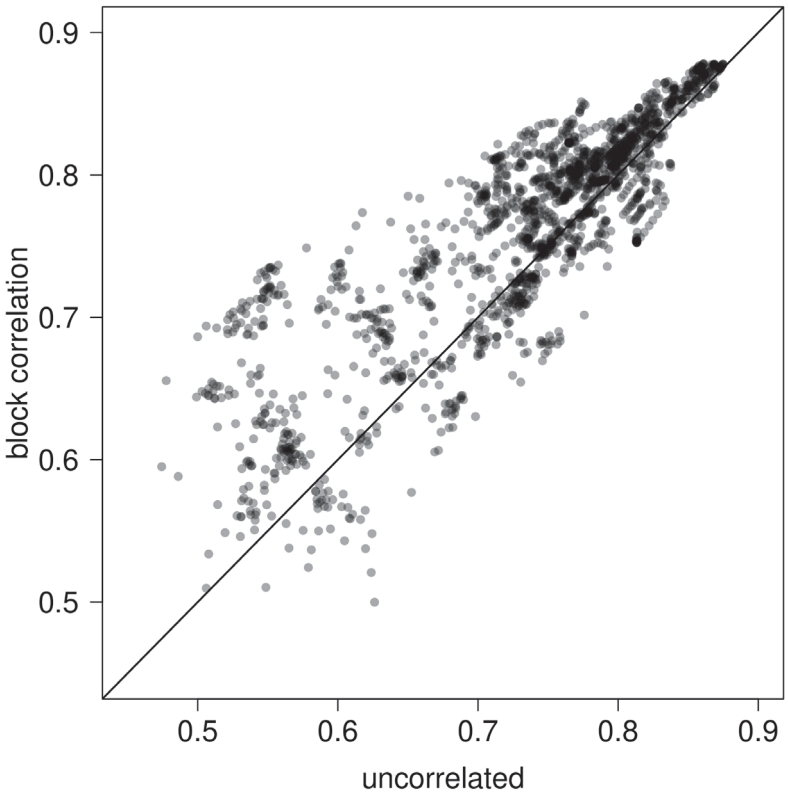


FIGURE B.14: *Scatterplot of mean C-index for two correlation structures across all combinations of model types and parameter settings for data simulation and weights estimation.*

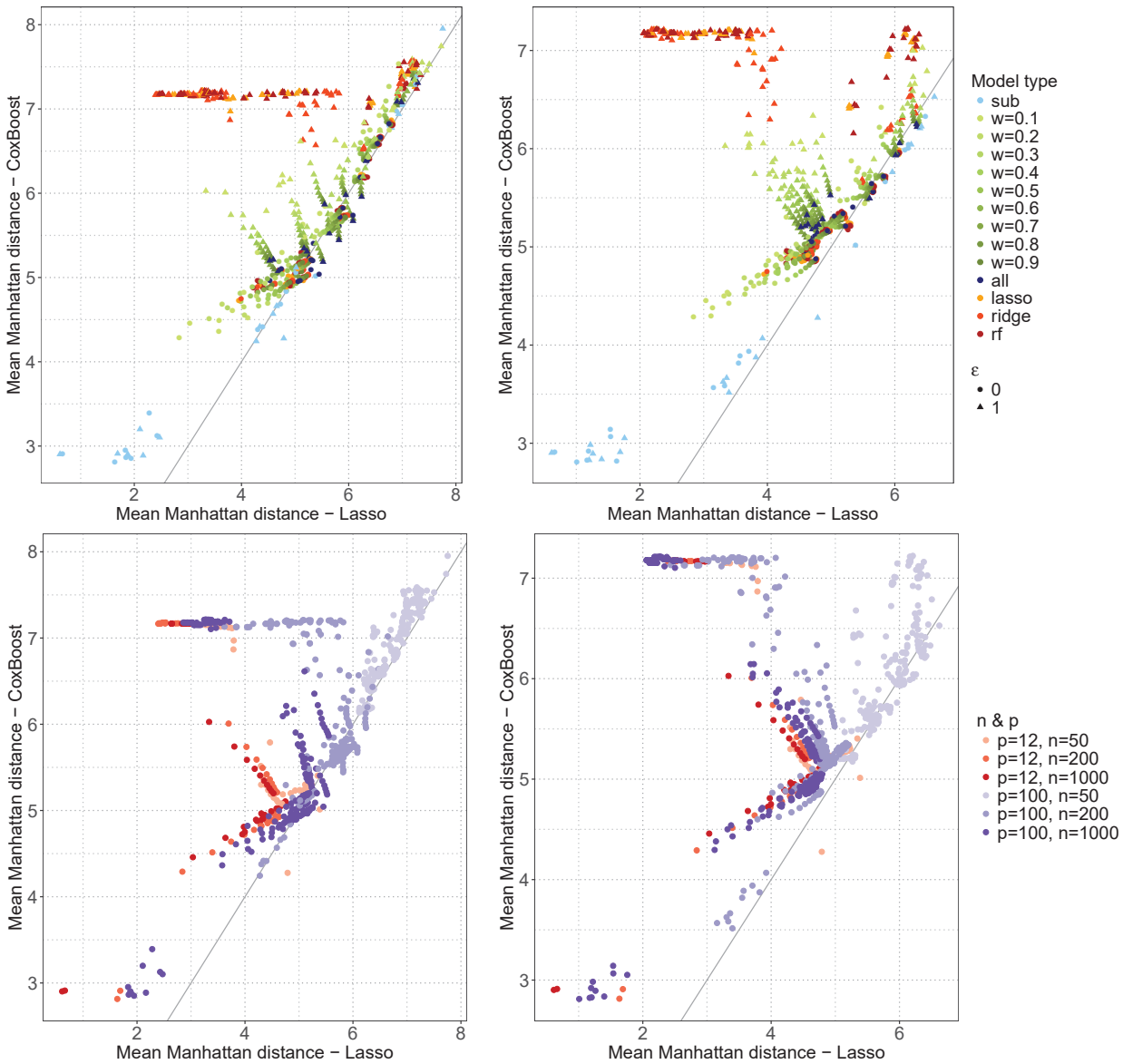


FIGURE B.15: *Mean Manhattan distance between true and estimated regression coefficients (averaged over all subgroups and training sets) for CoxBoost vs. Cox model with lasso penalty. Distances are computed using all covariates (left column), or just the first 12 prognostic covariates (right column).*

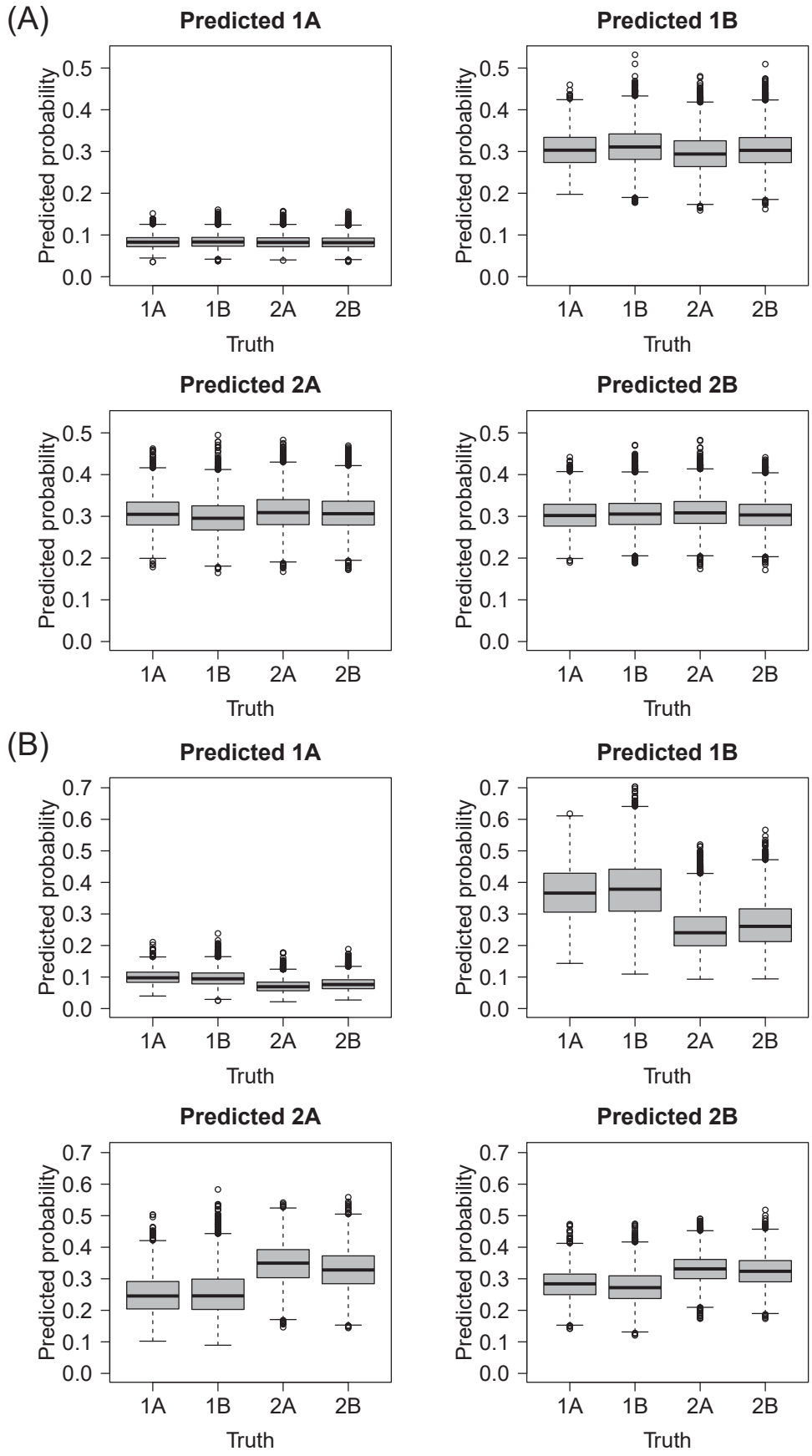


FIGURE B.16: *Predicted probabilities estimated with random forest without interactions, cumulative HR and oversampling, for simulated data with* $n^{(2)} = 200$, *and (A)* $\epsilon = 0$, *(B)* $\epsilon = 0.2$.

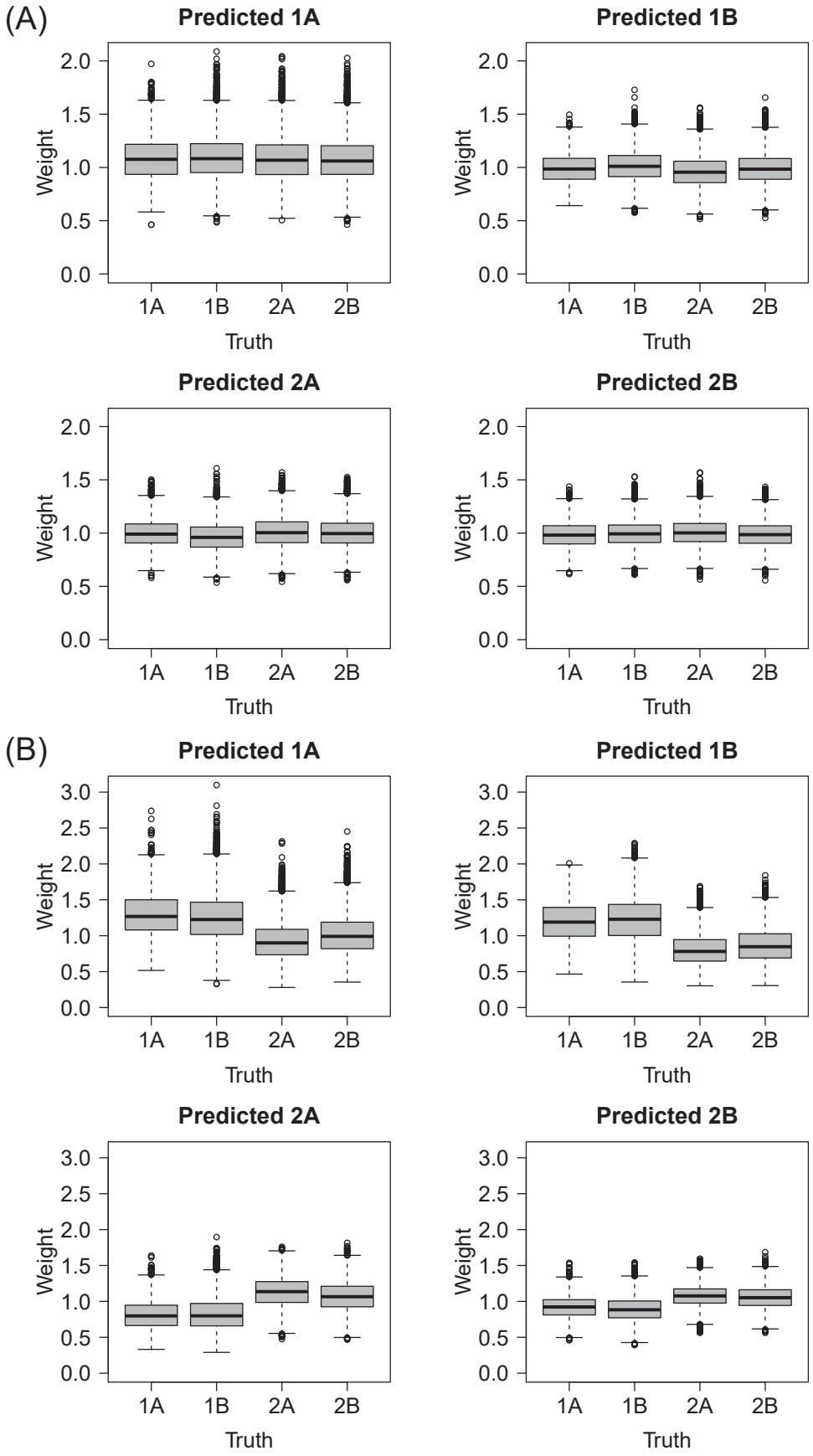


FIGURE B.17: *Weights estimated with random forest without interactions, cumulative HR and oversampling, for simulated data with* $n^{(2)} = 200$, *and (A)* $\epsilon = 0$, *(B)* $\epsilon = 0.2$.

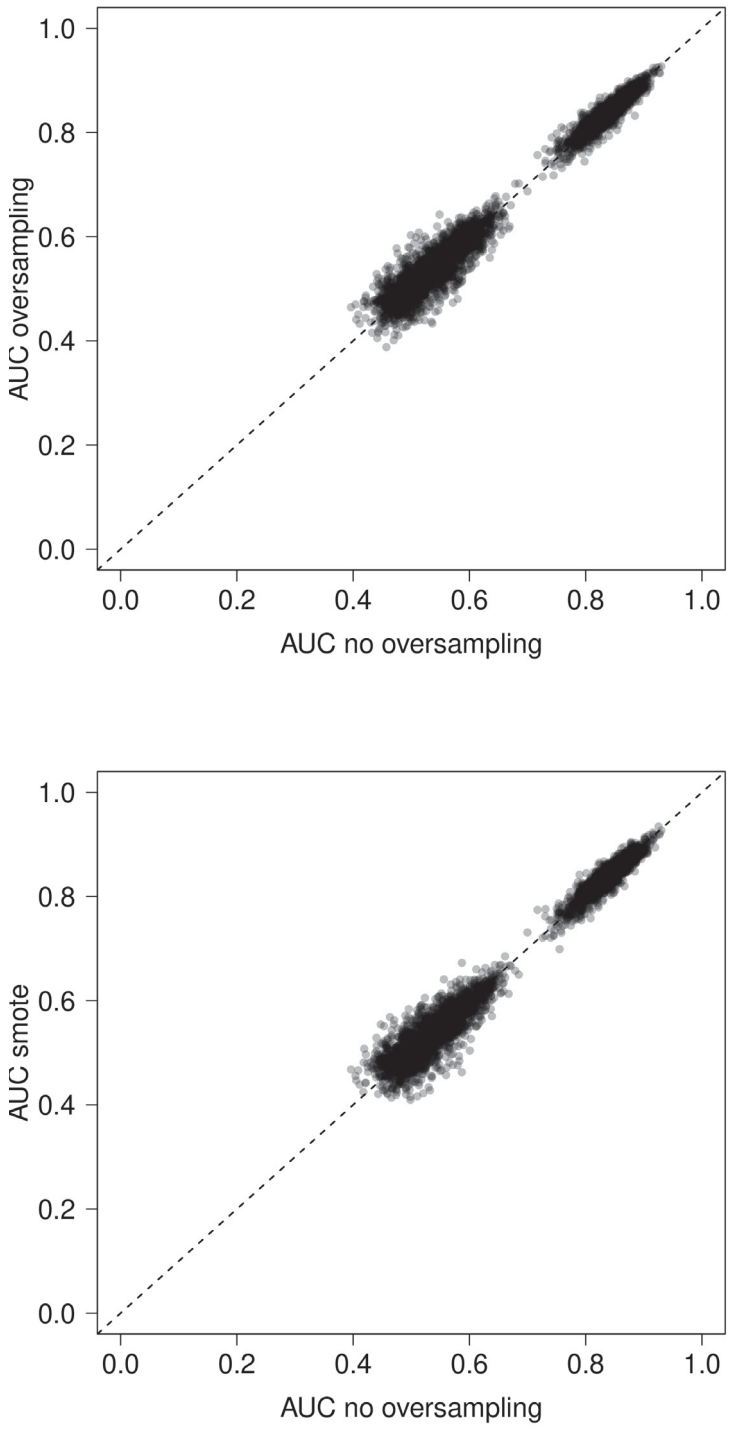


FIGURE B.18: *Scatterplot of AUC values based on cross-validated training data for oversampling or smote compared to no oversampling.*

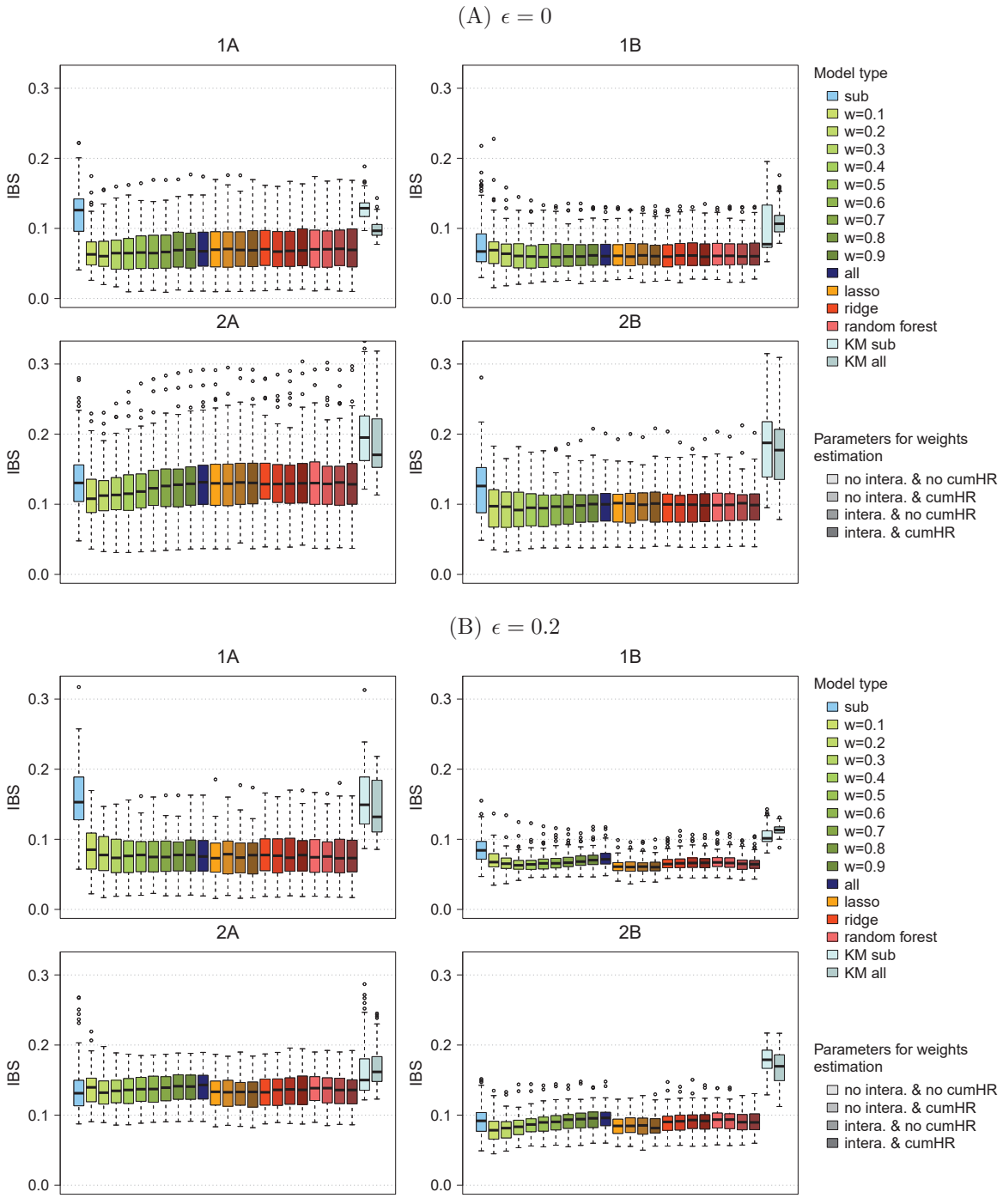


FIGURE B.19: *Boxplots of the integrated Brier score (IBS) based on all test sets for the prediction of each subgroup, with* $n^{(2)} = 100$ *and weights estimation without oversampling. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (KM sub) or combined (KM all) test data.*

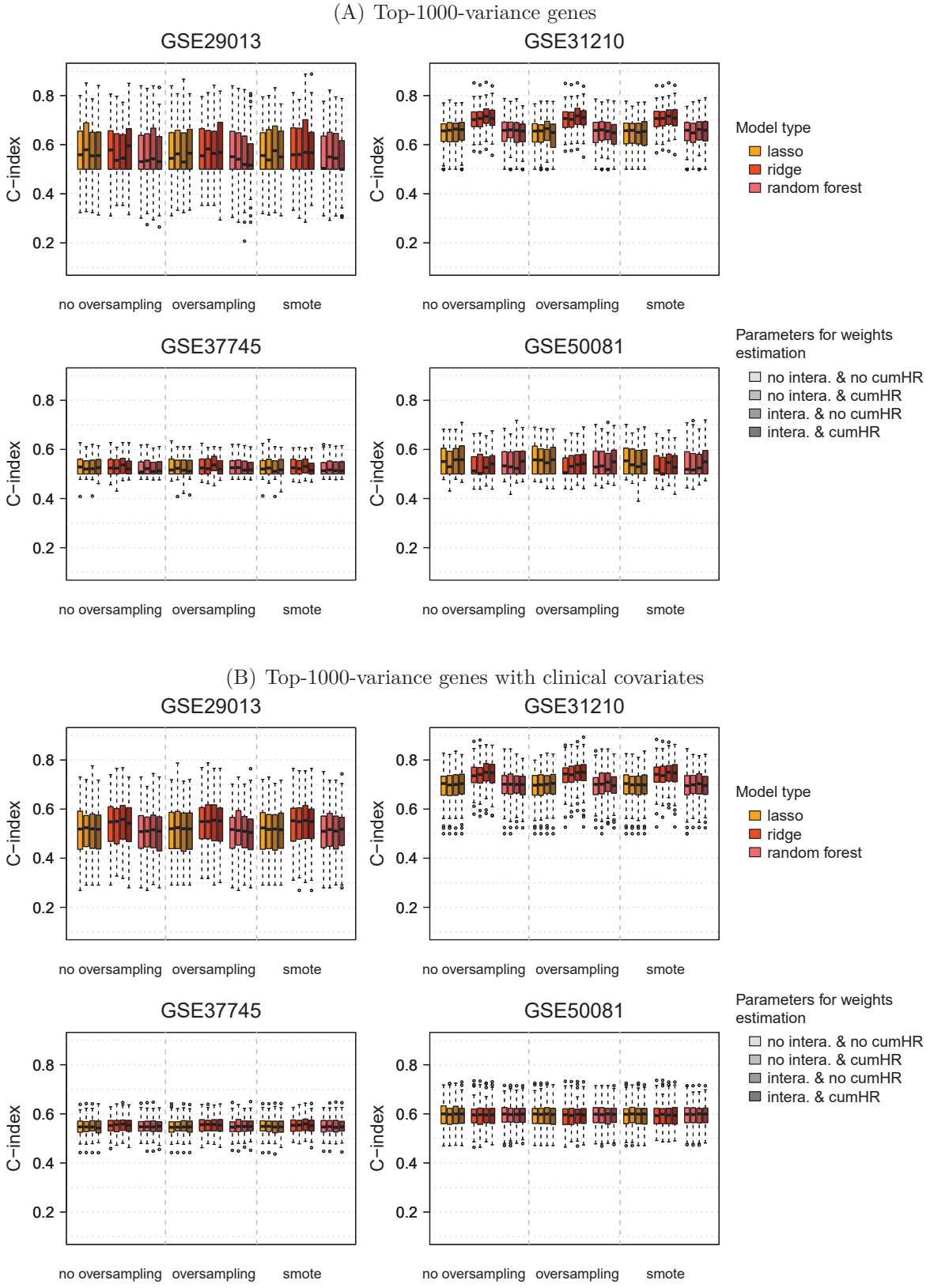


FIGURE B.20: *Boxplots of C-index based on all test sets for the prediction of each subgroup under varying parameters for weights estimation.*

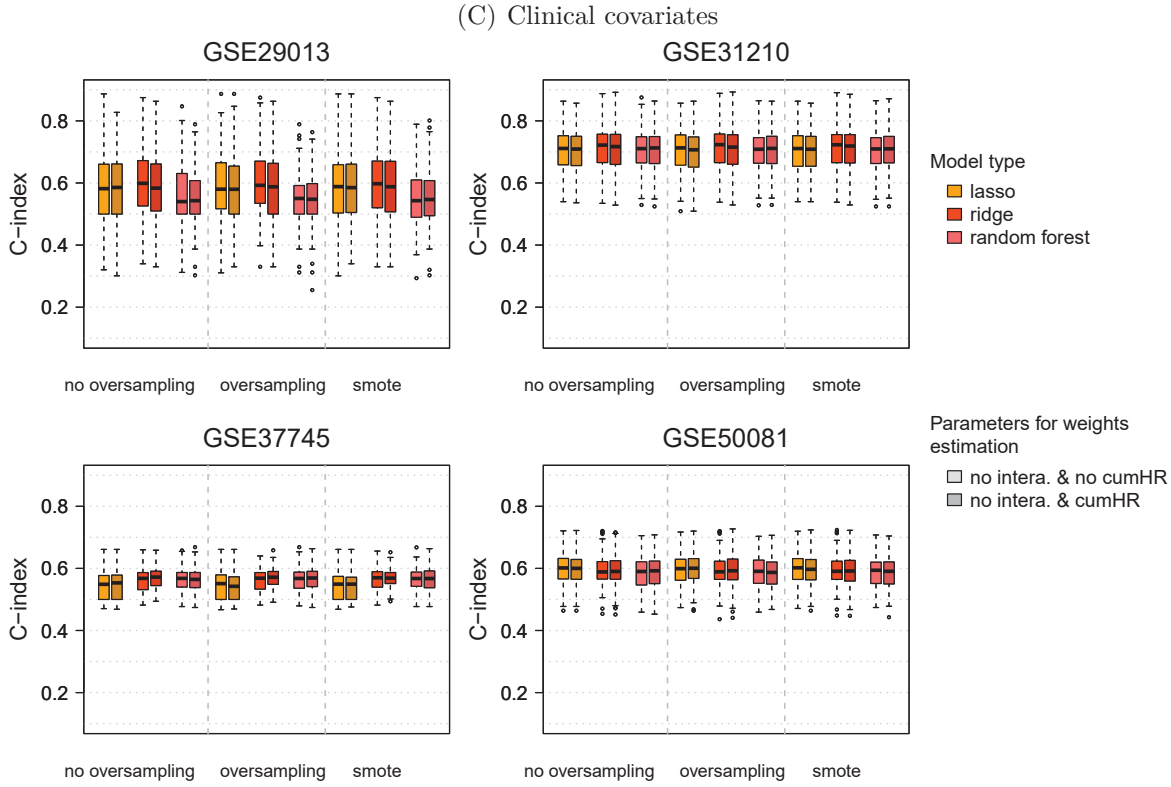


FIGURE B.20: *Boxplots of C-index based on all test sets for the prediction of each subgroup under varying parameters for weights estimation (cont.).*

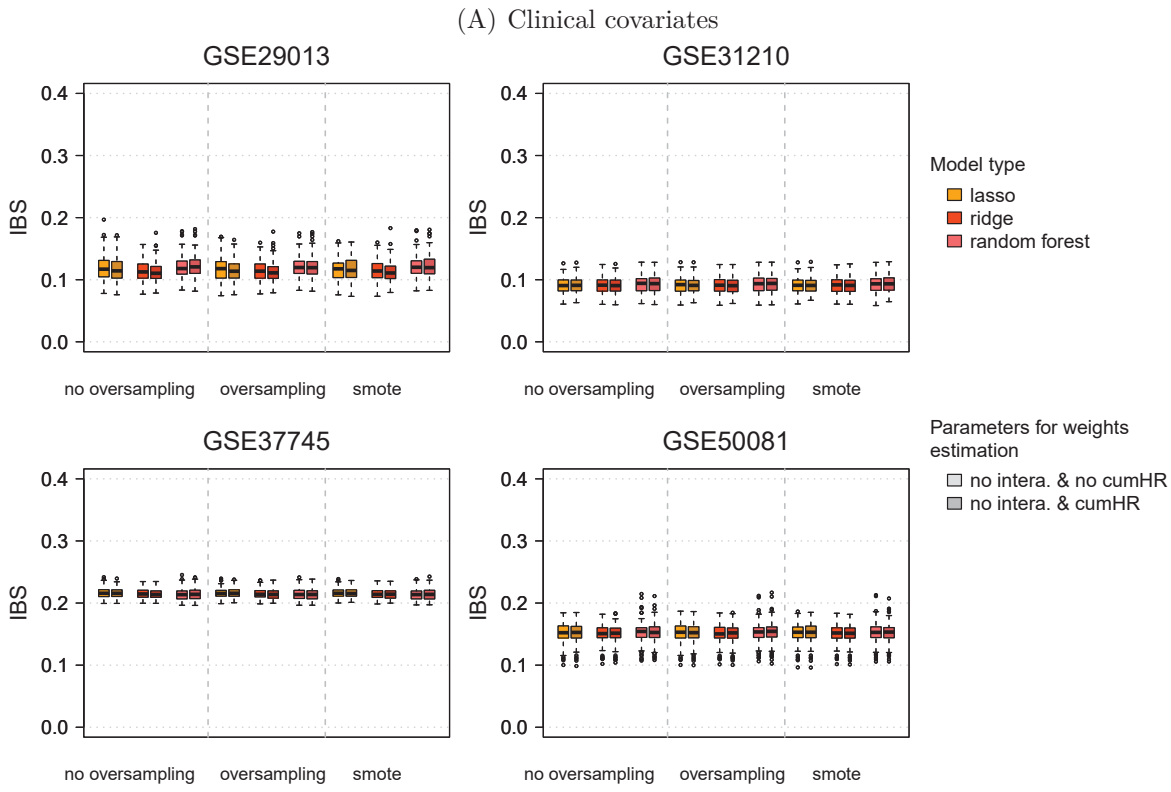


FIGURE B.21: *Boxplots of IBS based on all test sets for the prediction of each subgroup under varying parameters for weights estimation.*

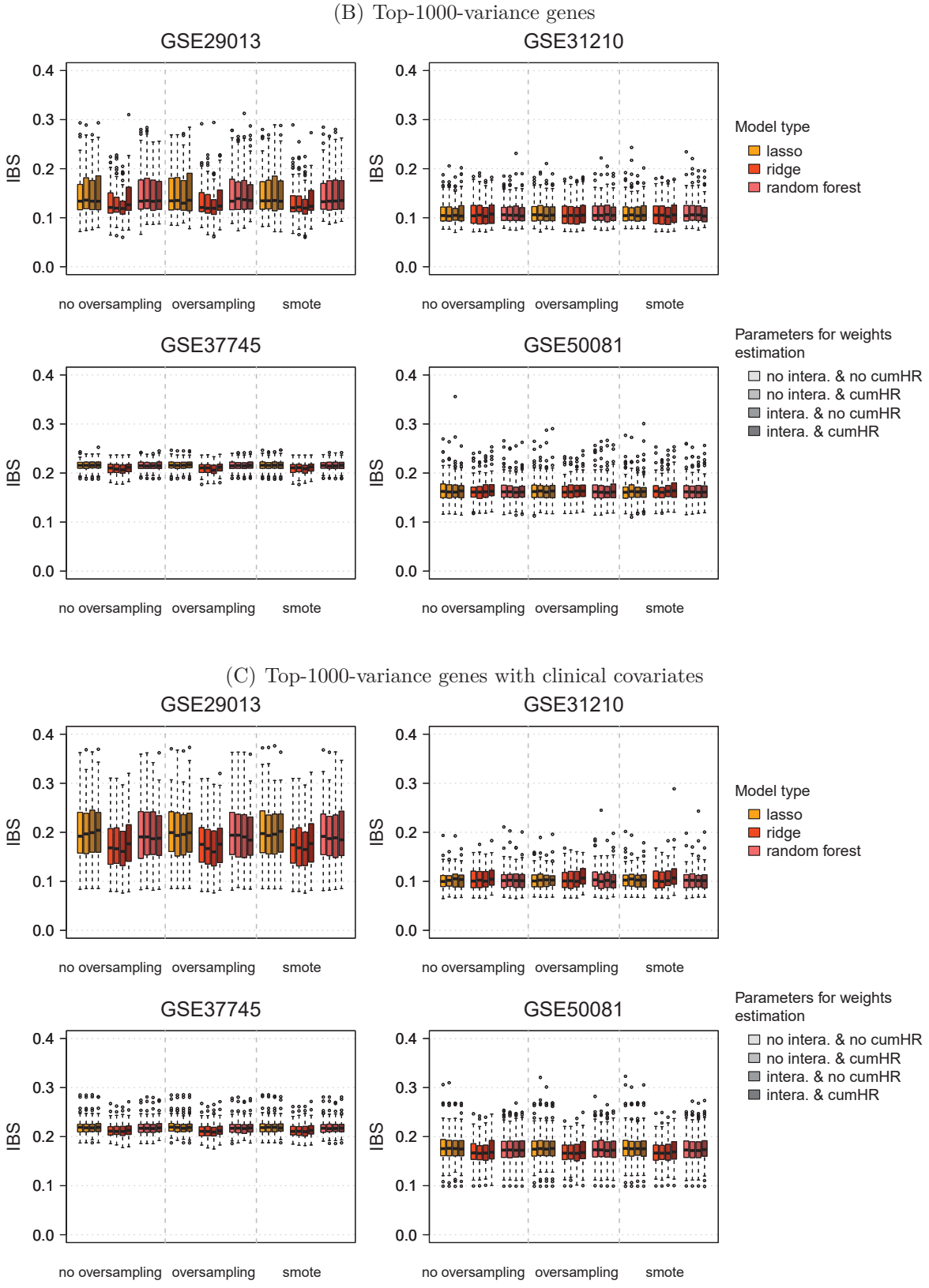


FIGURE B.21: *Boxplots of IBS based on all test sets for the prediction of each subgroup under varying parameters for weights estimation (cont.).*

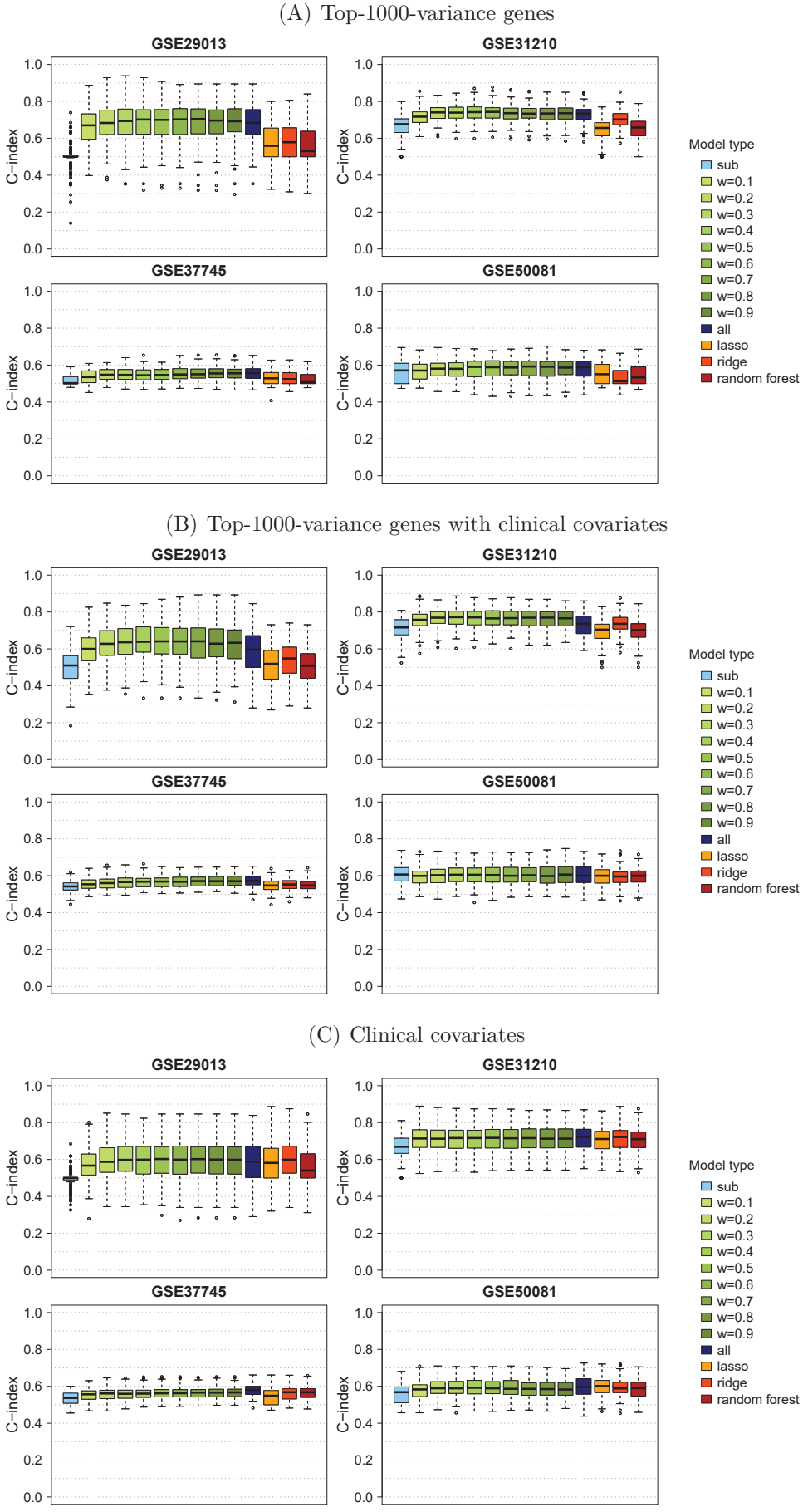


FIGURE B.22: *Boxplots of C-index based on all test sets for the prediction of each subgroup under different Cox model types and covariate sets.*

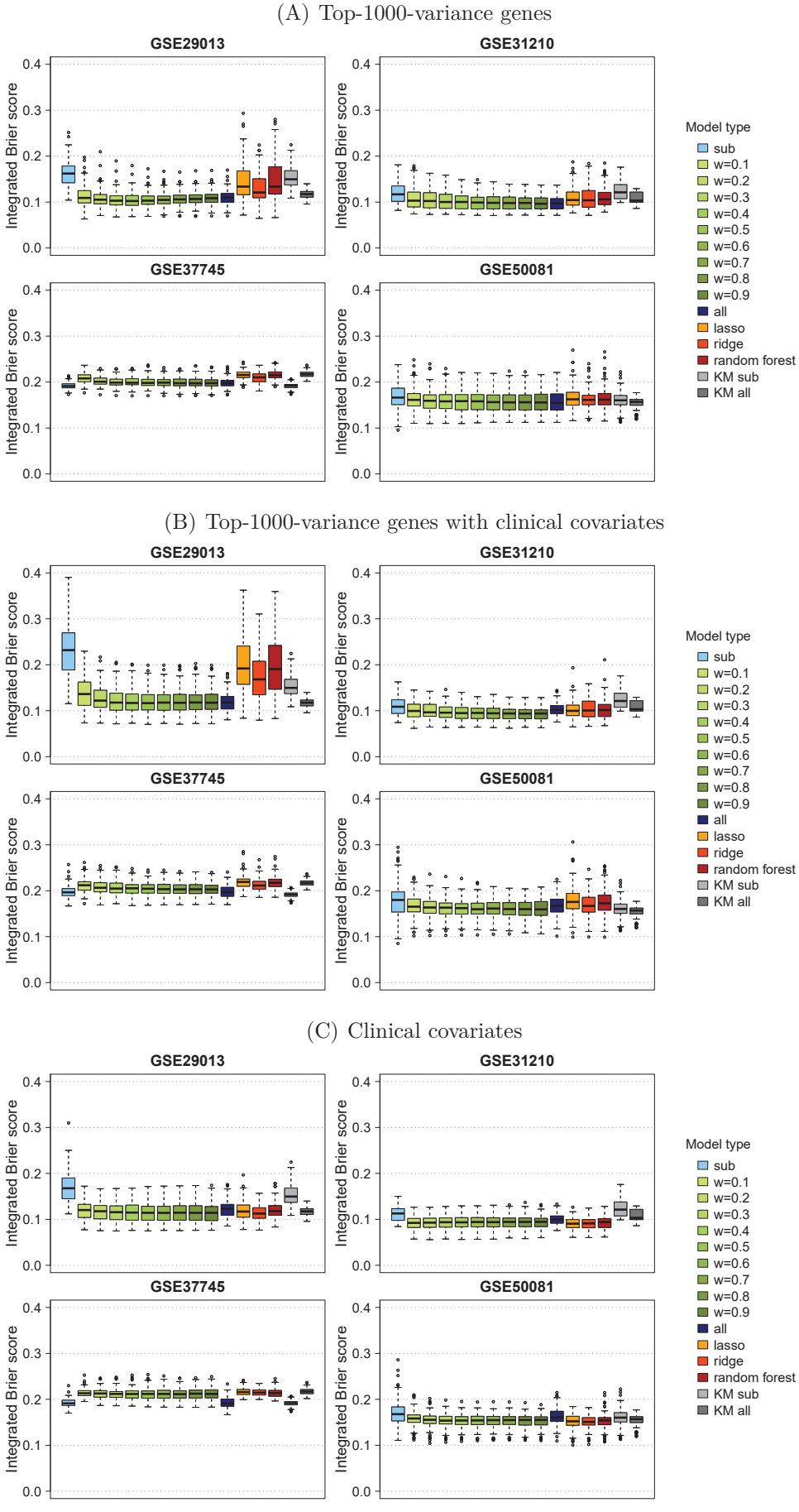
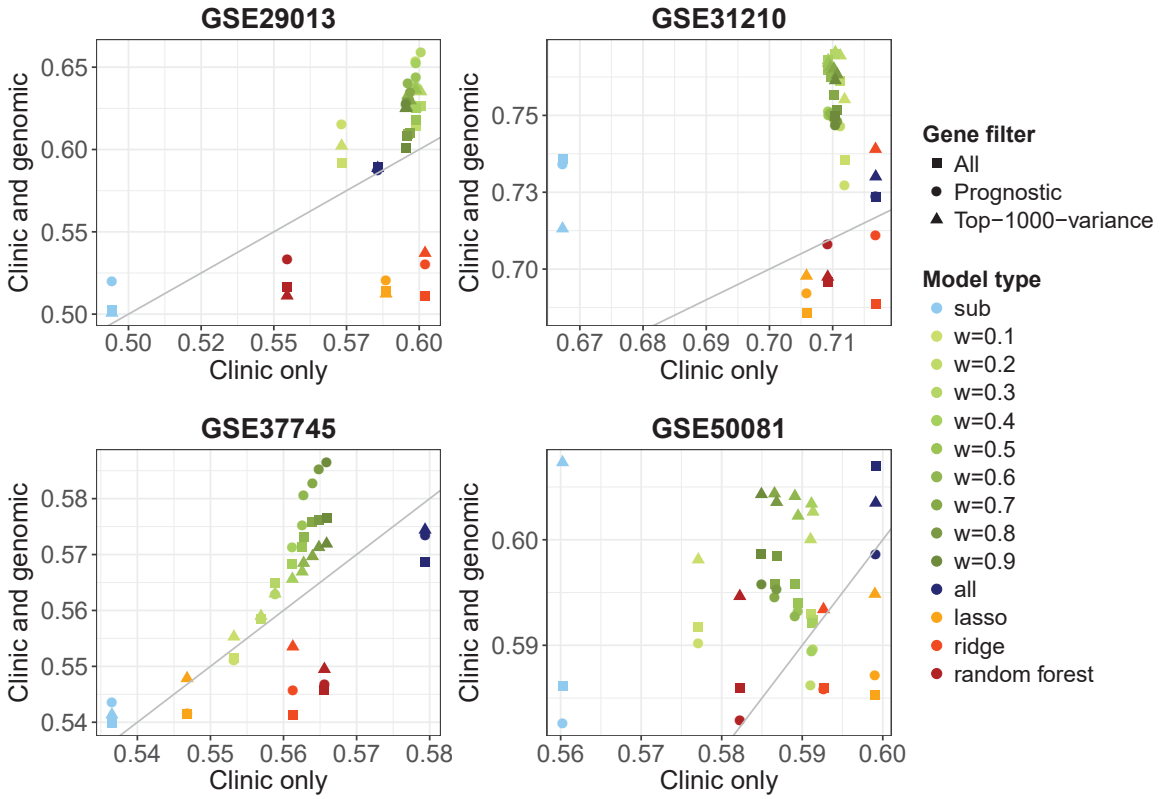


FIGURE B.23: *Boxplots of IBS for different Cox models and covariate sets. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (KM sub) or combined (KM all) test data.*

(A) Mean C-index



(B) Mean integrated Brier score

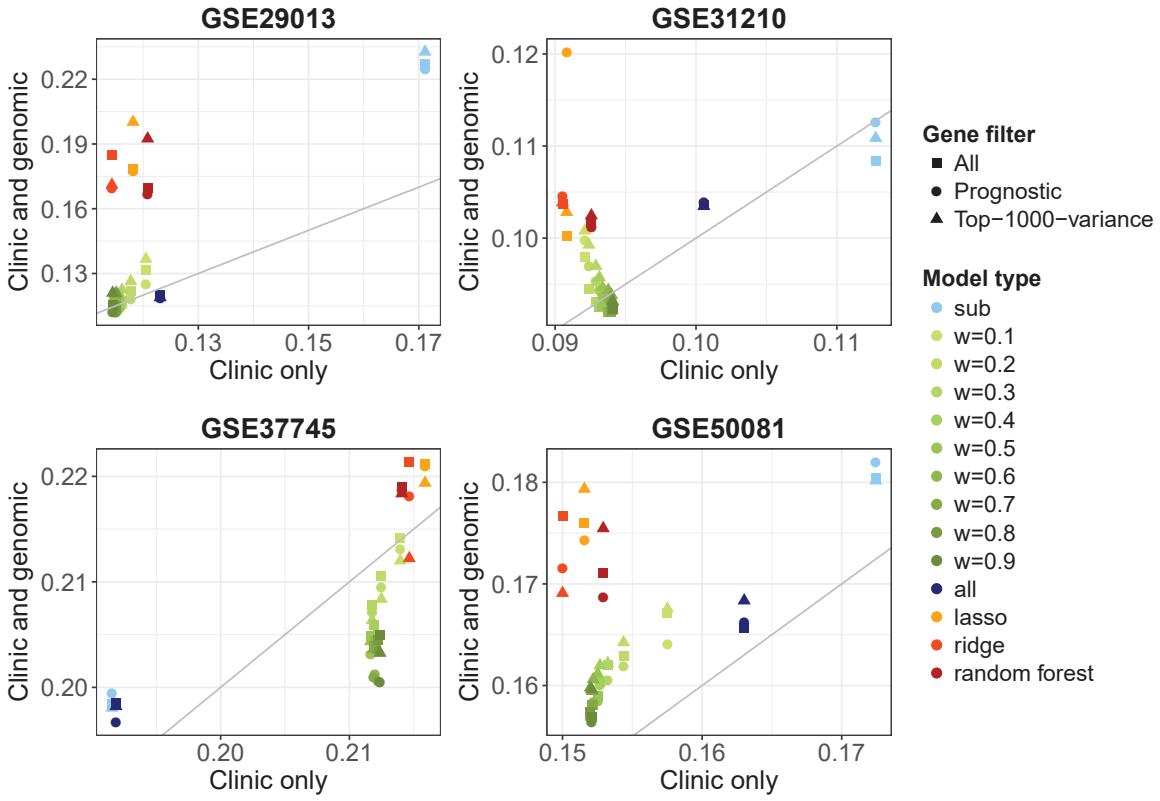


FIGURE B.24: *Comparison of clinical covariates and combination of clinical and genomic covariates with respect to mean prediction performance (averaged across all test sets).*

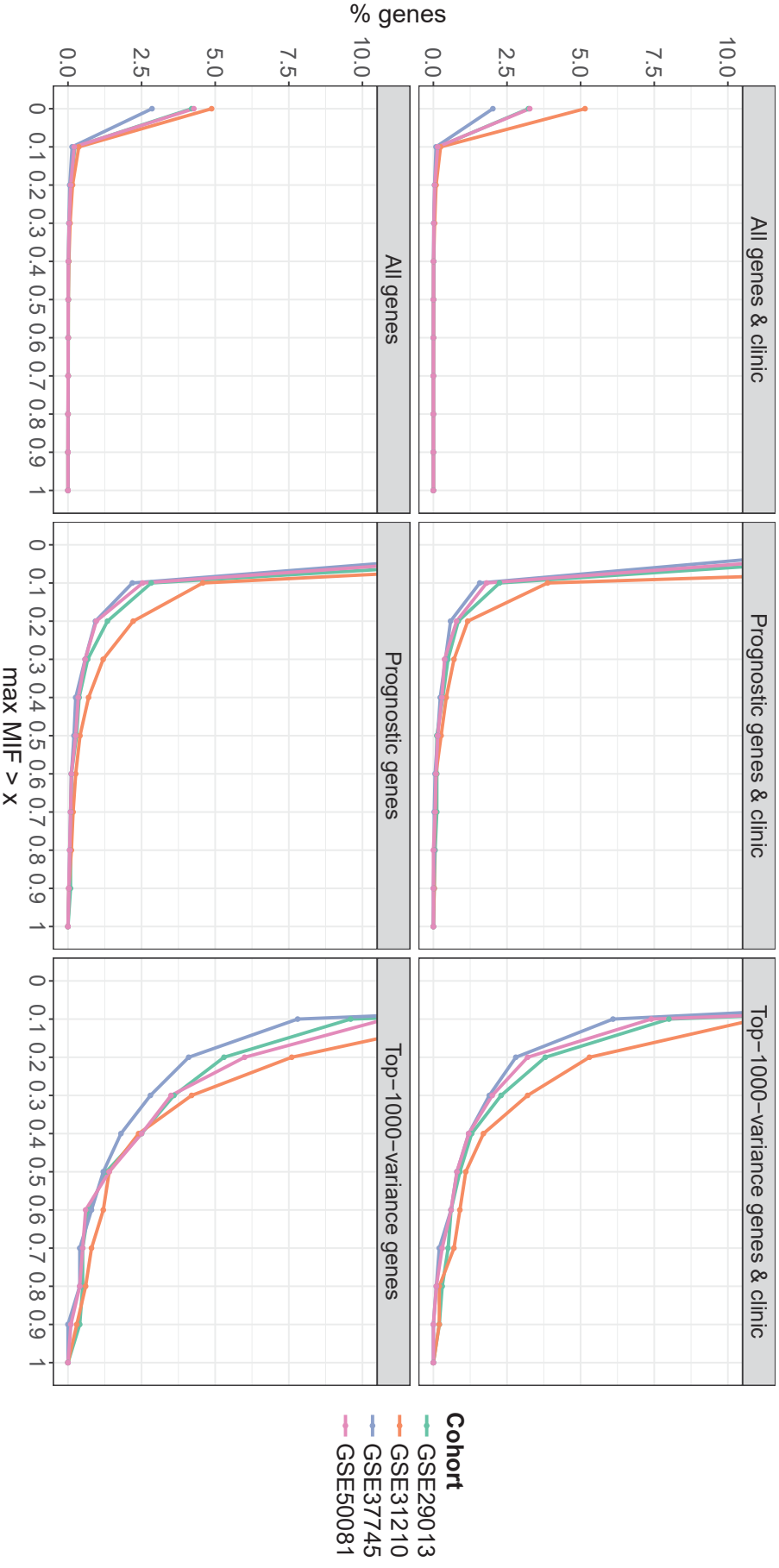


FIGURE B.25: *Percentage of genes with maximum MIF in any Cox model larger than* $x = 0, 0.1, ..., 1$.

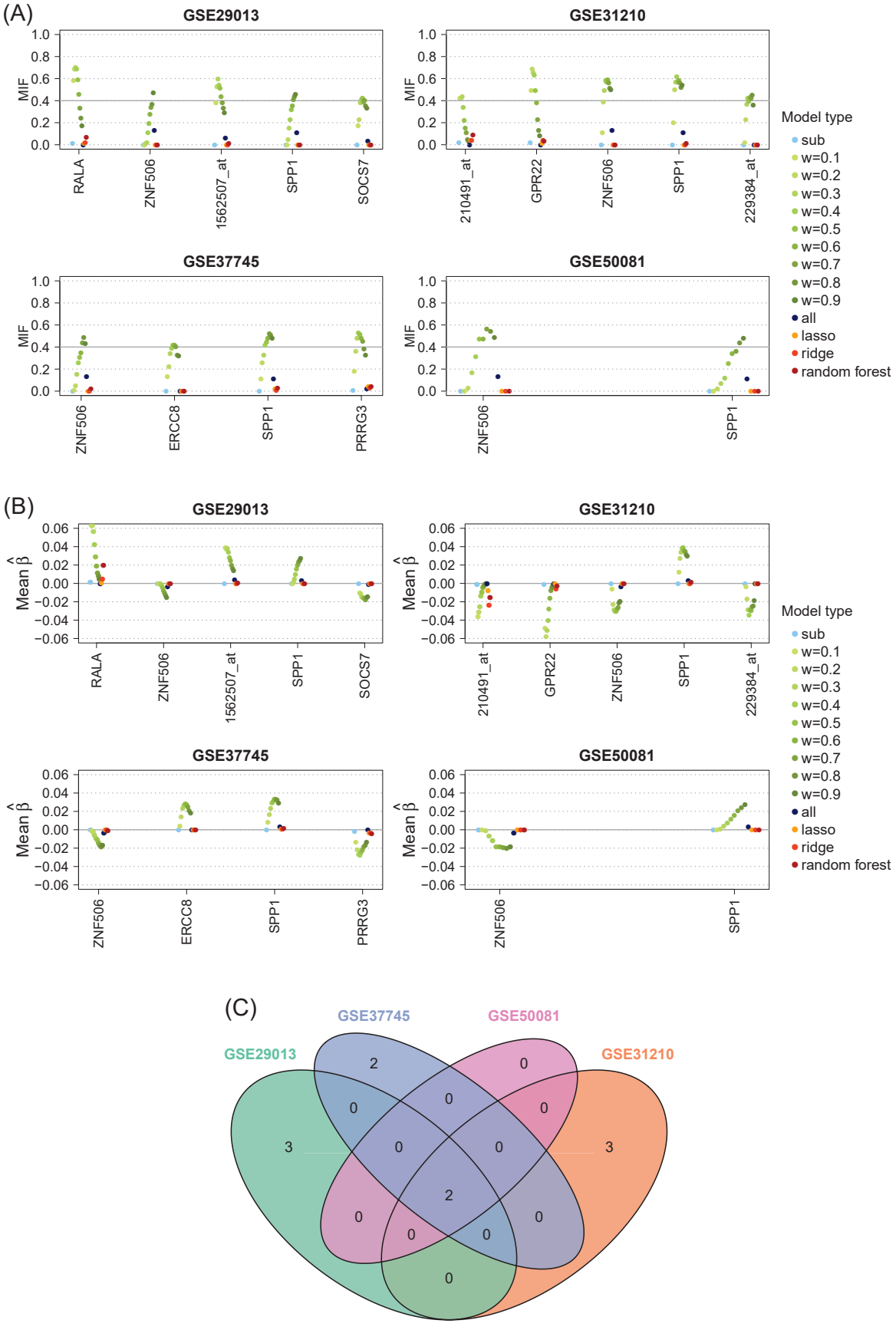


FIGURE B.26: *Results of variable selection for Cox models including all genes and mandatory clinical covariates. For each subgroup genes with a mean inclusion frequency (MIF) larger than 0.4 in any model type are selected. (A) Mean inclusion frequencies, (B) mean estimated regression coefficients, and (C) Venn diagram of selected genes in all subgroups.*

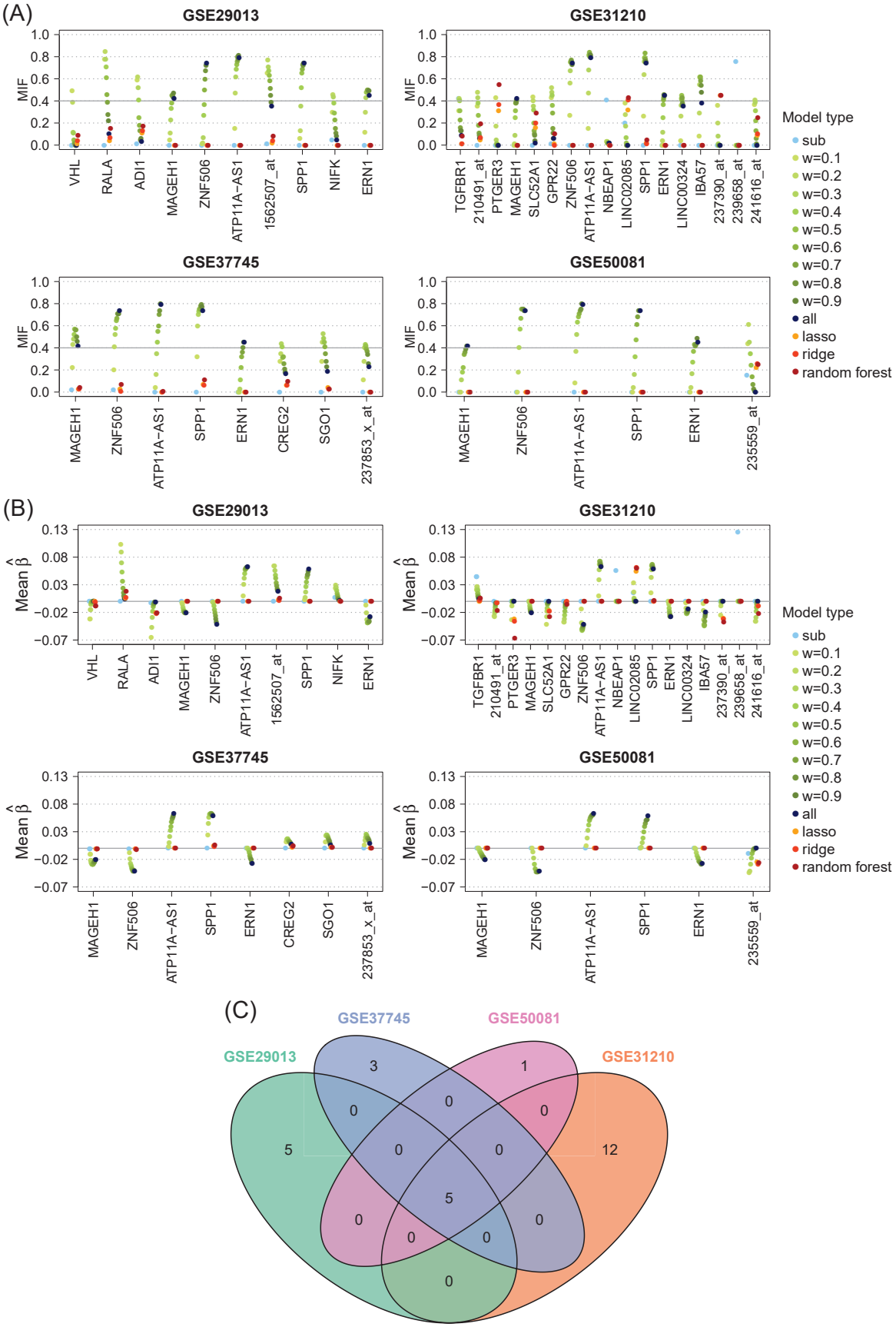


FIGURE B.27: *Results of variable selection for Cox models including all genes as covariates. For each subgroup genes with a mean inclusion frequency (MIF) larger than 0.4 in any model type are selected. (A) Mean inclusion frequencies, (B) mean estimated regression coefficients, and (C) Venn diagram of selected genes in all subgroups.*

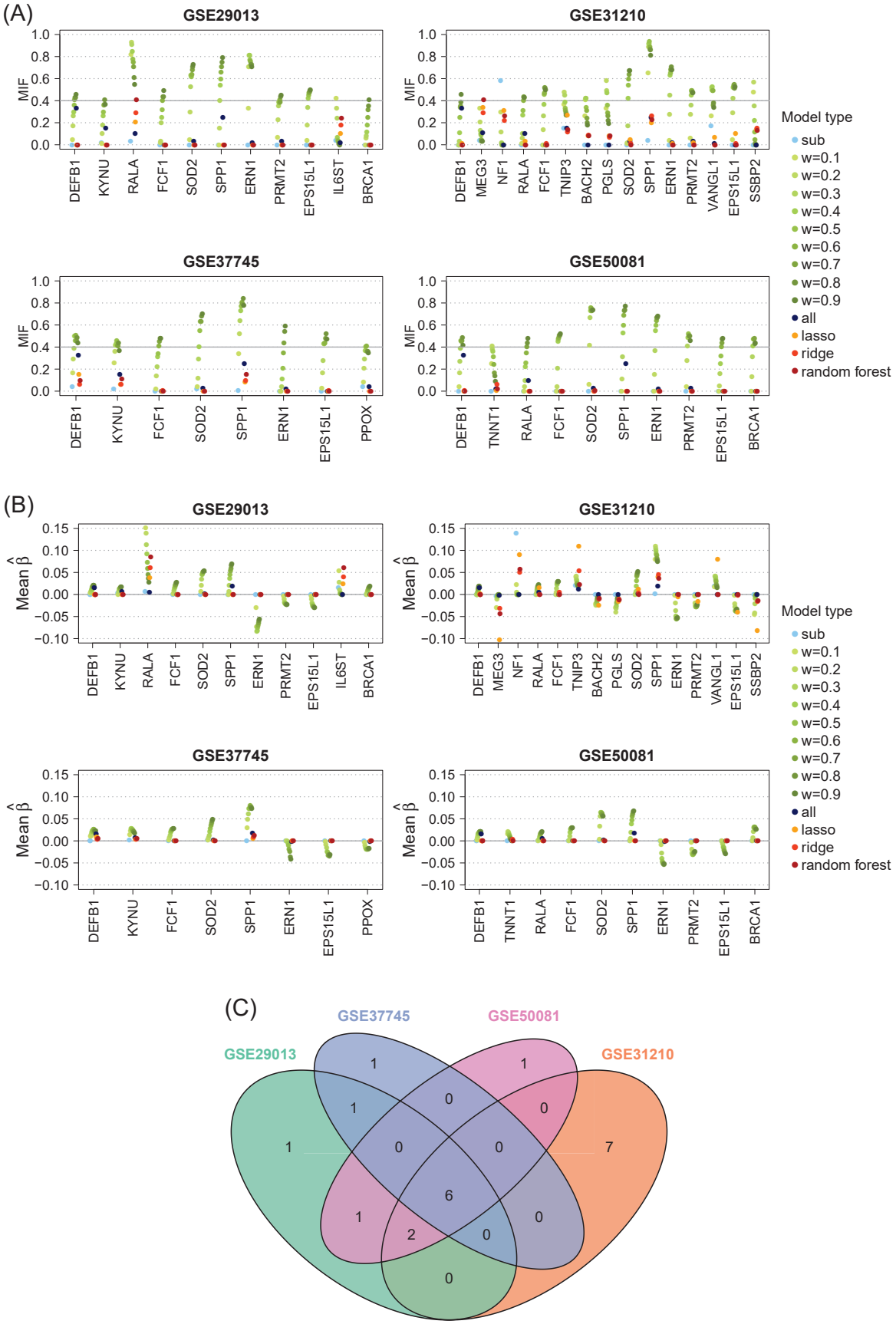


FIGURE B.28: *Results of variable selection for Cox models including prognostic genes and mandatory clinical covariates. For each subgroup genes with a mean inclusion frequency (MIF) larger than 0.4 in any model type are selected. (A) Mean inclusion frequencies, (B) mean estimated regression coefficients, and (C) Venn diagram of selected genes in all subgroups.*

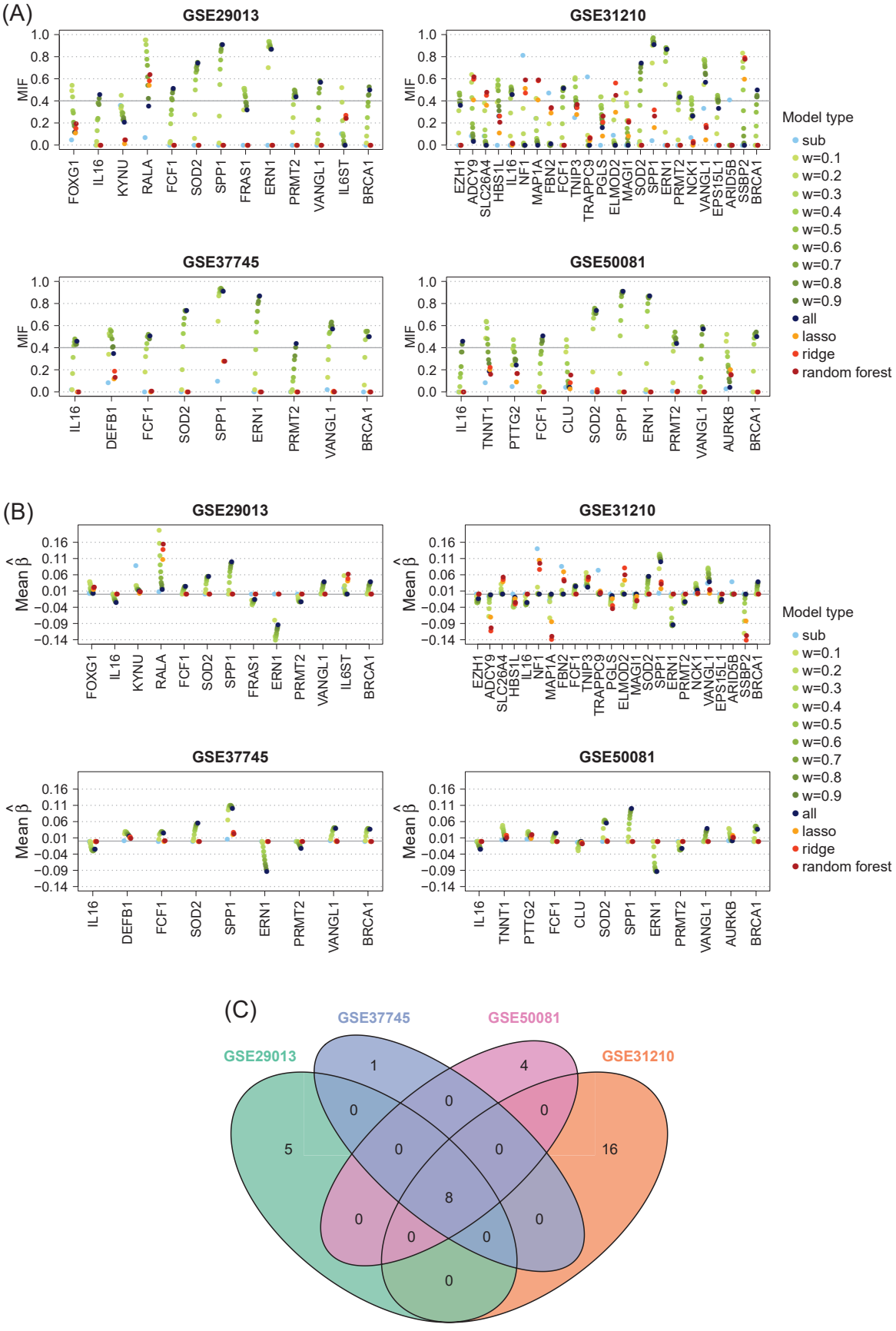


FIGURE B.29: *Results of variable selection for Cox models including prognostic genes as covariates. For each subgroup genes with a mean inclusion frequency (MIF) larger than 0.4 in any model type are selected. (A) Mean inclusion frequencies, (B) mean estimated regression coefficients, and (C) Venn diagram of selected genes in all subgroups.*

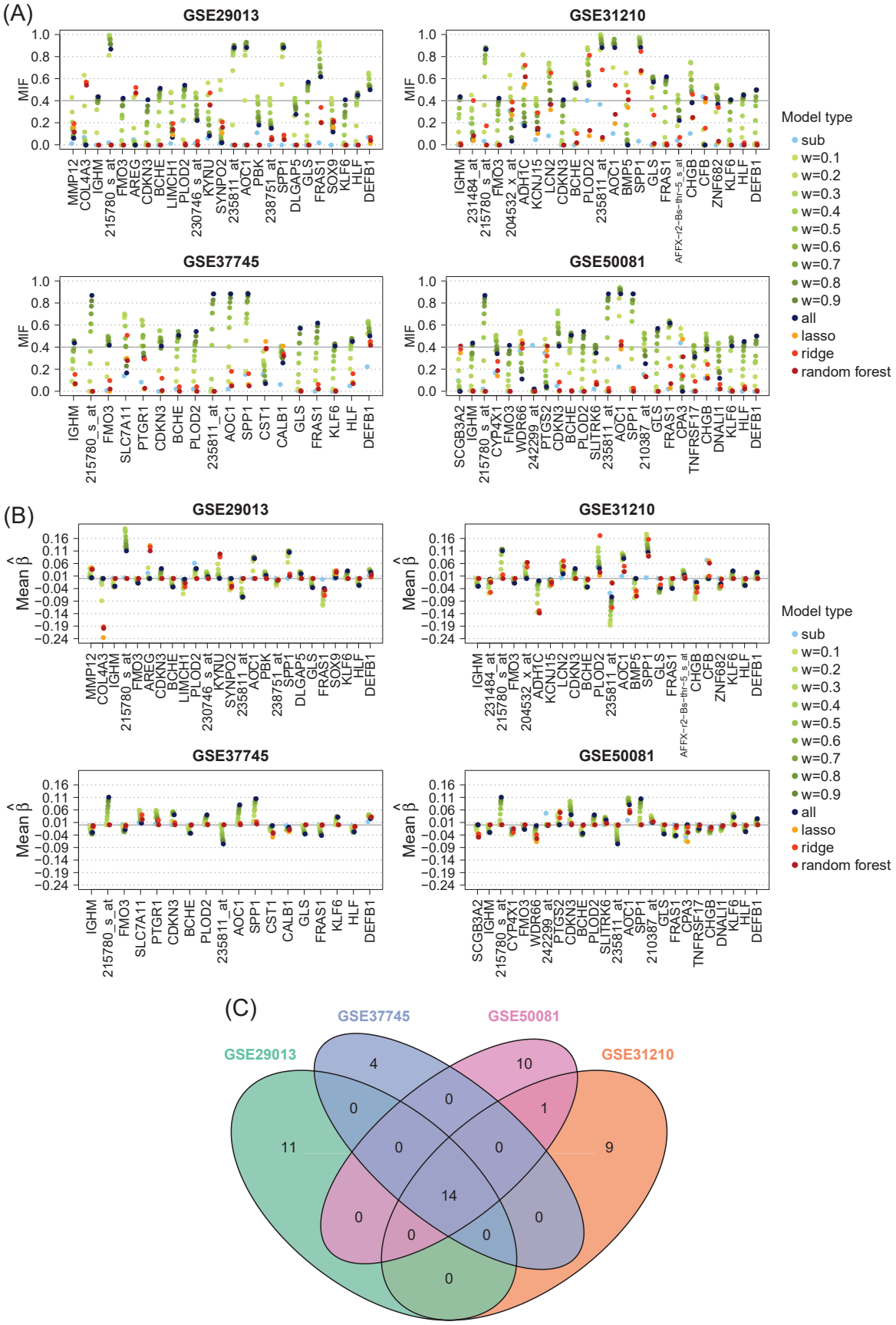


FIGURE B.30: *Results of variable selection for Cox models including top-1000-variance genes as covariates. For each subgroup genes with a mean inclusion frequency (MIF) larger than 0.4 in any model type are selected. (A) Mean inclusion frequencies, (B) mean estimated regression coefficients, and (C) Venn diagram of selected genes in all subgroups.*

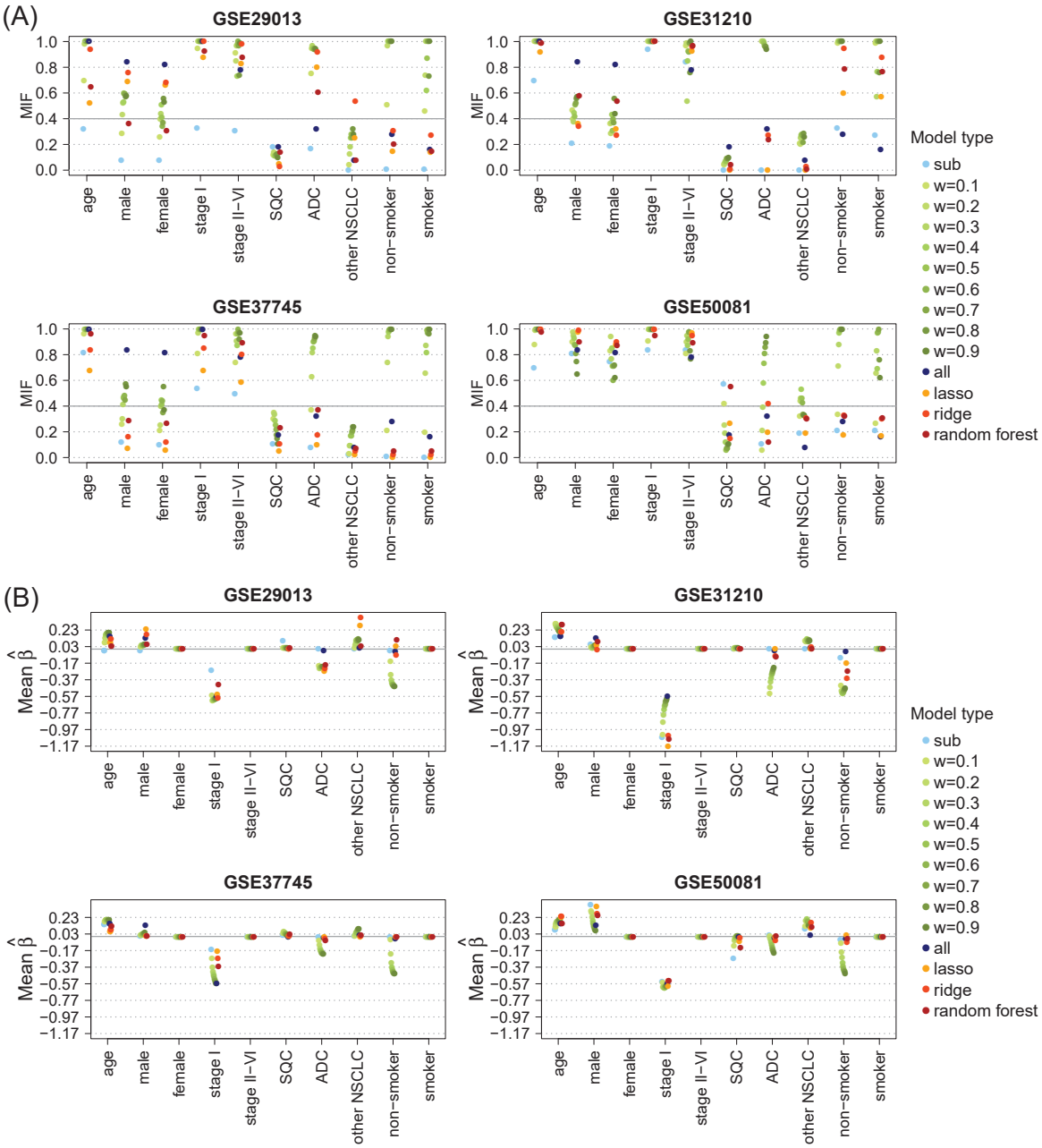


FIGURE B.31: *Results of Cox models including only clinical covariates. (A) Mean inclusion frequencies, and (B) mean estimated regression coefficients of clinical covariates in all subgroups.*

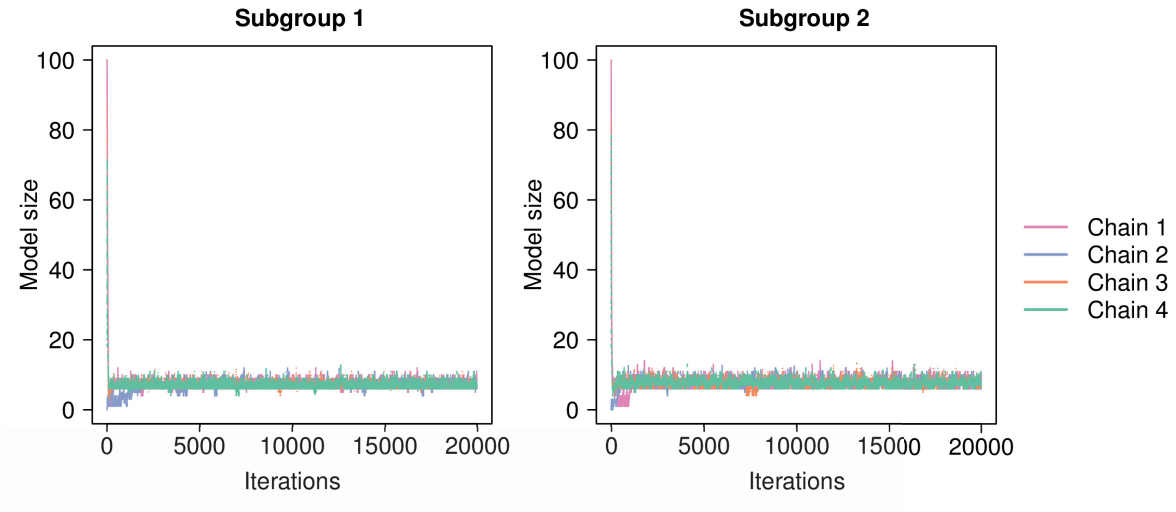


FIGURE B.32: *Trace plots of the model size (number of selected variables), comparing four independent Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected.*

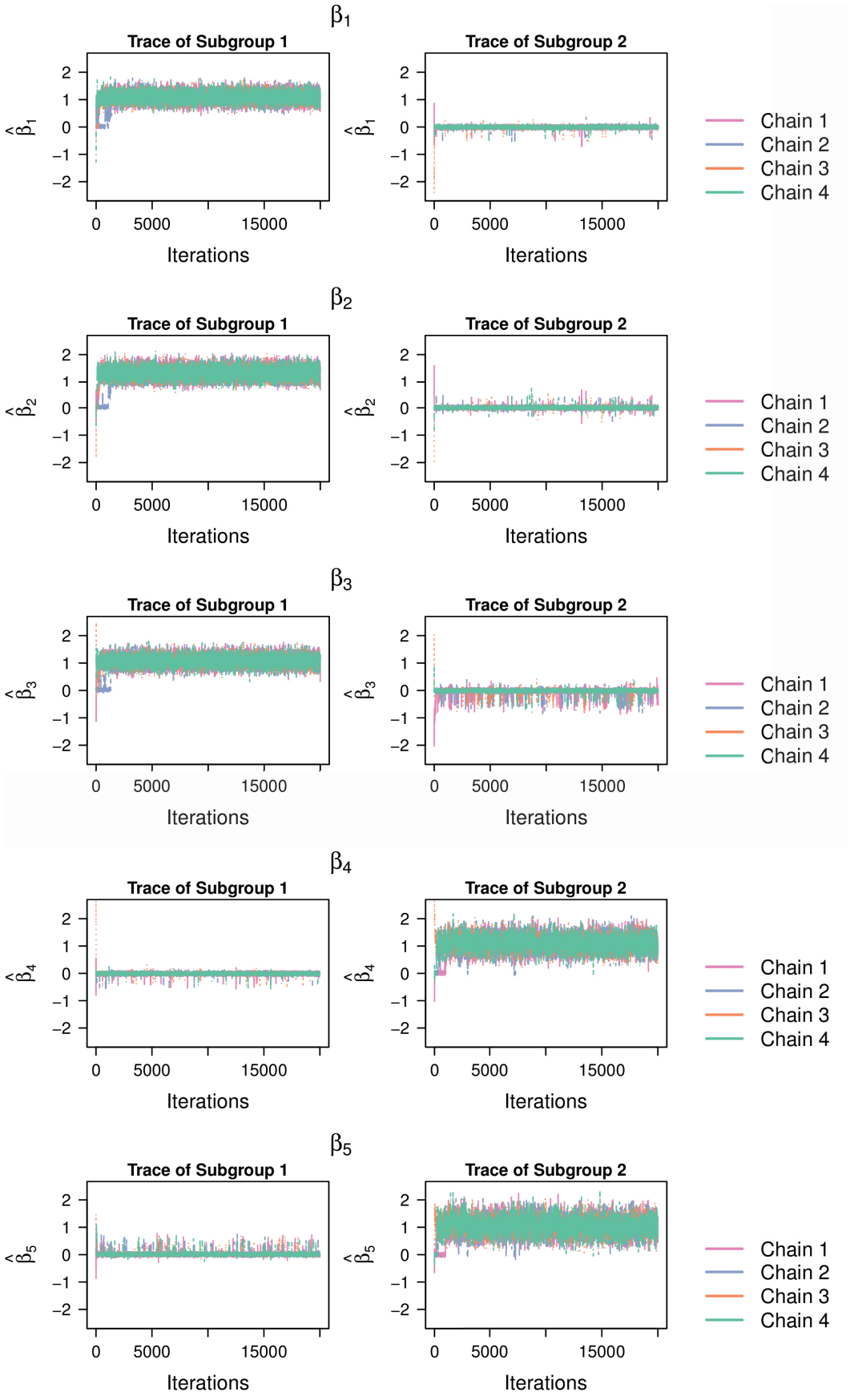


FIGURE B.33: *Trace plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable), comparing four independent Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected.*

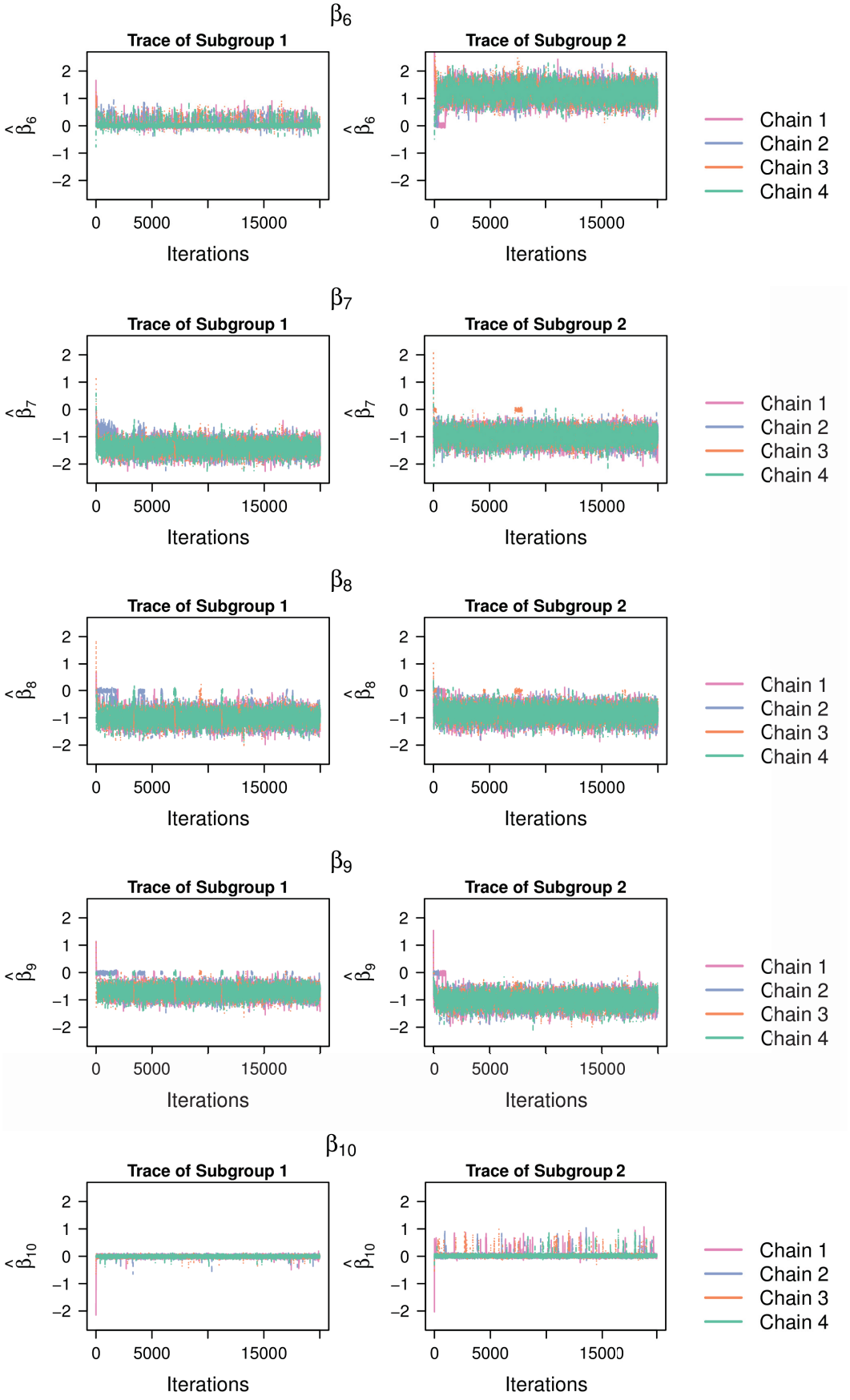


FIGURE B.33: *Trace plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable), comparing four independent Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

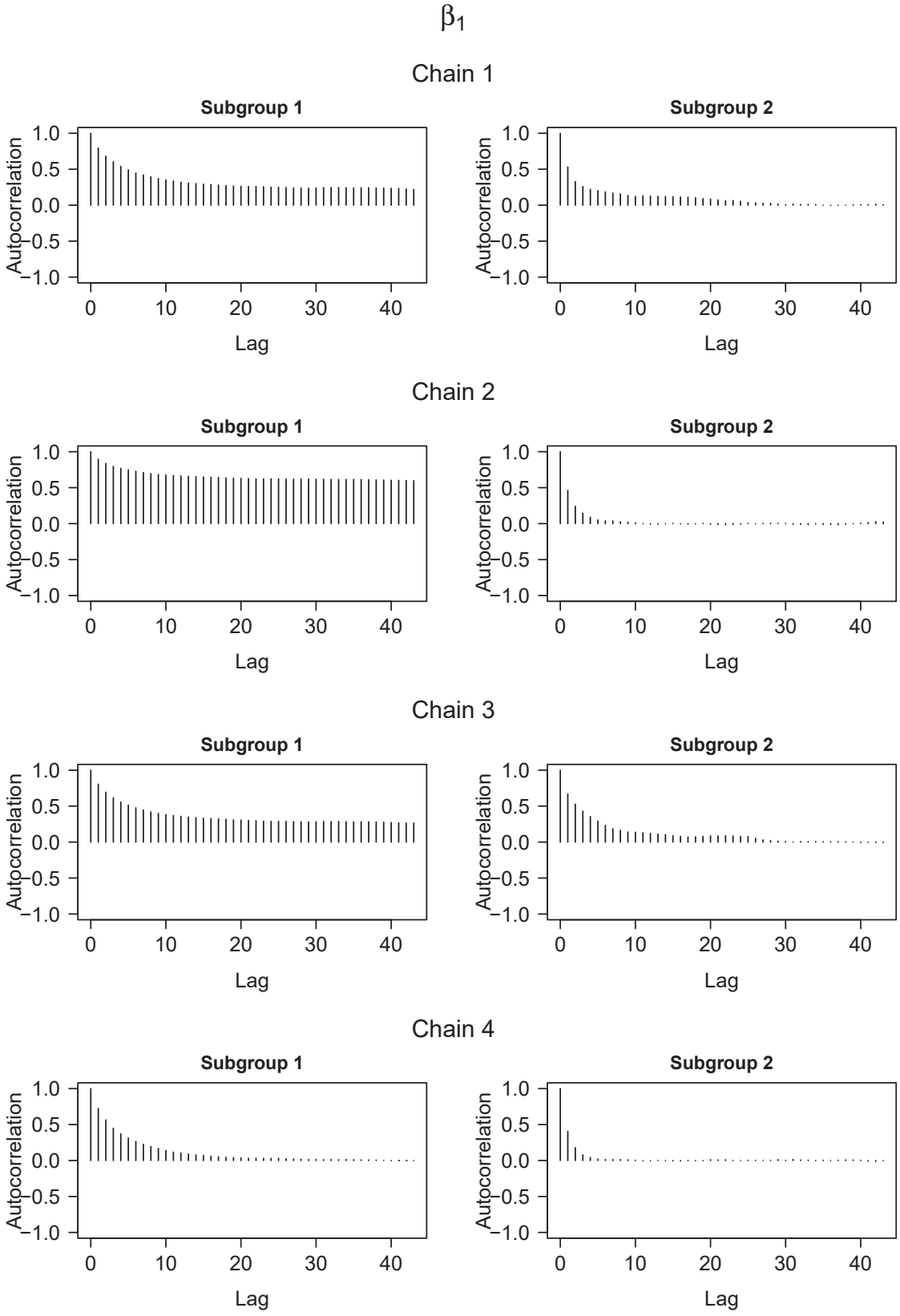


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected.*

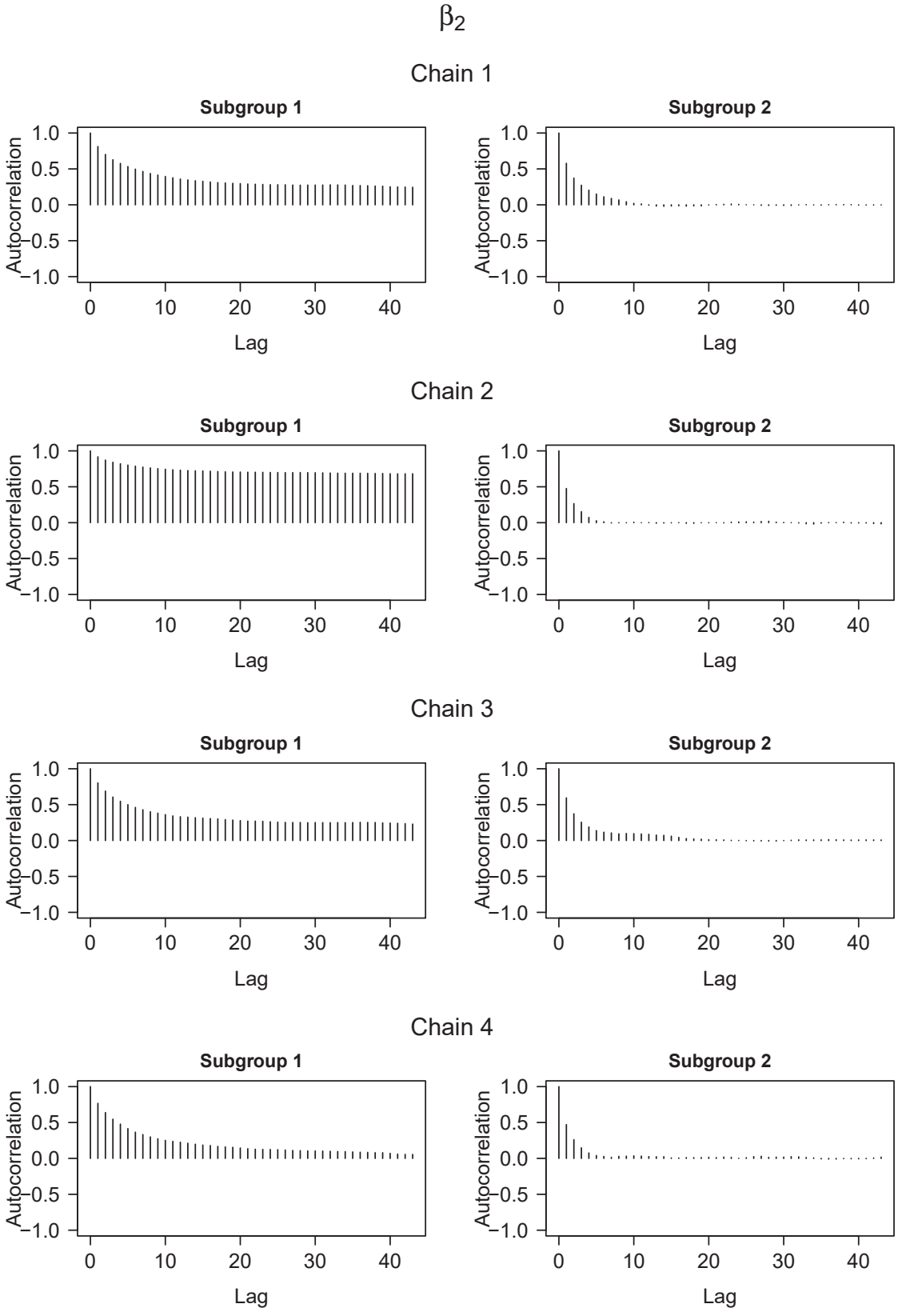


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

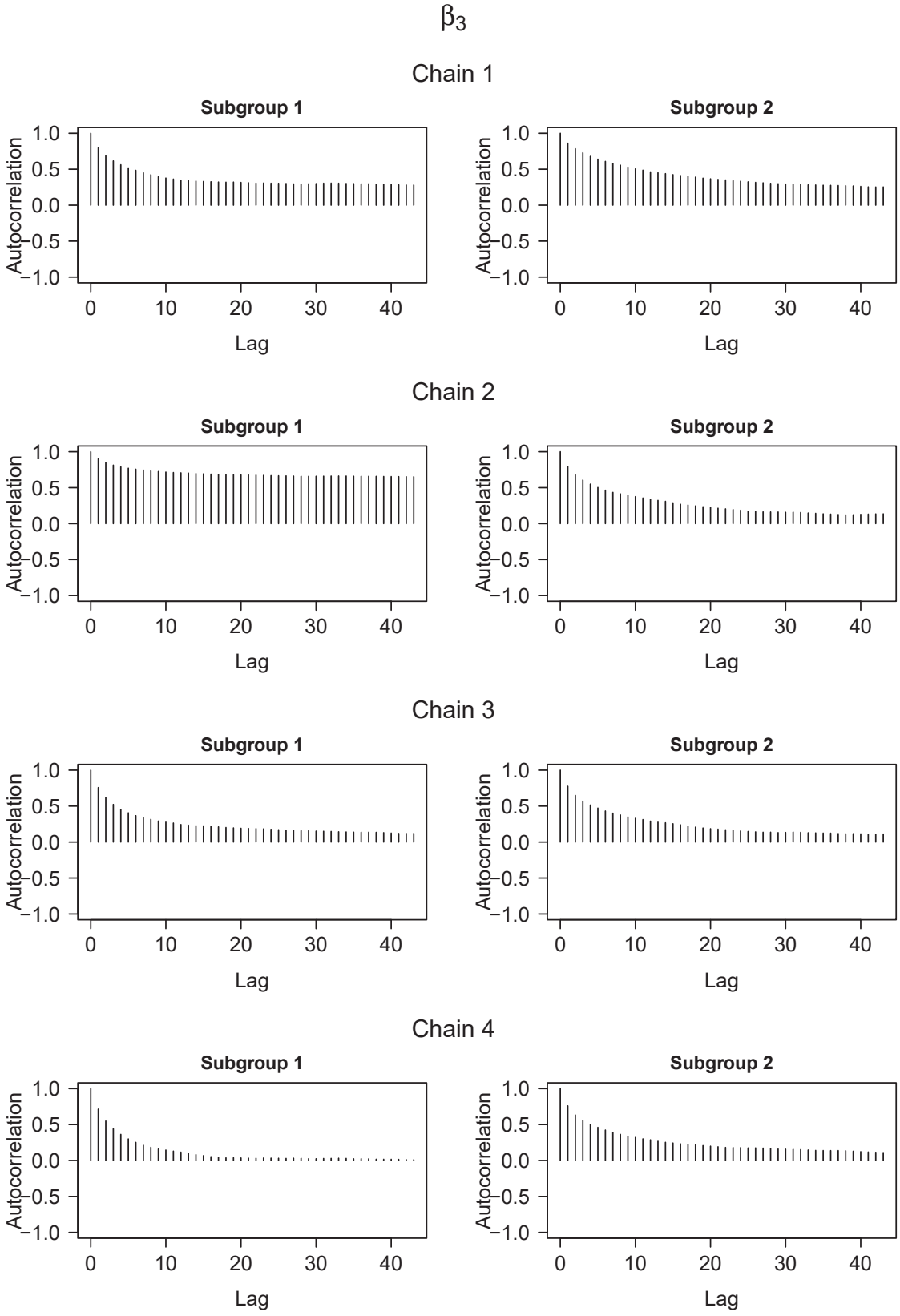


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

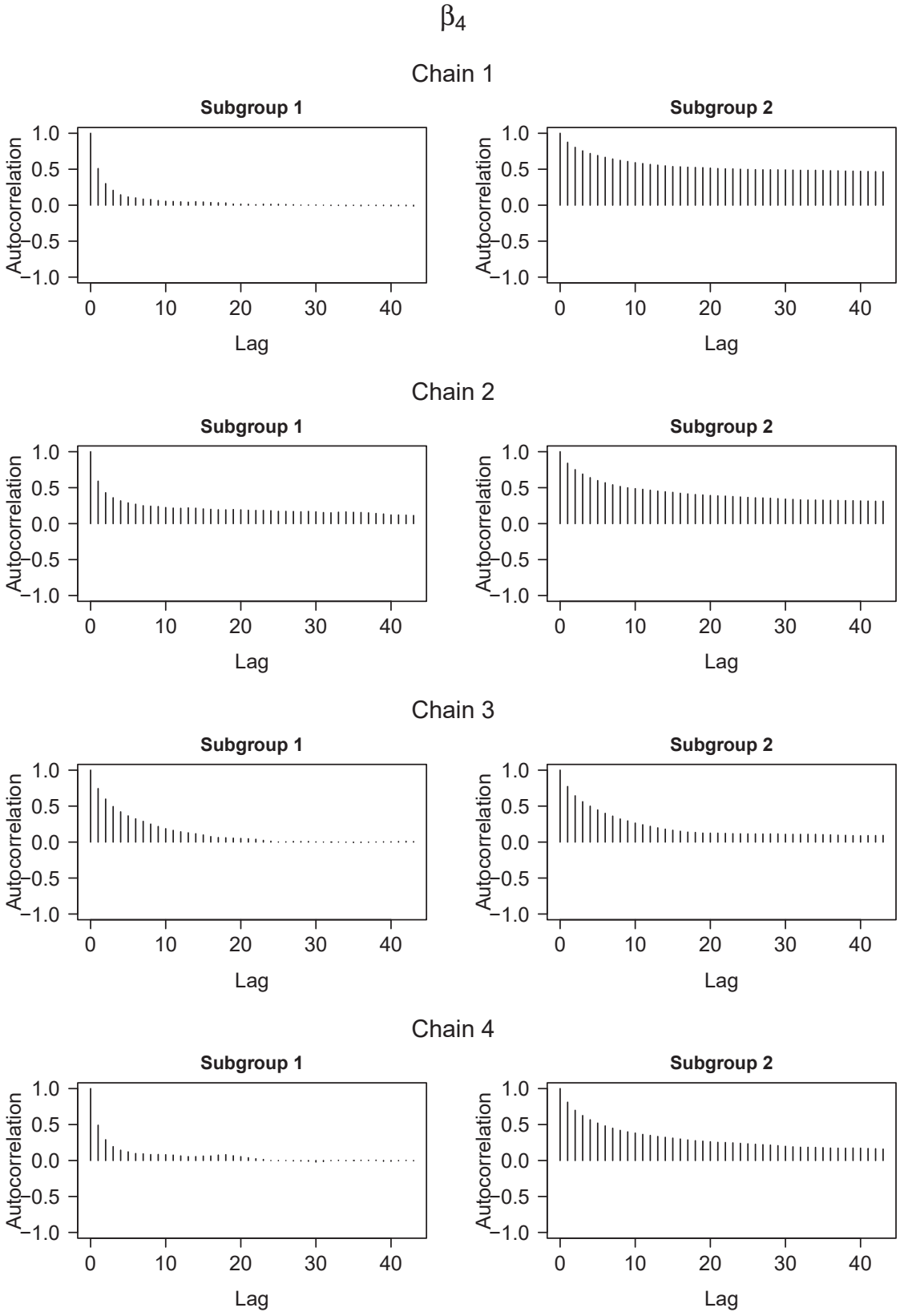


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

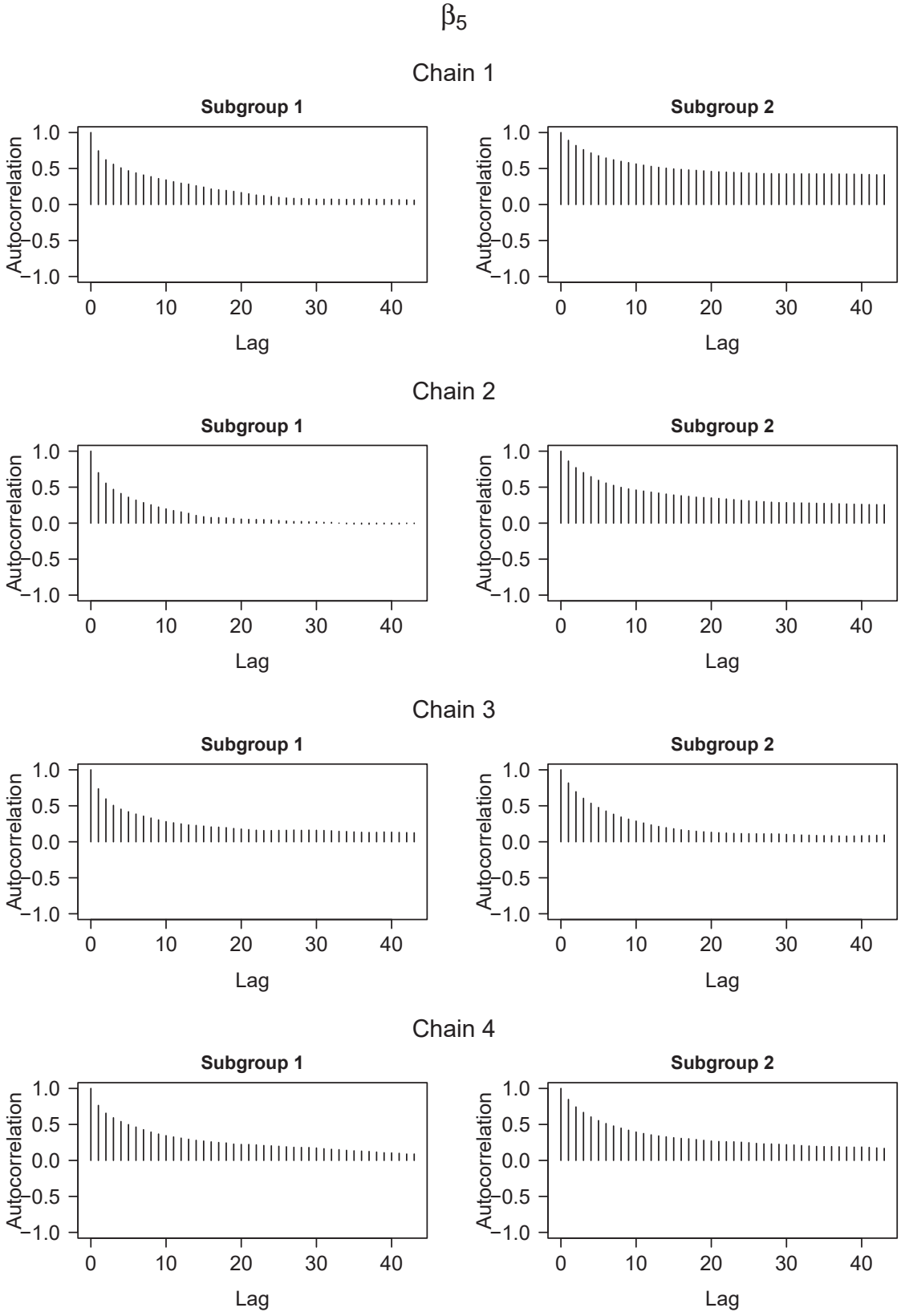


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

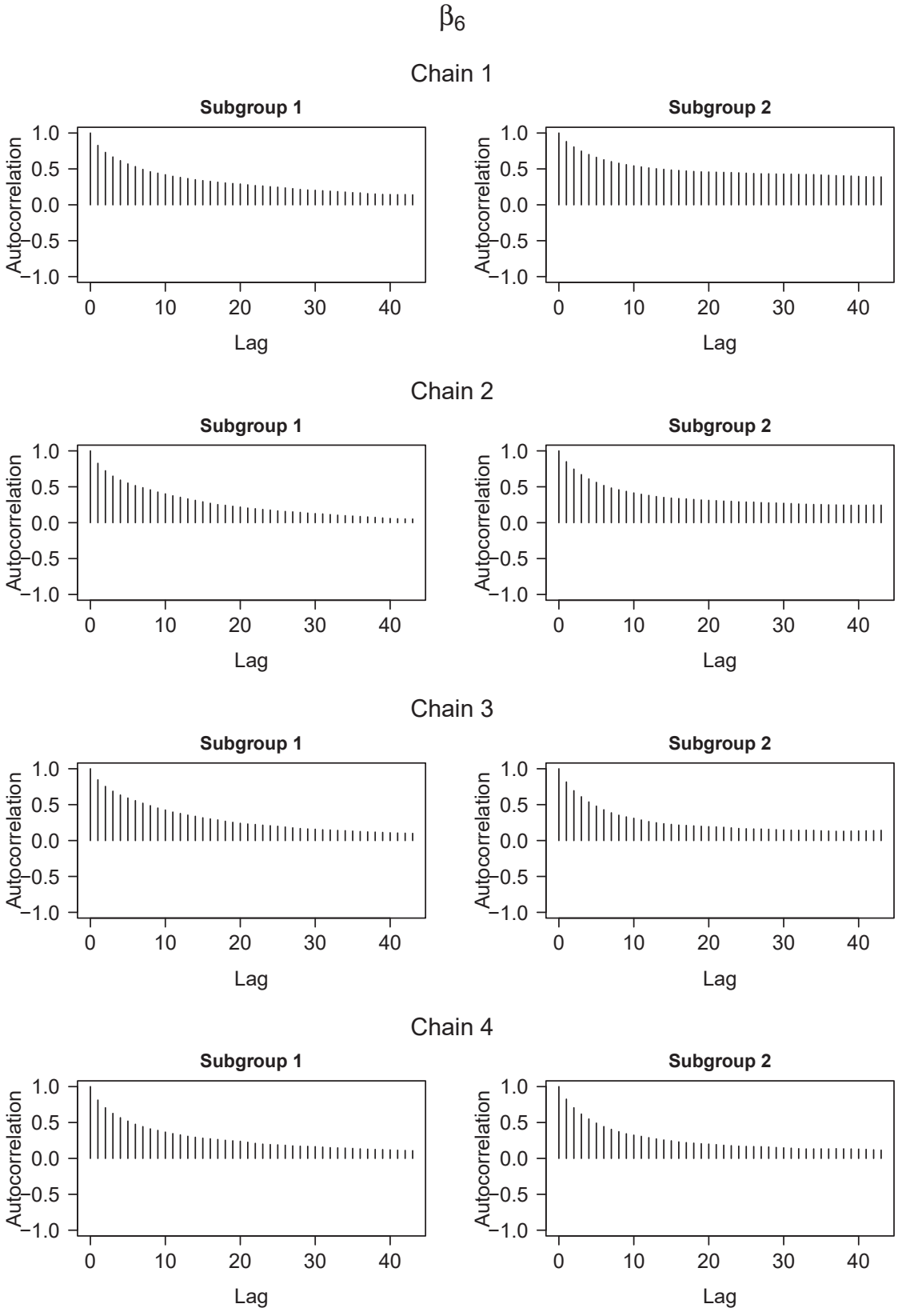


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

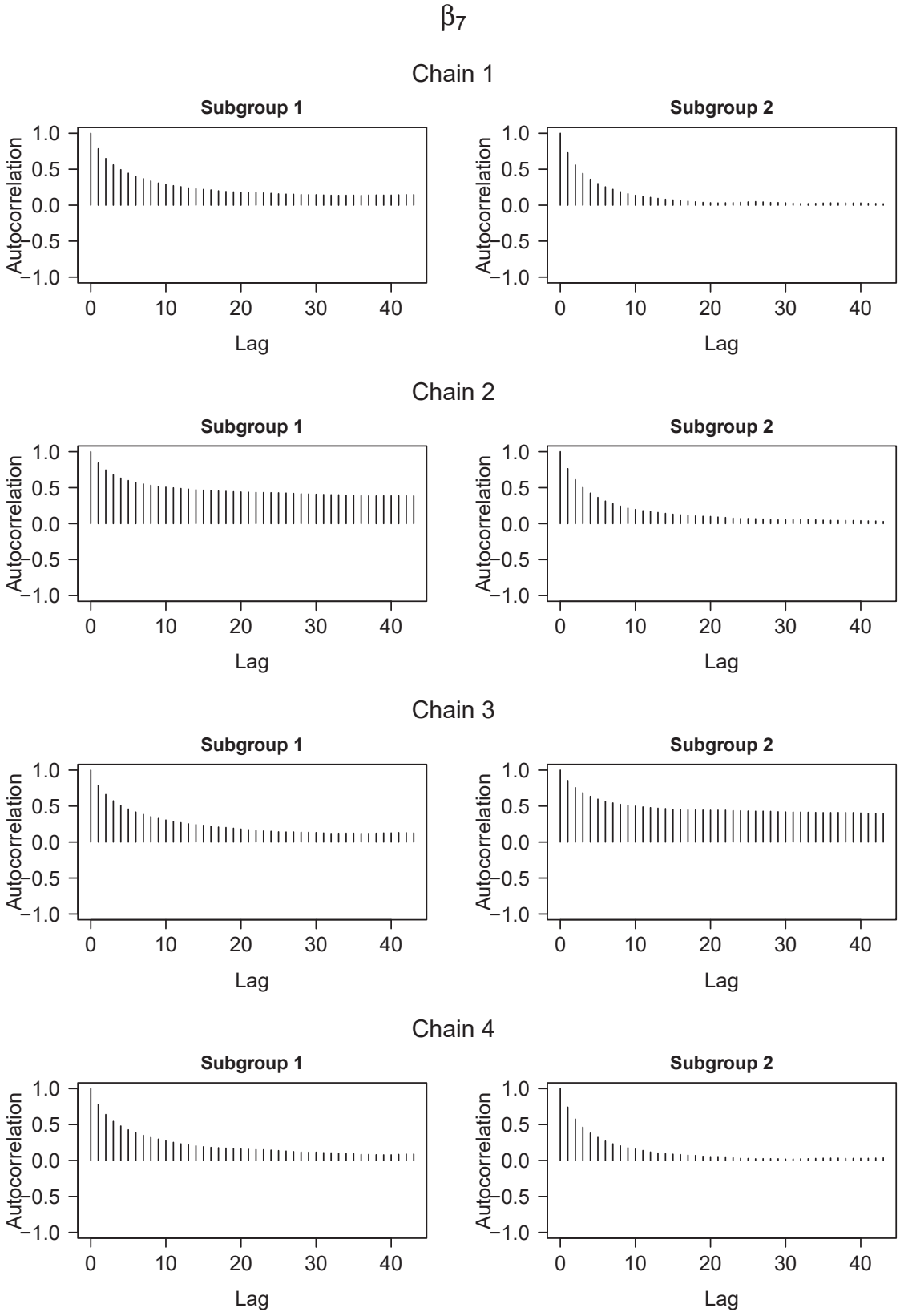


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

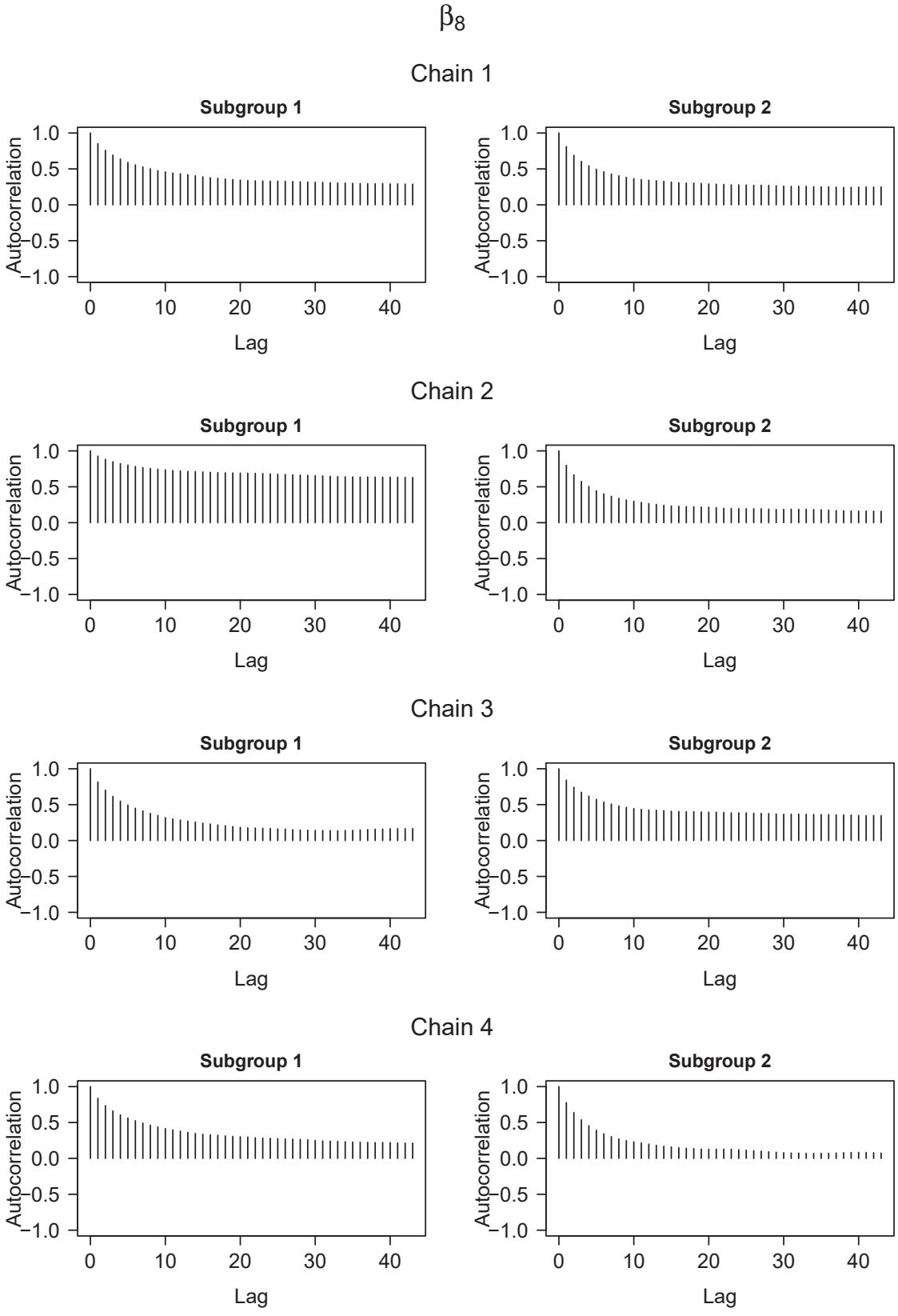


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

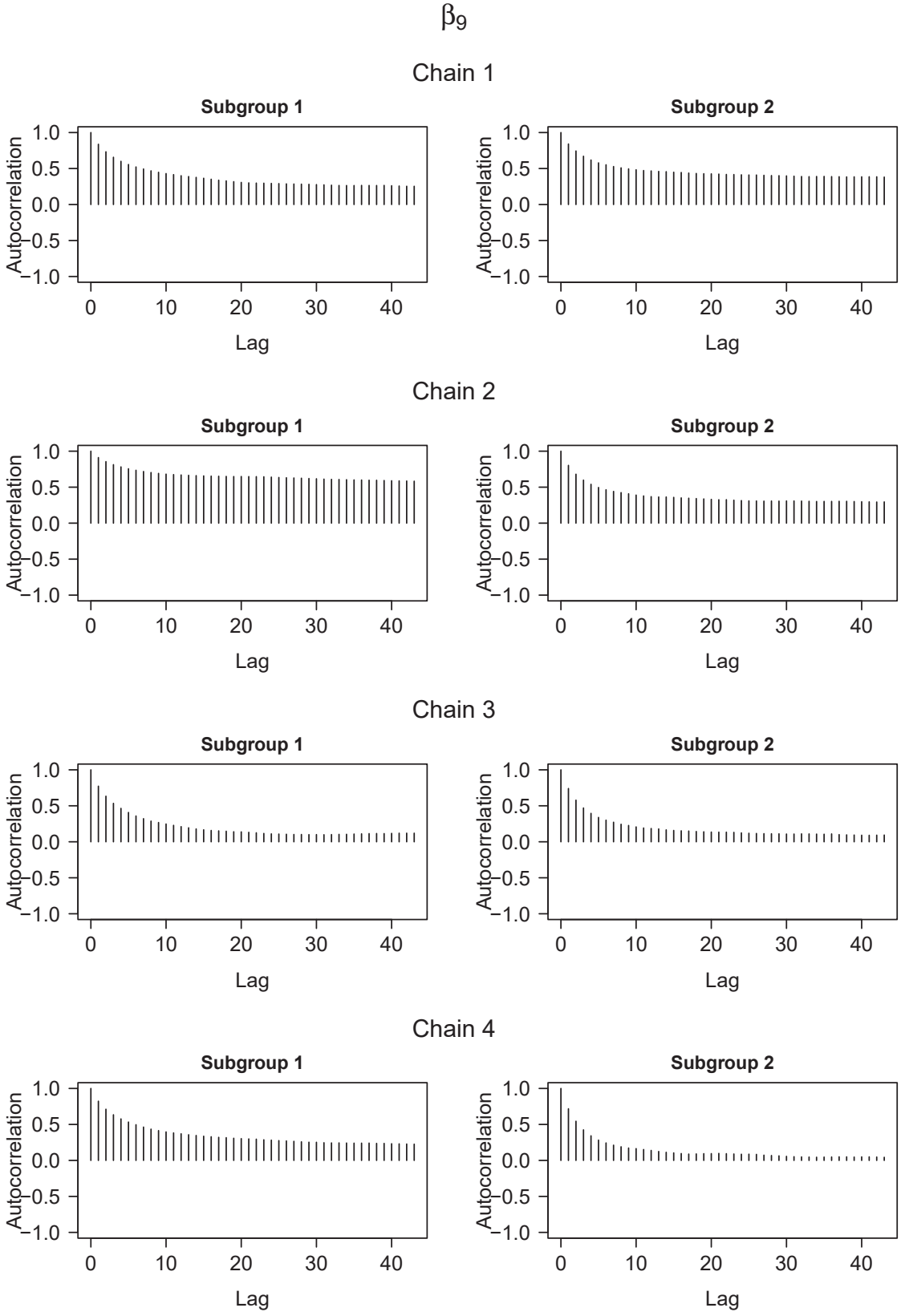


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

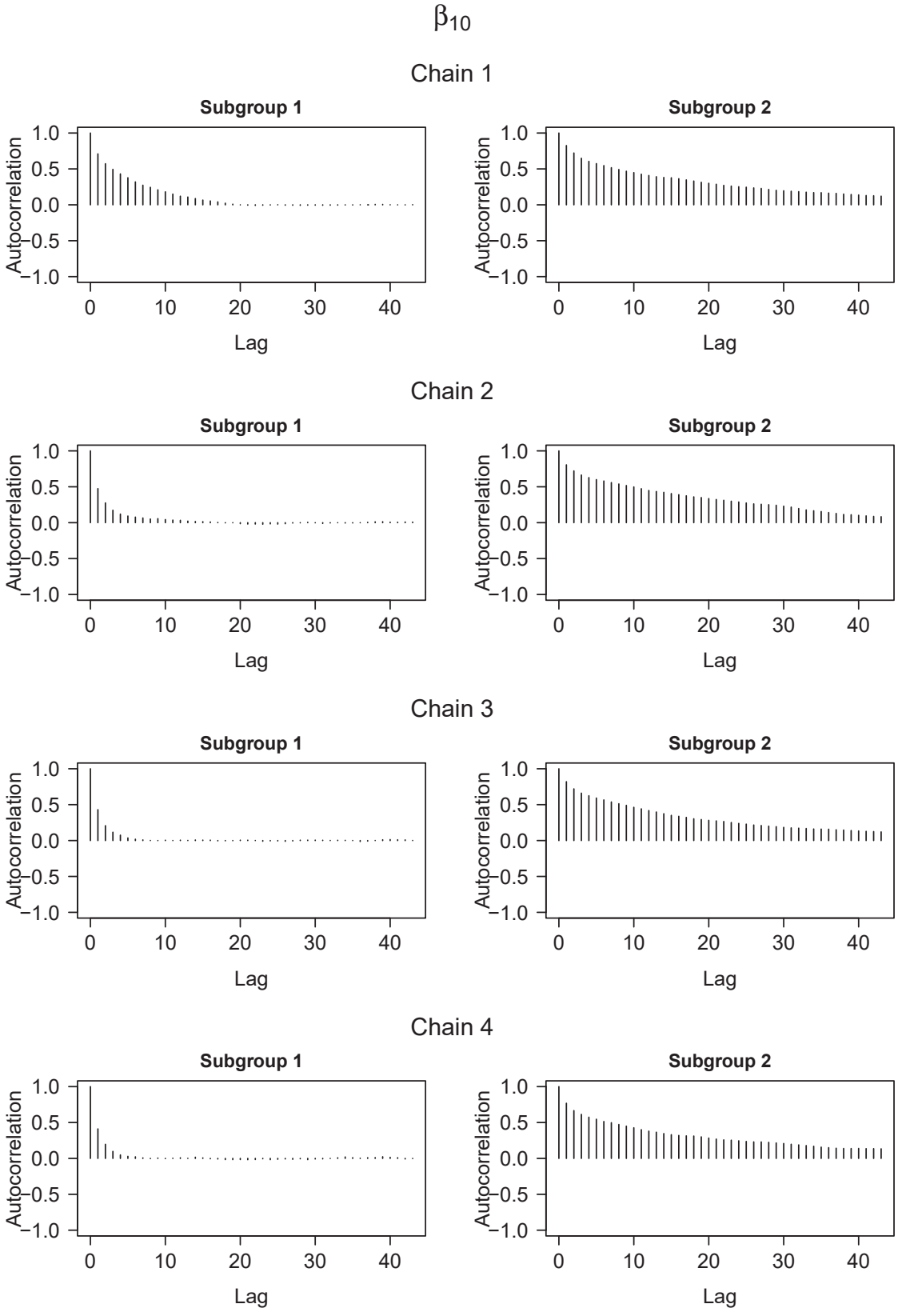


FIGURE B.34: *Autocorrelation plots of the estimated regression coefficients of the first ten variables (first nine prognostic variables and tenth variable as representative non-prognostic variable) in each of the four Markov chains. Chain 1: full model; Chain 2: empty model; Chain 3: 50% of variables and edges selected; Chain 4: 20% of variables and edges selected (cont.)*

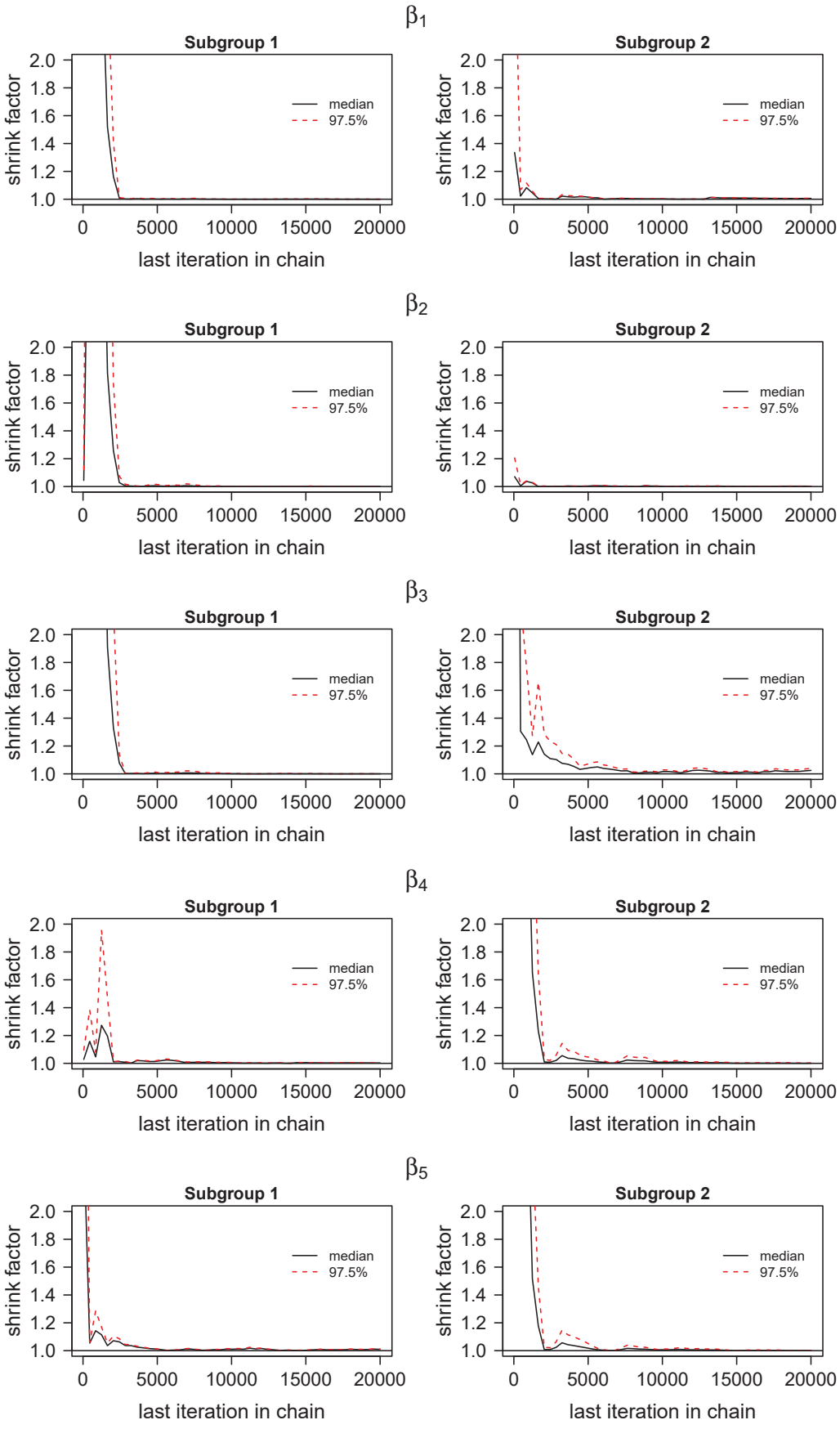


FIGURE B.35: *Plots of the corrected potential scale reduction factor by Brooks and Gelman (1998) of the estimated regression coefficients of the first ten variables.*

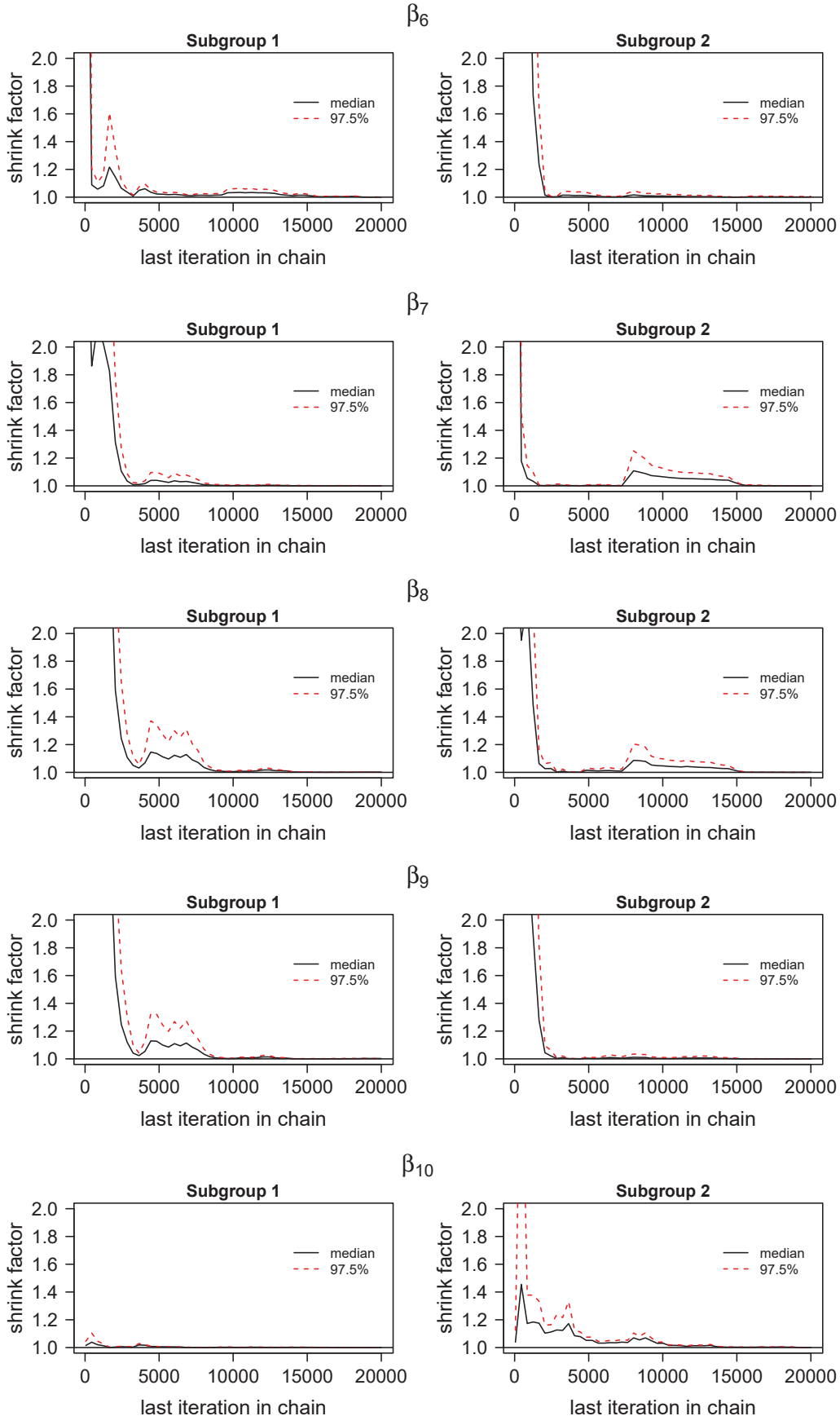


FIGURE B.35: *Plots of the corrected potential scale reduction factor by Brooks and Gelman (1998) of the estimated regression coefficients of the first ten variables (cont.)*

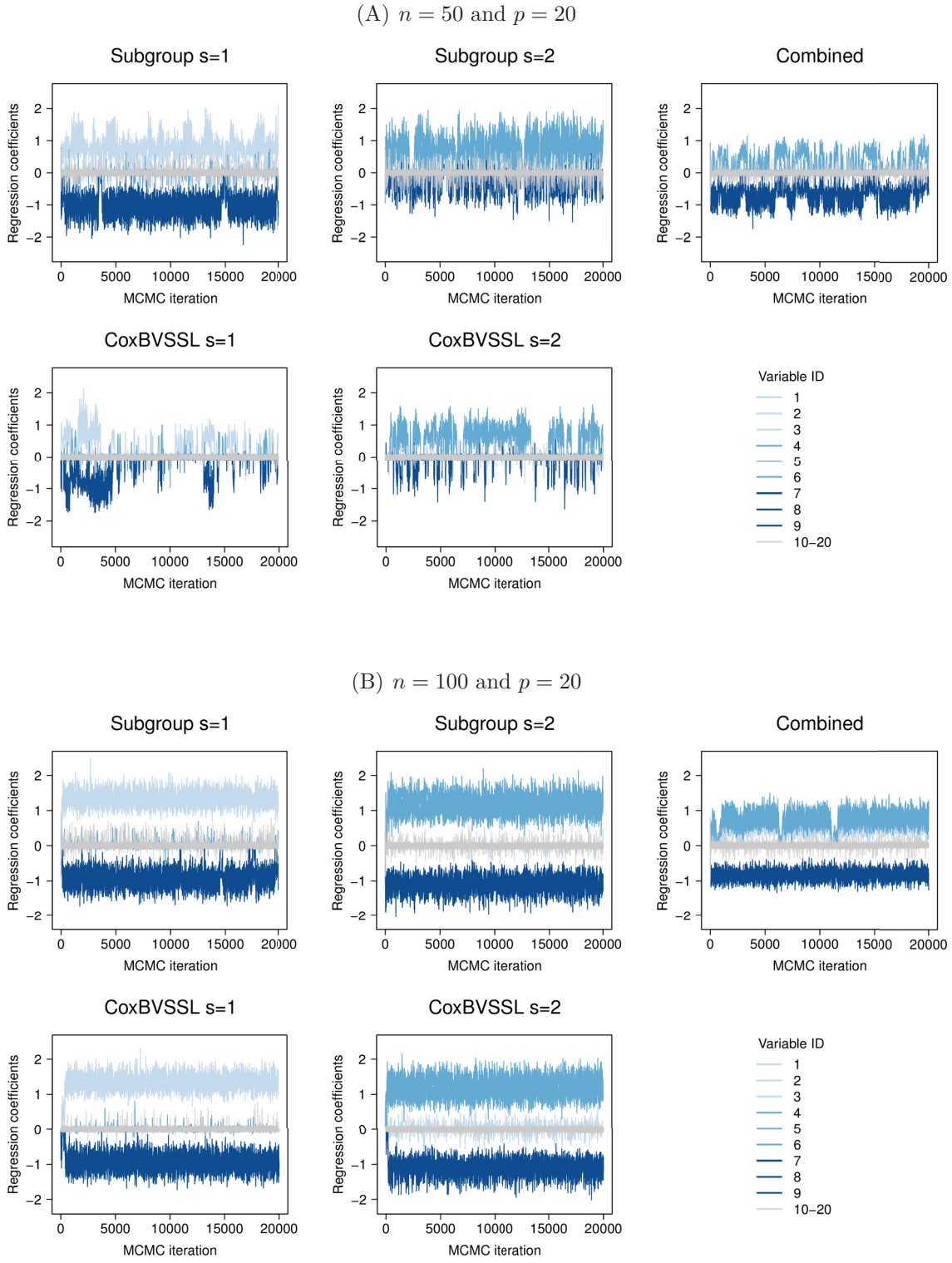


FIGURE B.36: *Trace plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and for varying n and p. (A) $n = 50$, $p = 20$; (B) $n = 100$, $p = 20$; (C) $n = 25$, $p = 100$; (D) $n = 50$, $p = 100$; (E) $n = 75$, $p = 100$; (F) $n = 100$, $p = 100$; (G) $n = 150$, $p = 100$.*

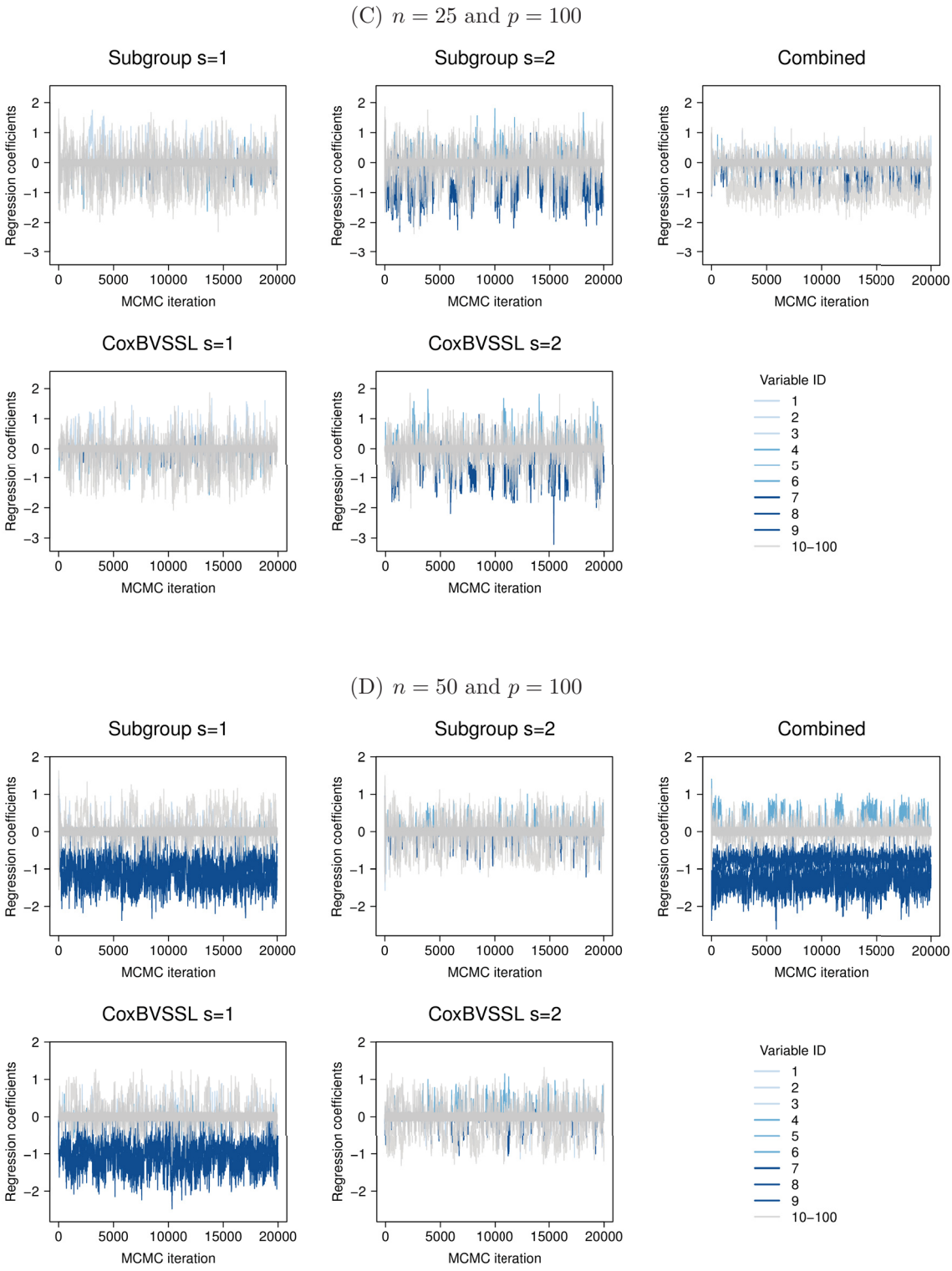


FIGURE B.36: *Trace plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and for varying n and p. (A) $n = 50$, $p = 20$; (B) $n = 100$, $p = 20$; (C) $n = 25$, $p = 100$; (D) $n = 50$, $p = 100$; (E) $n = 75$, $p = 100$; (F) $n = 100$, $p = 100$; (G) $n = 150$, $p = 100$ (cont.).*

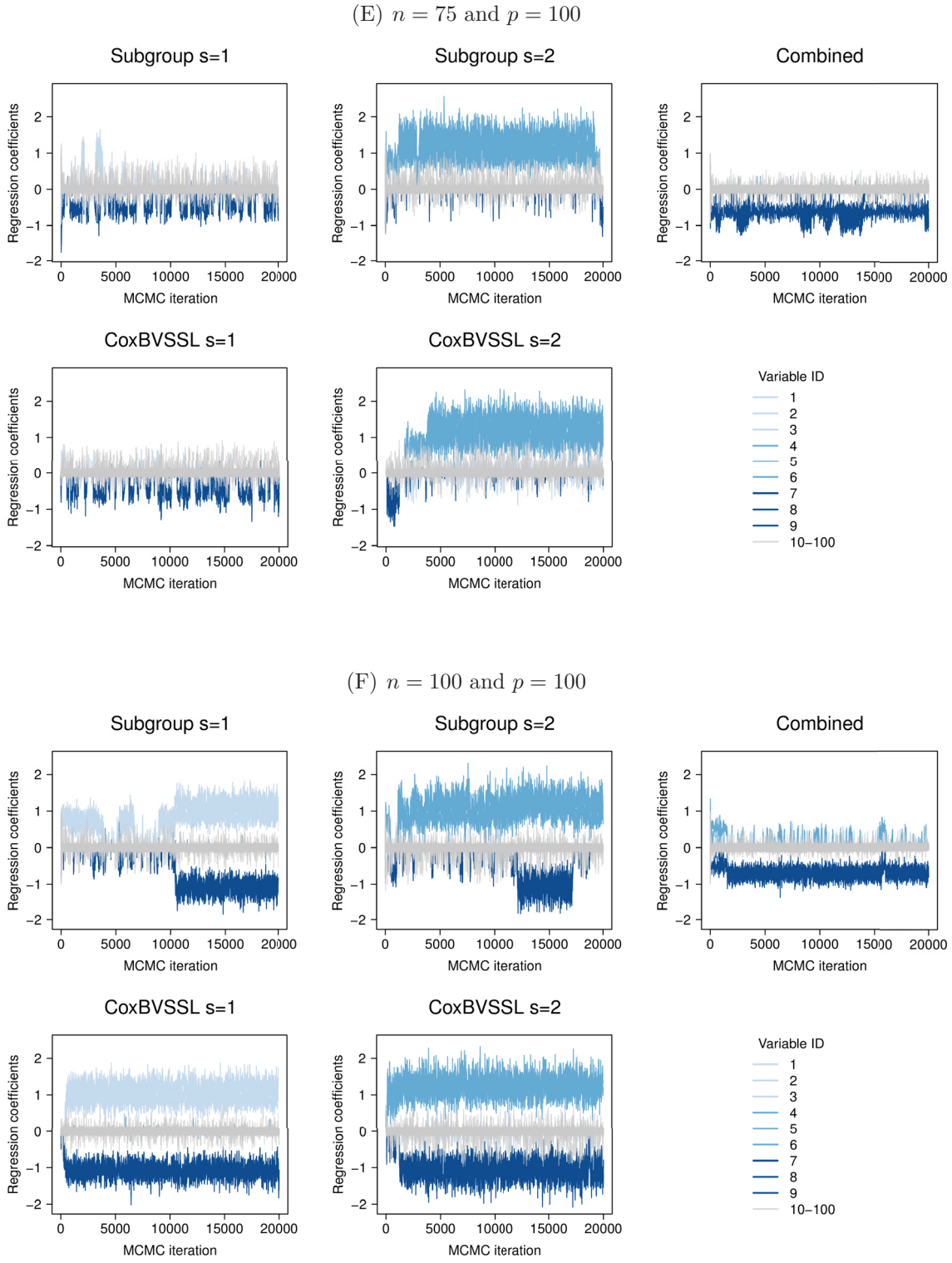


FIGURE B.36: *Trace plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and for varying n and p. (A) $n = 50$, $p = 20$; (B) $n = 100$, $p = 20$; (C) $n = 25$, $p = 100$; (D) $n = 50$, $p = 100$; (E) $n = 75$, $p = 100$; (F) $n = 100$, $p = 100$; (G) $n = 150$, $p = 100$ (cont.).*

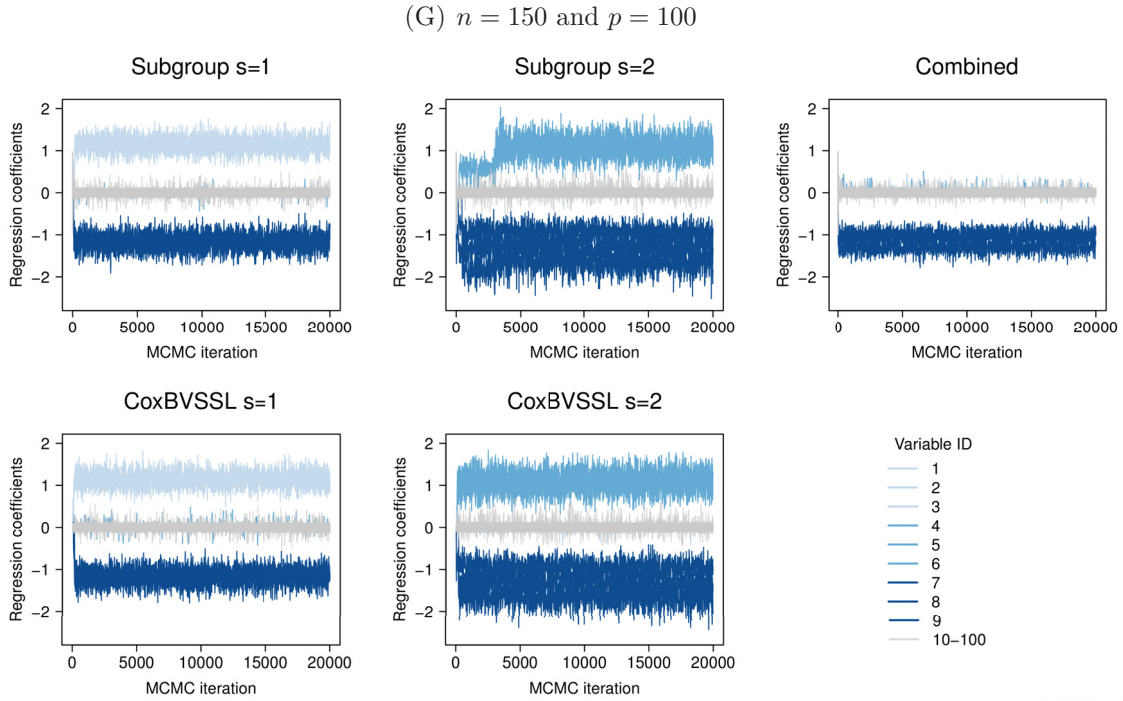


FIGURE B.36: *Trace plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and for varying n and p. (A) $n = 50$, $p = 20$; (B) $n = 100$, $p = 20$; (C) $n = 25$, $p = 100$; (D) $n = 50$, $p = 100$; (E) $n = 75$, $p = 100$; (F) $n = 100$, $p = 100$; (G) $n = 150$, $p = 100$ (cont.).*

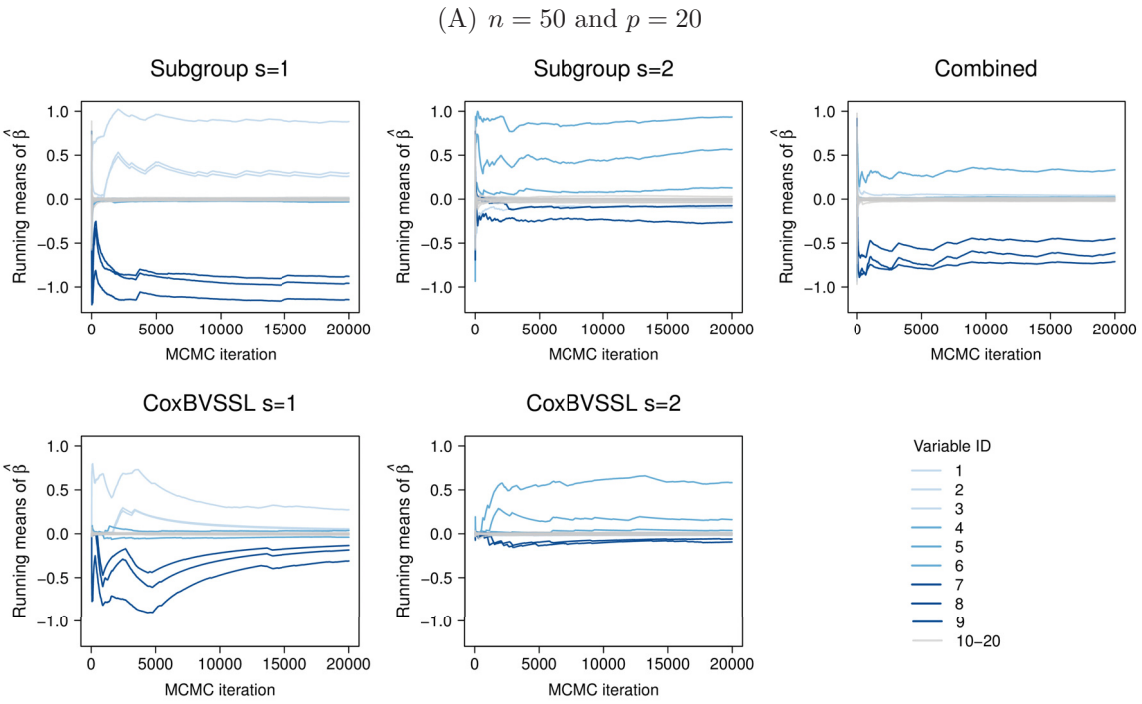
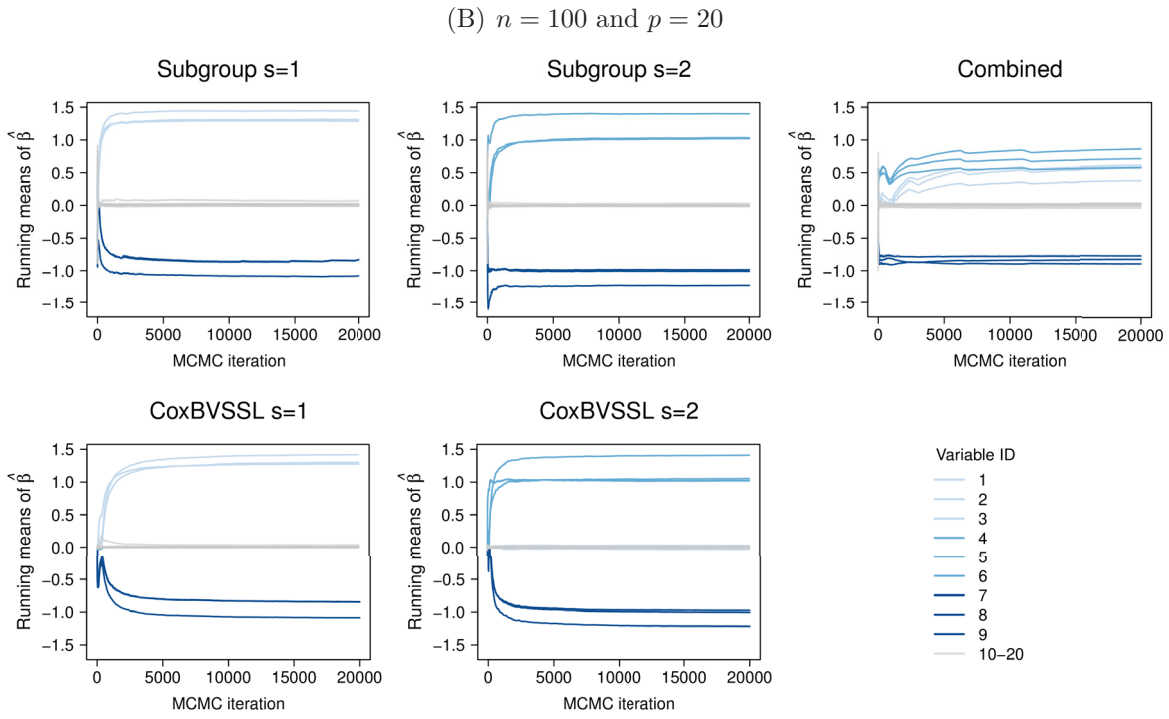


FIGURE B.37: *Running mean plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and for varying n and p.*



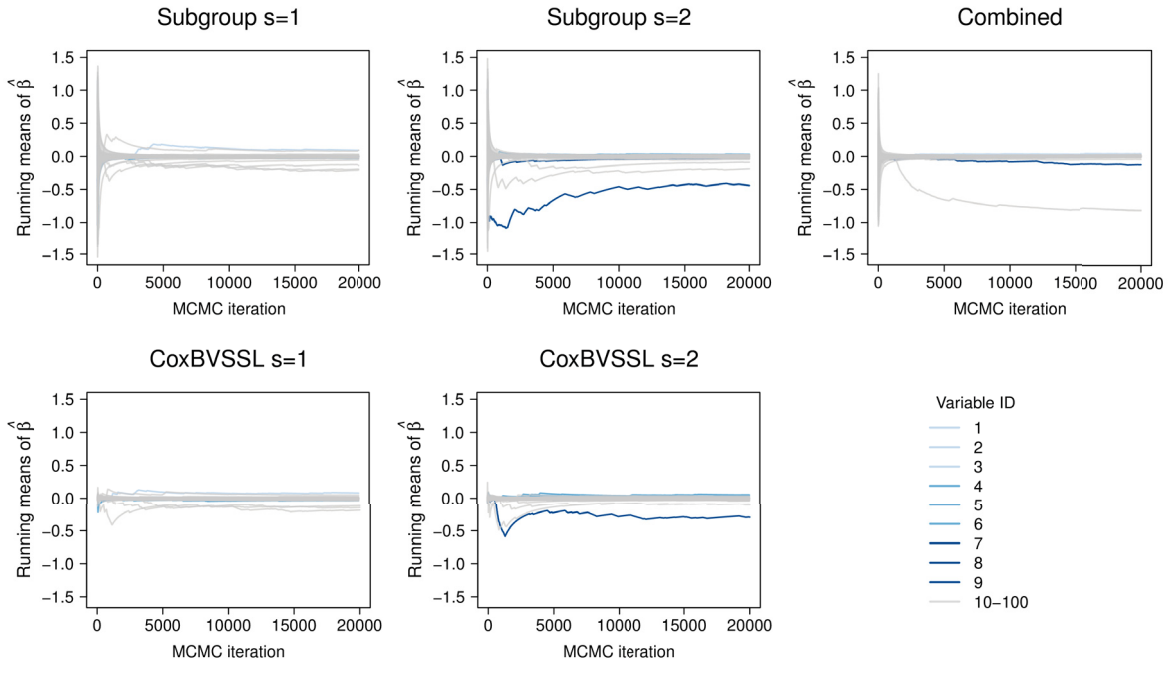
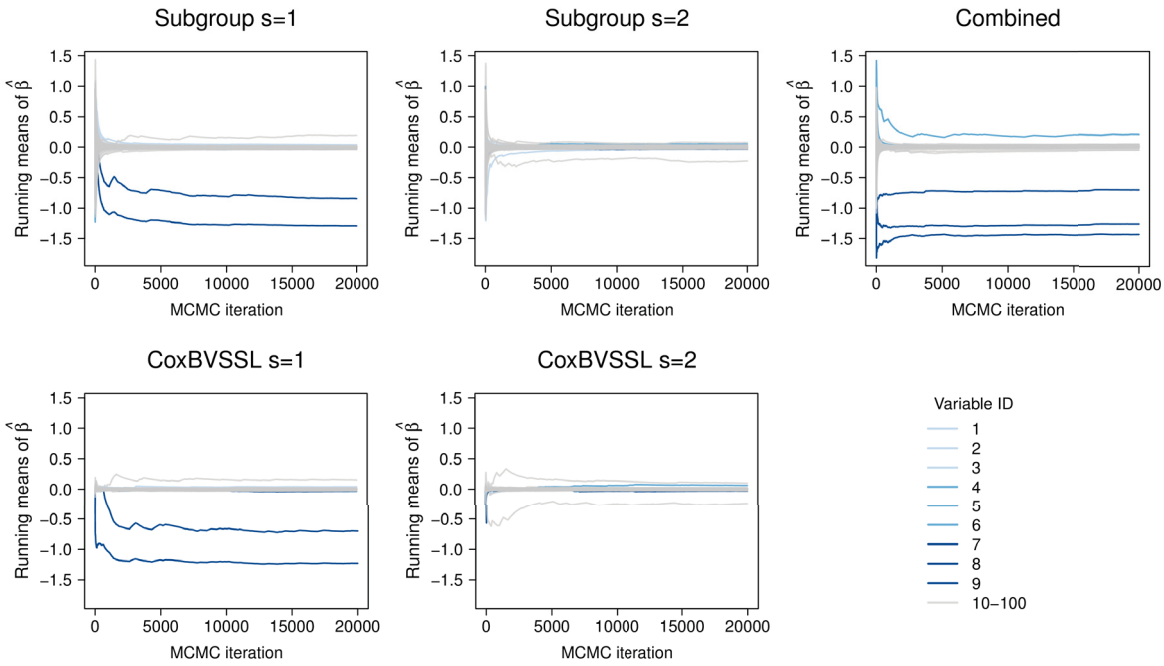


FIGURE B.37: *Running mean plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and for varying n and p (cont.).*

(D) $n = 50$ and $p = 100$



(E) $n = 75$ and $p = 100$

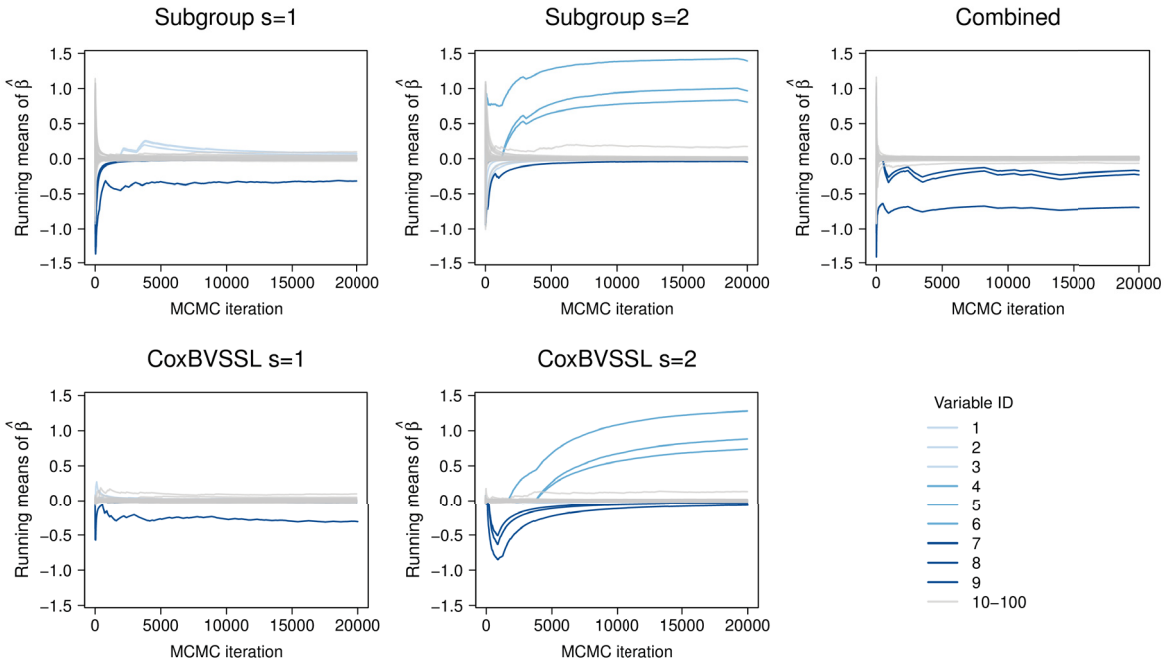


FIGURE B.37: *Running mean plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and for varying n and p (cont.).*

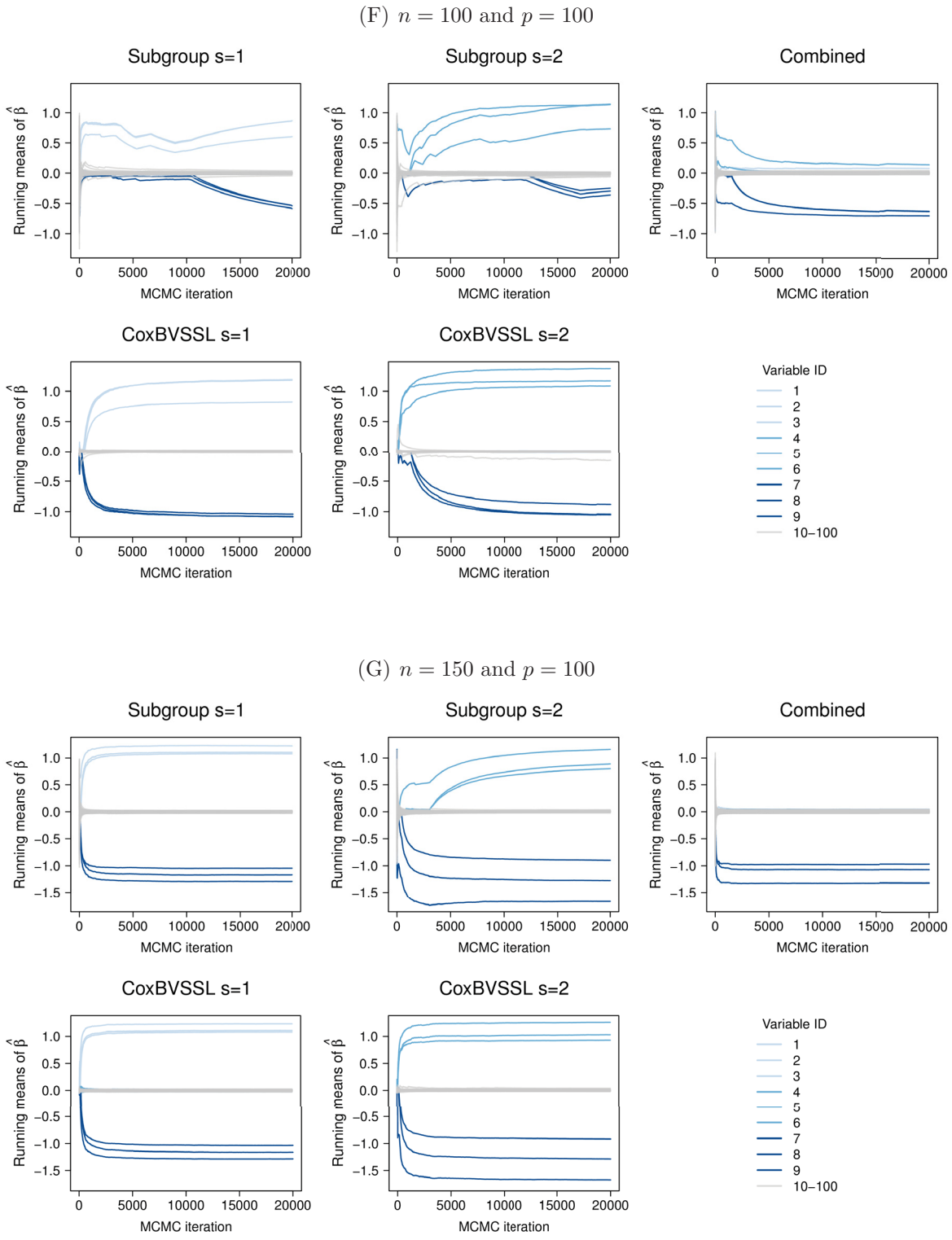


FIGURE B.37: *Running mean plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and for varying n and p (cont.).*

(A) $n = 50$ and $p = 20$

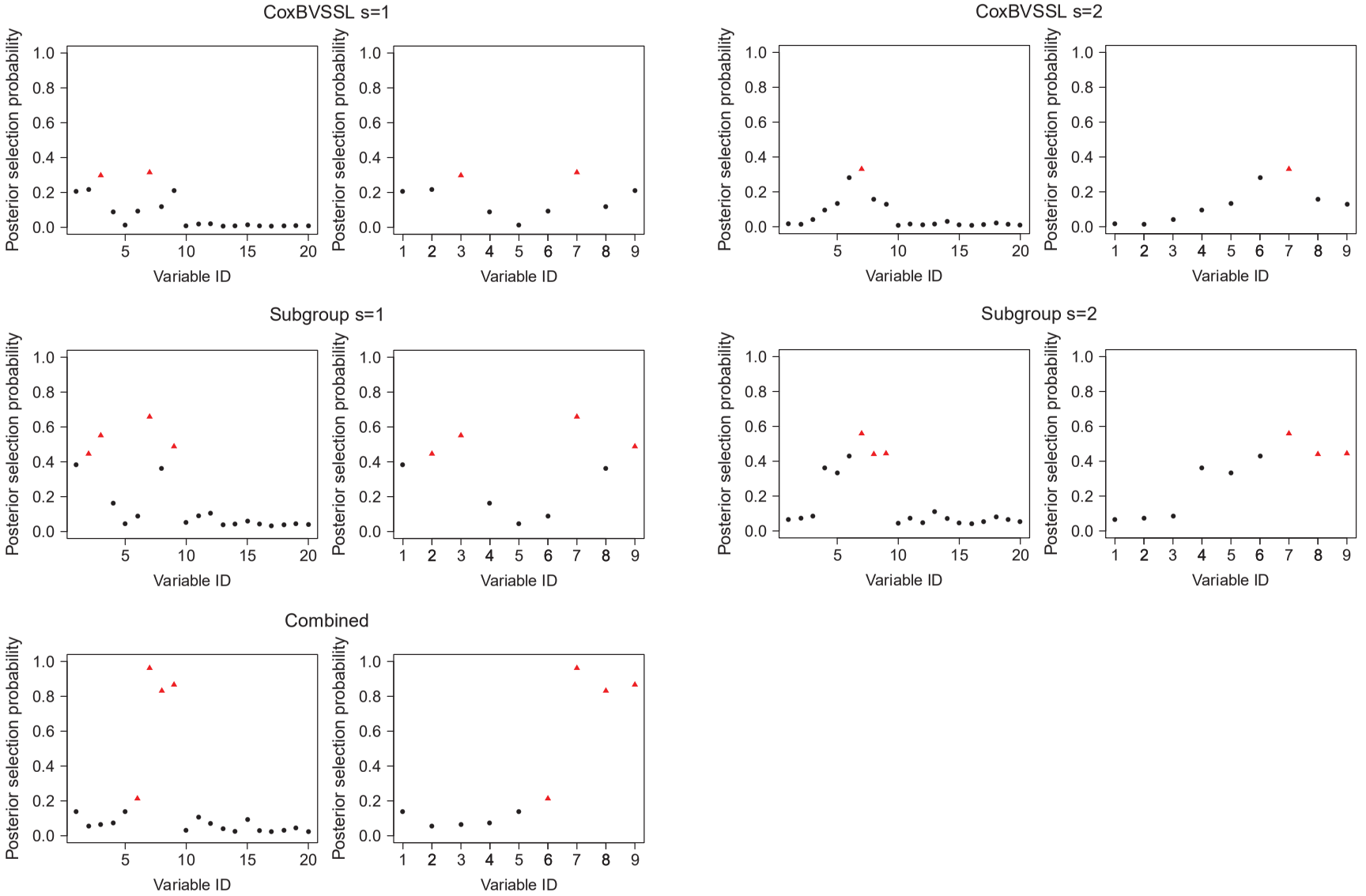


FIGURE B.38: *Mean posterior inclusion probabilities (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles.*

(B) $n = 100$ and $p = 20$

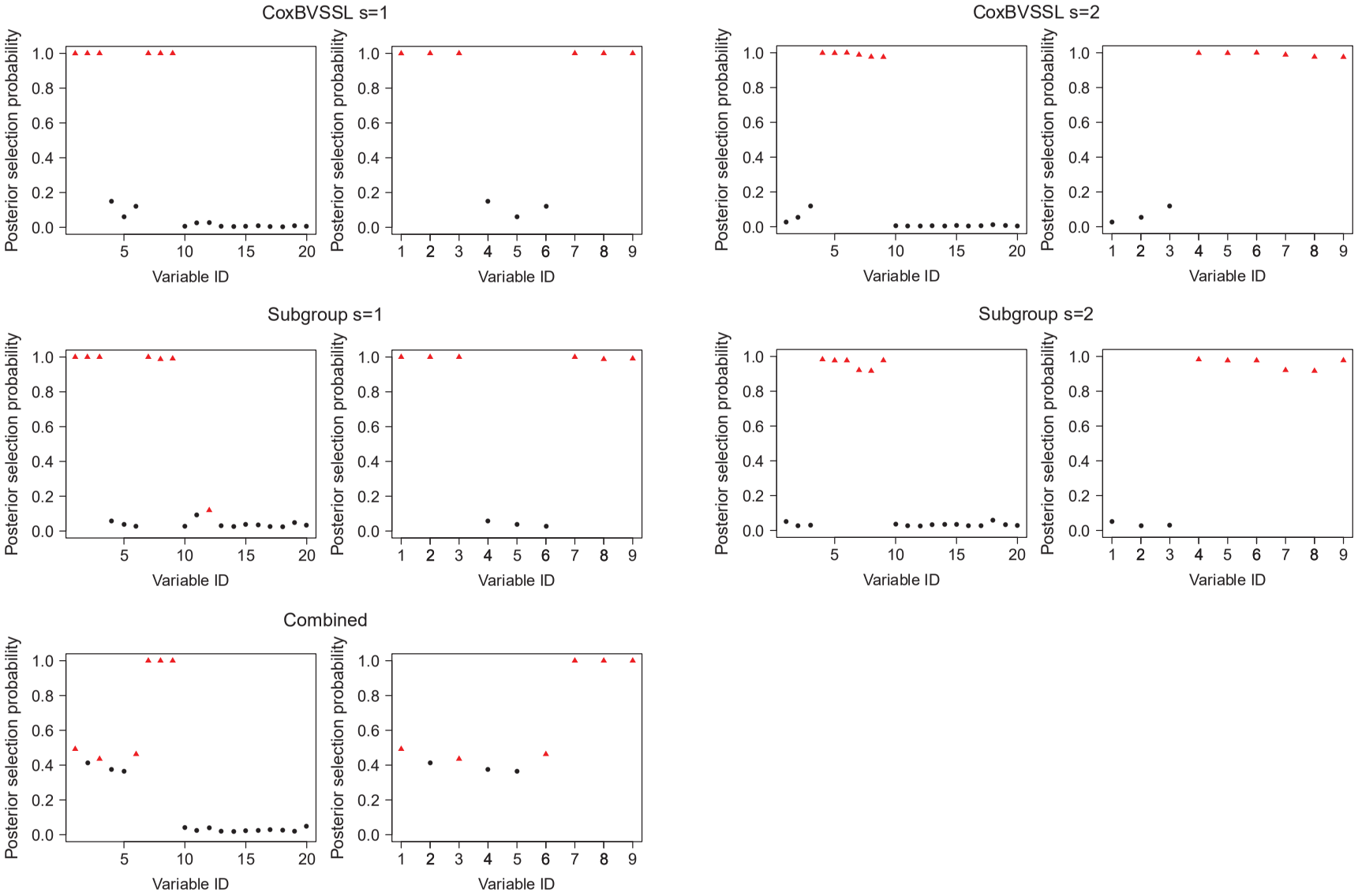


FIGURE B.38: *Mean posterior inclusion probabilities (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(C) $n = 25$ and $p = 100$

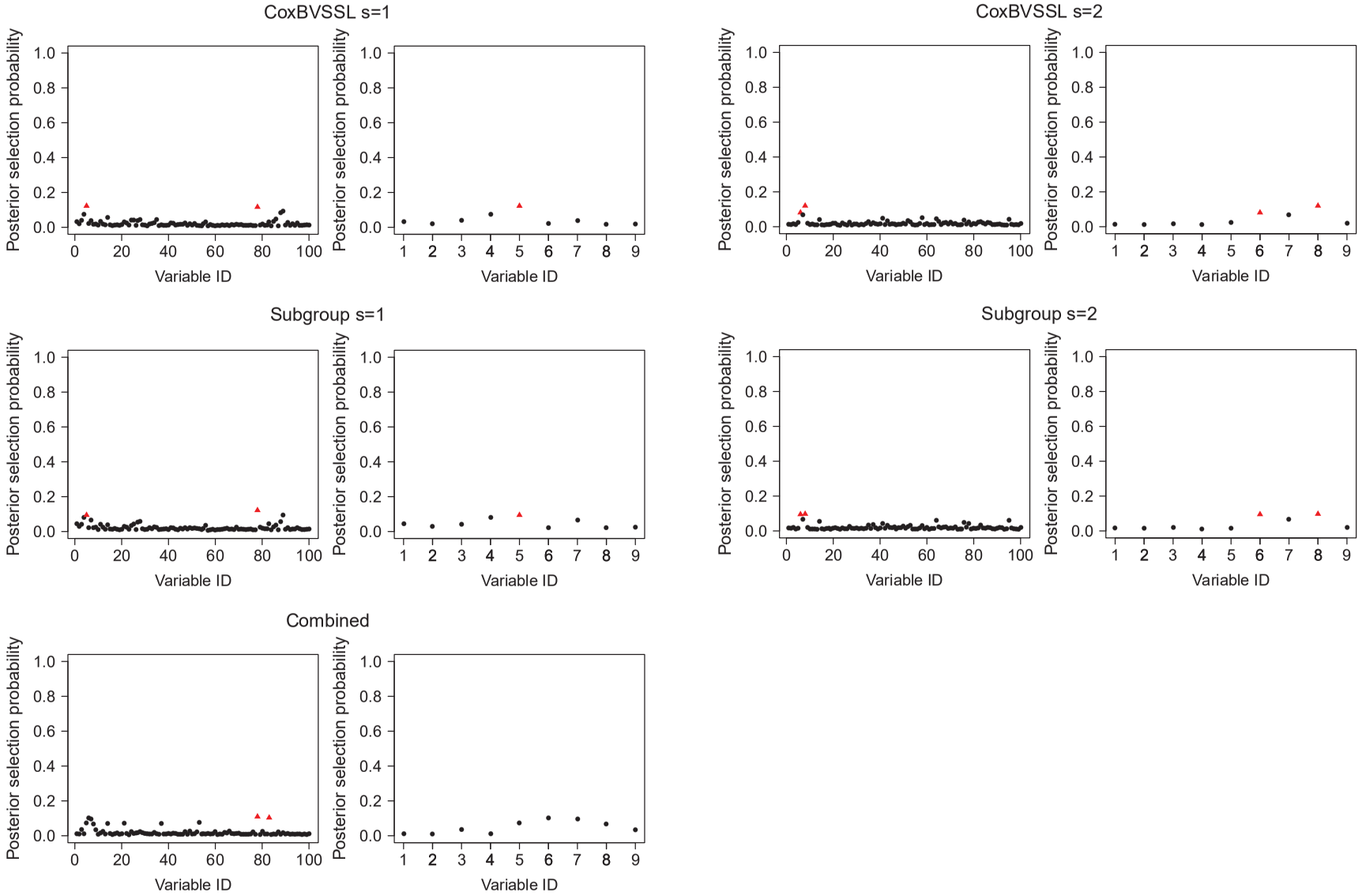


FIGURE B.38: *Mean posterior inclusion probabilities (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(D) $n = 50$ and $p = 100$

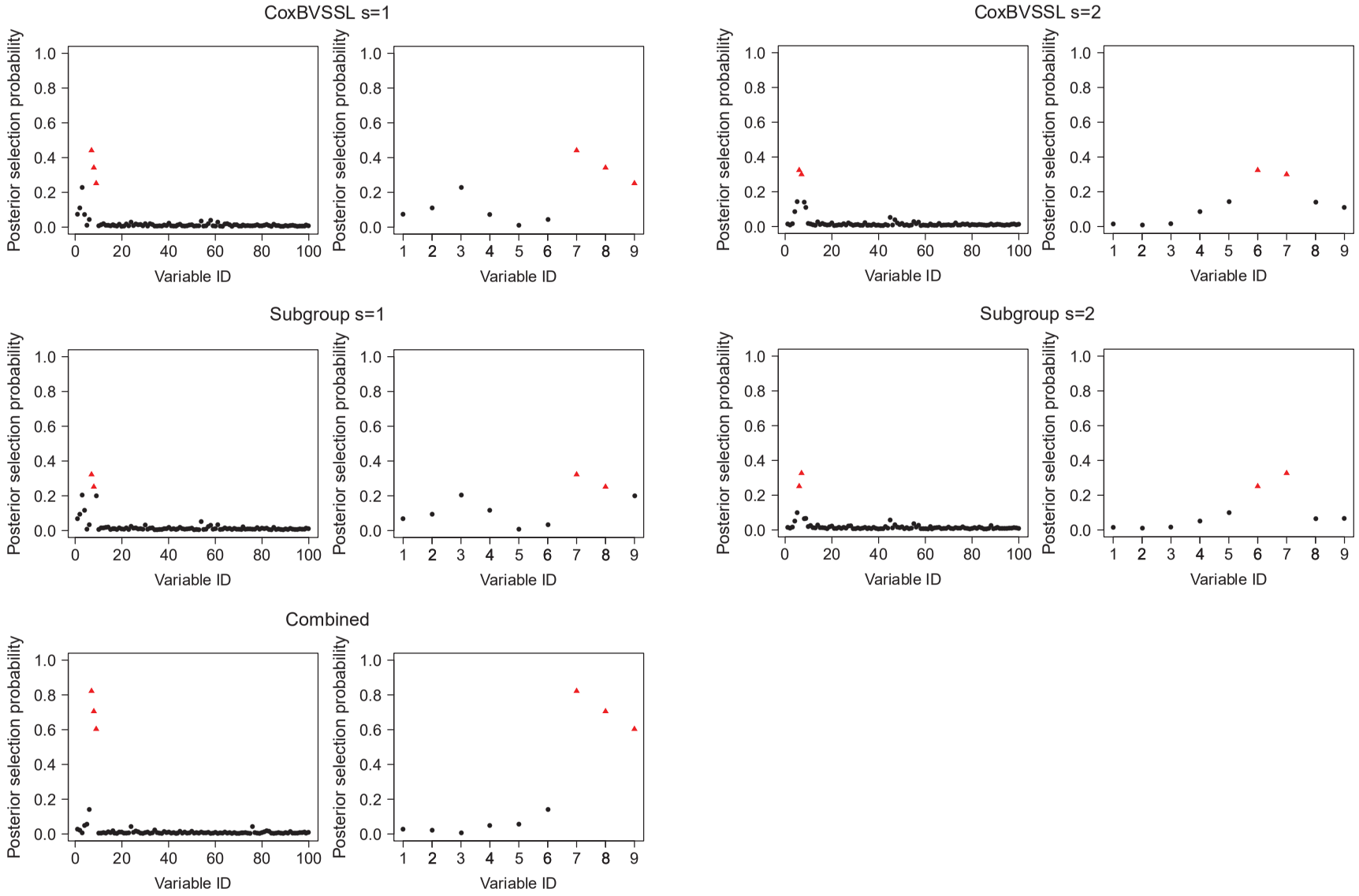


FIGURE B.38: *Mean posterior inclusion probabilities (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(E) $n = 75$ and $p = 100$

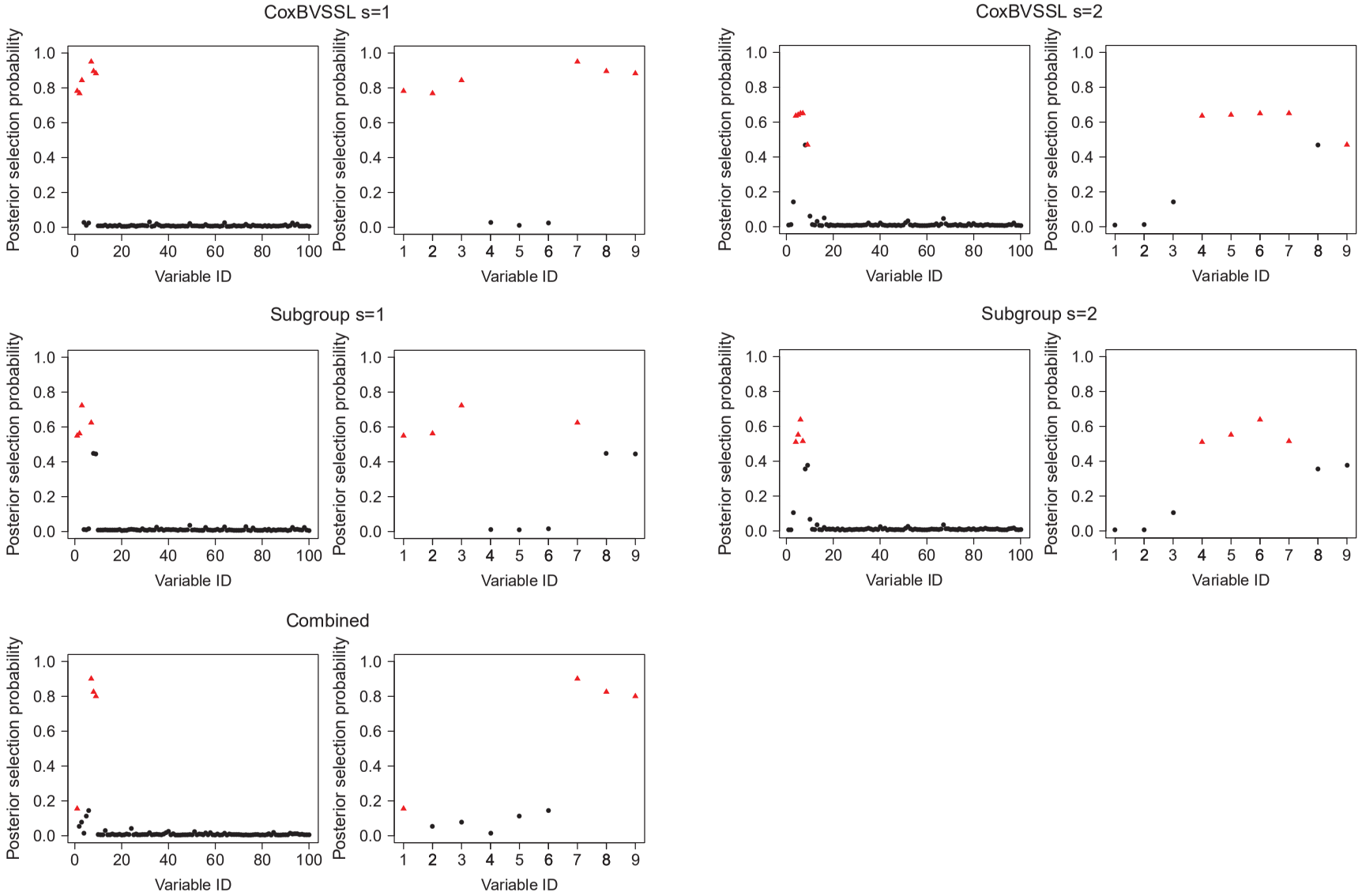


FIGURE B.38: *Mean posterior inclusion probabilities (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(F) $n = 100$ and $p = 100$

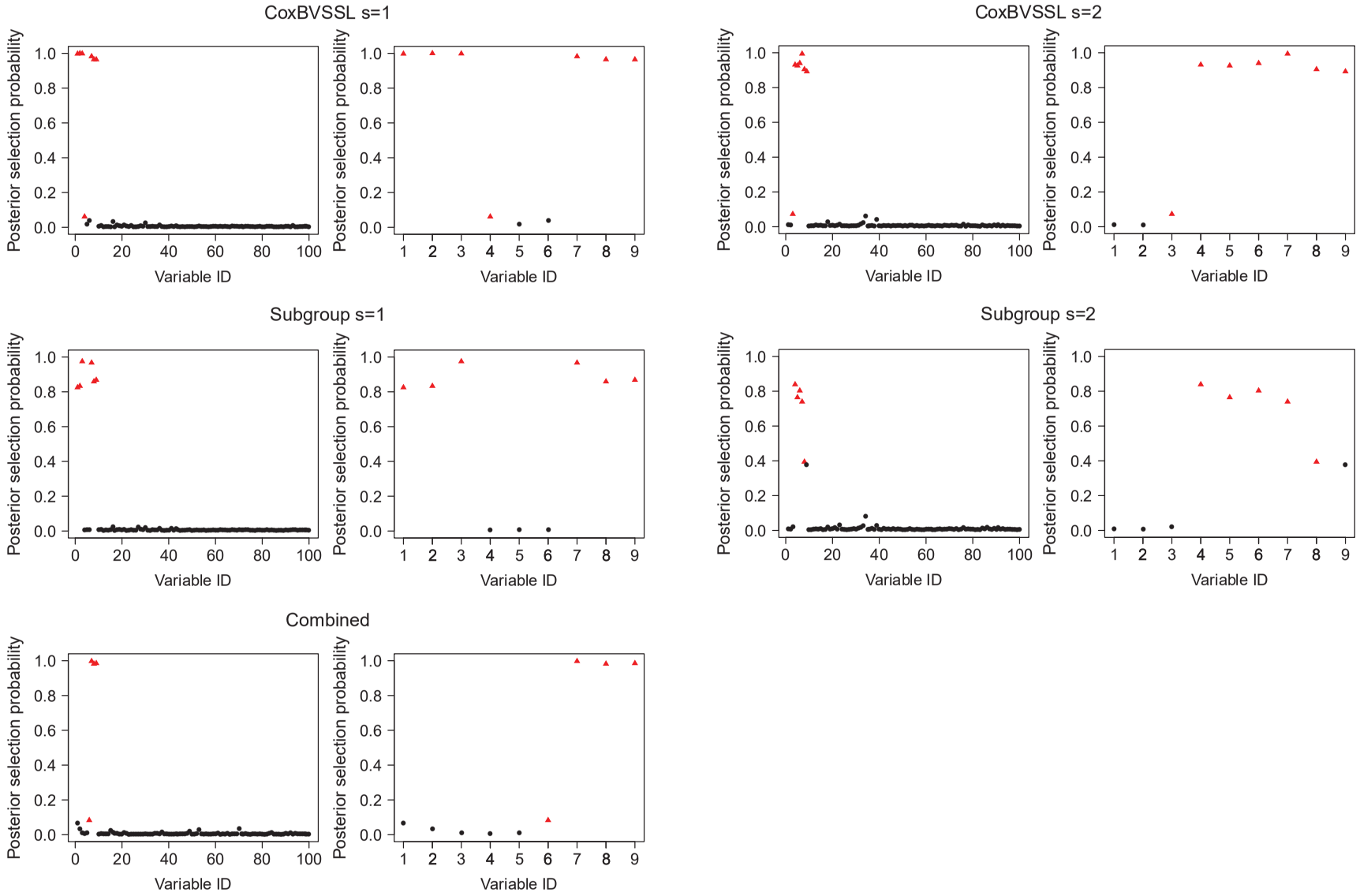


FIGURE B.38: *Mean posterior inclusion probabilities (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(G) $n = 150$ and $p = 100$

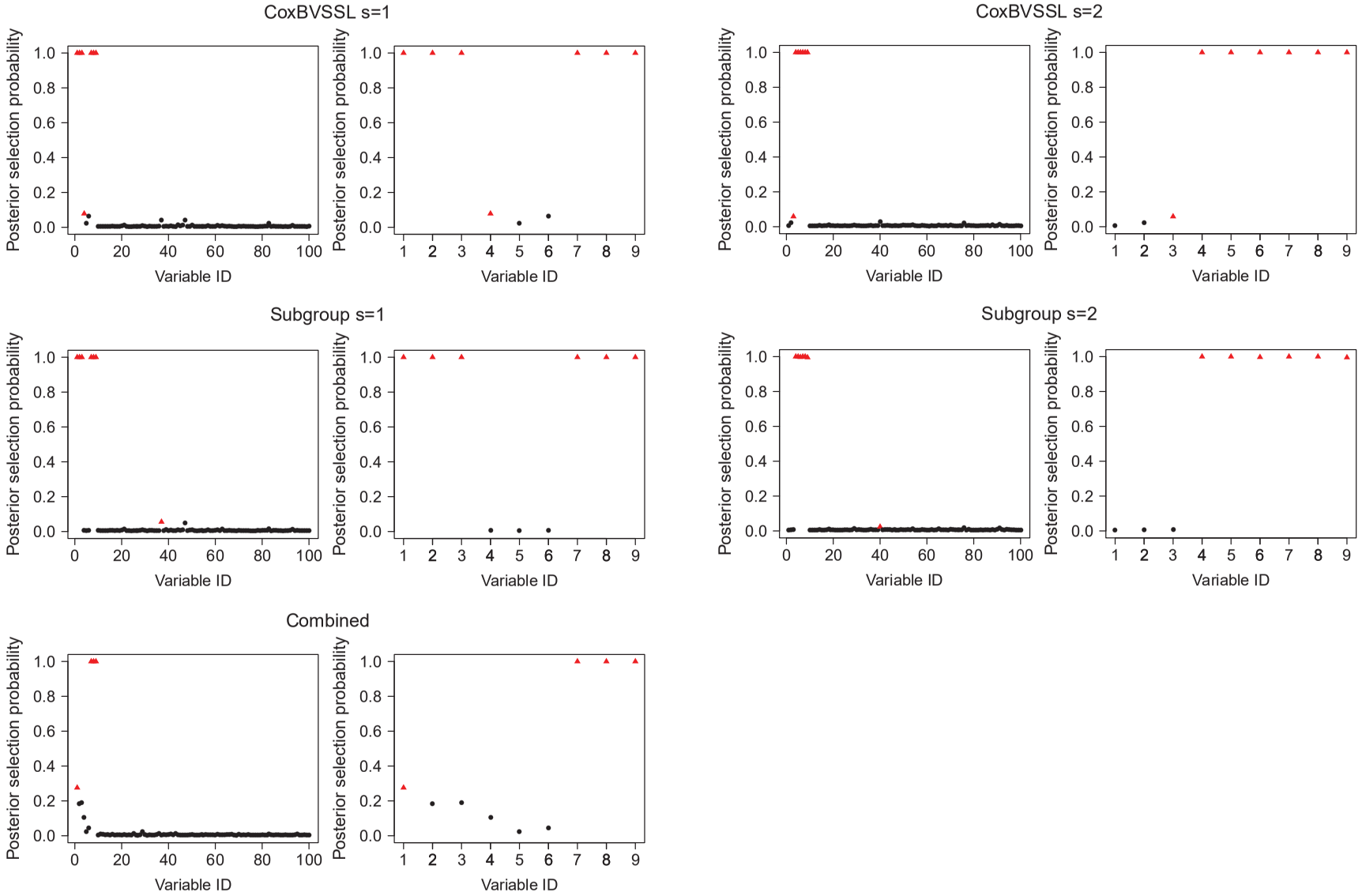


FIGURE B.38: *Mean posterior inclusion probabilities (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(A) $n = 50$ and $p = 20$

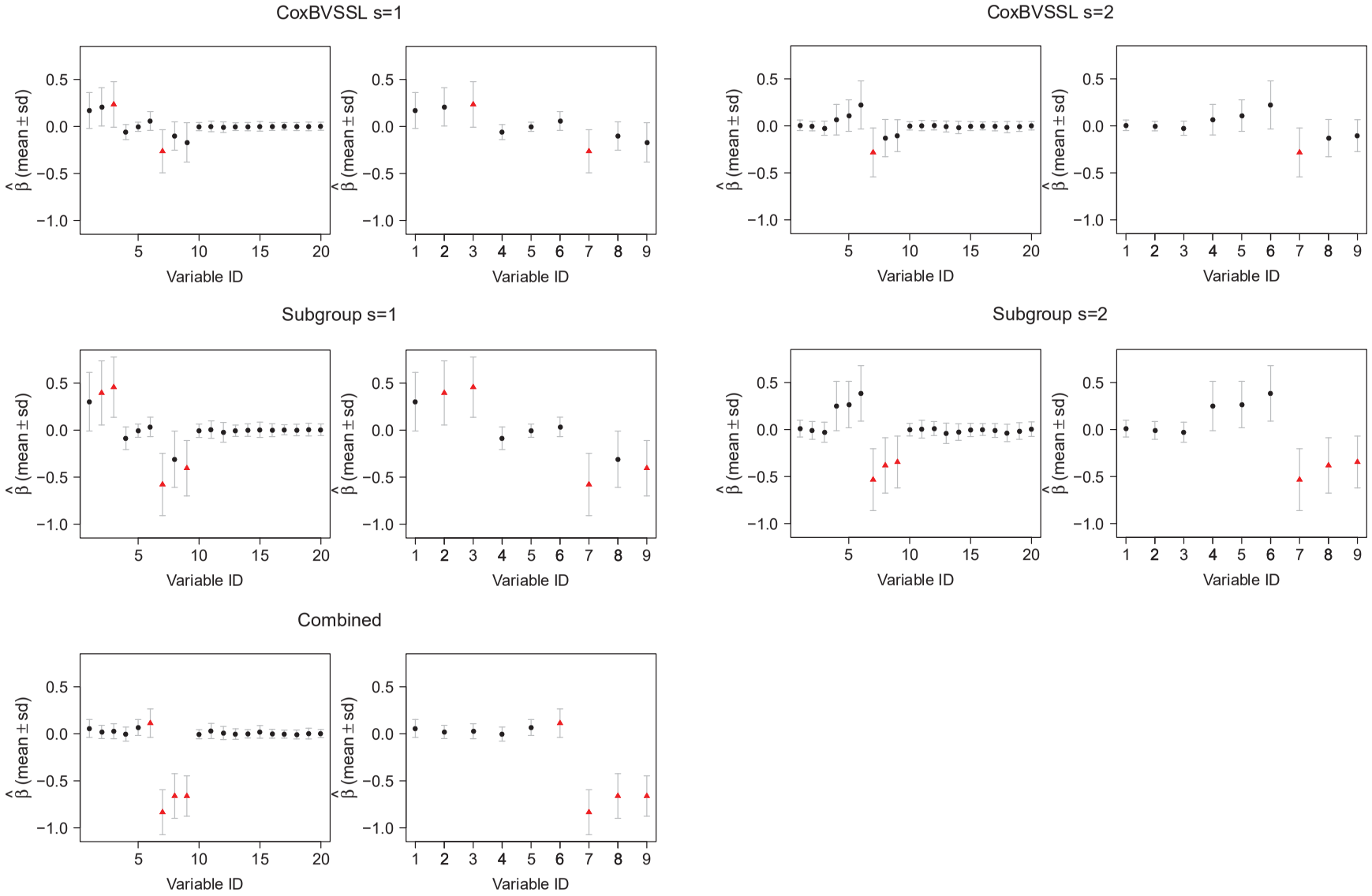


FIGURE B.39: *Posterior mean and standard deviation (sd) of regression coefficients (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles.*

(B) $n = 100$ and $p = 20$

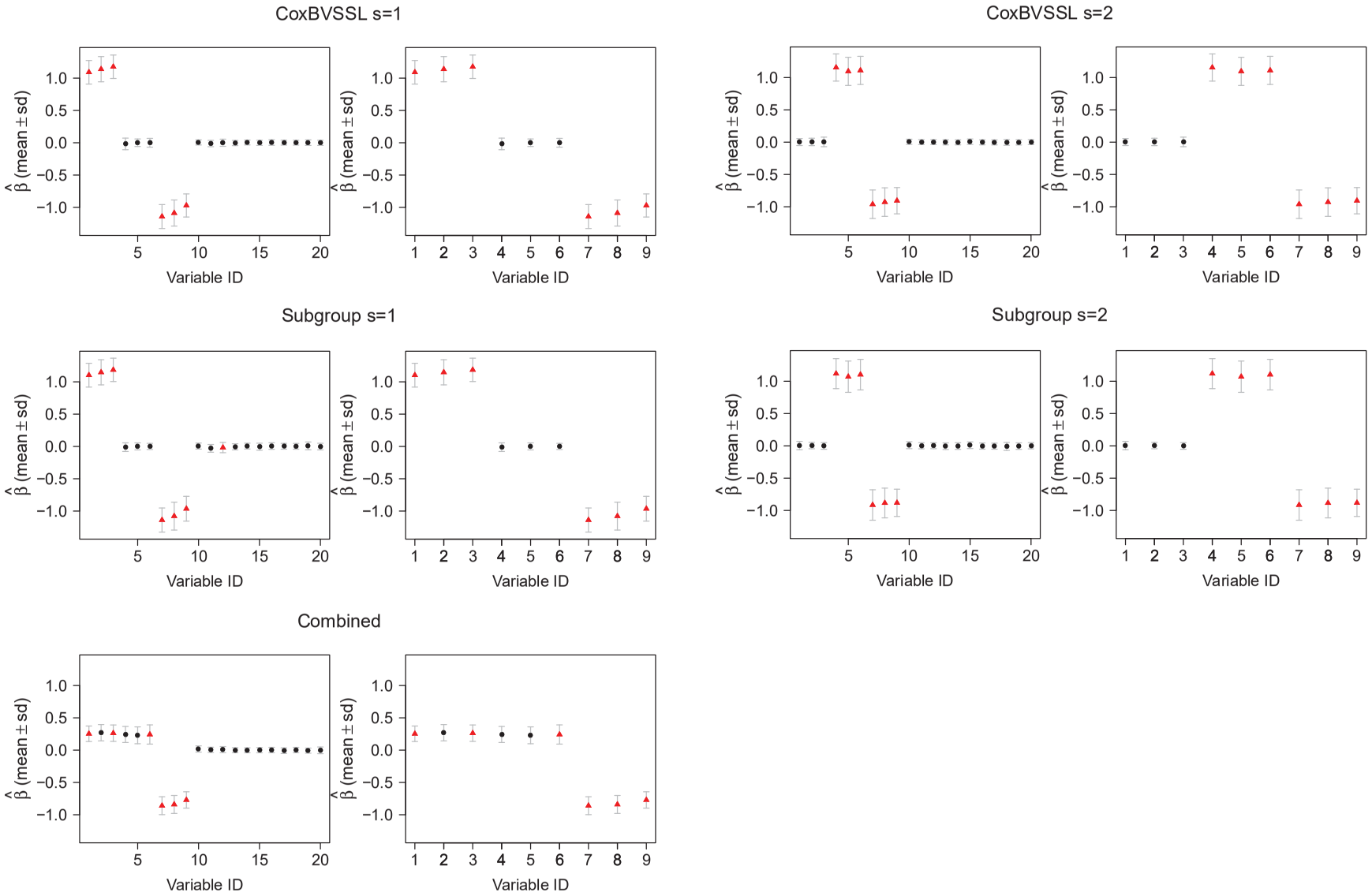


FIGURE B.39: *Posterior mean and standard deviation (sd) of regression coefficients (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(C) $n = 25$ and $p = 100$

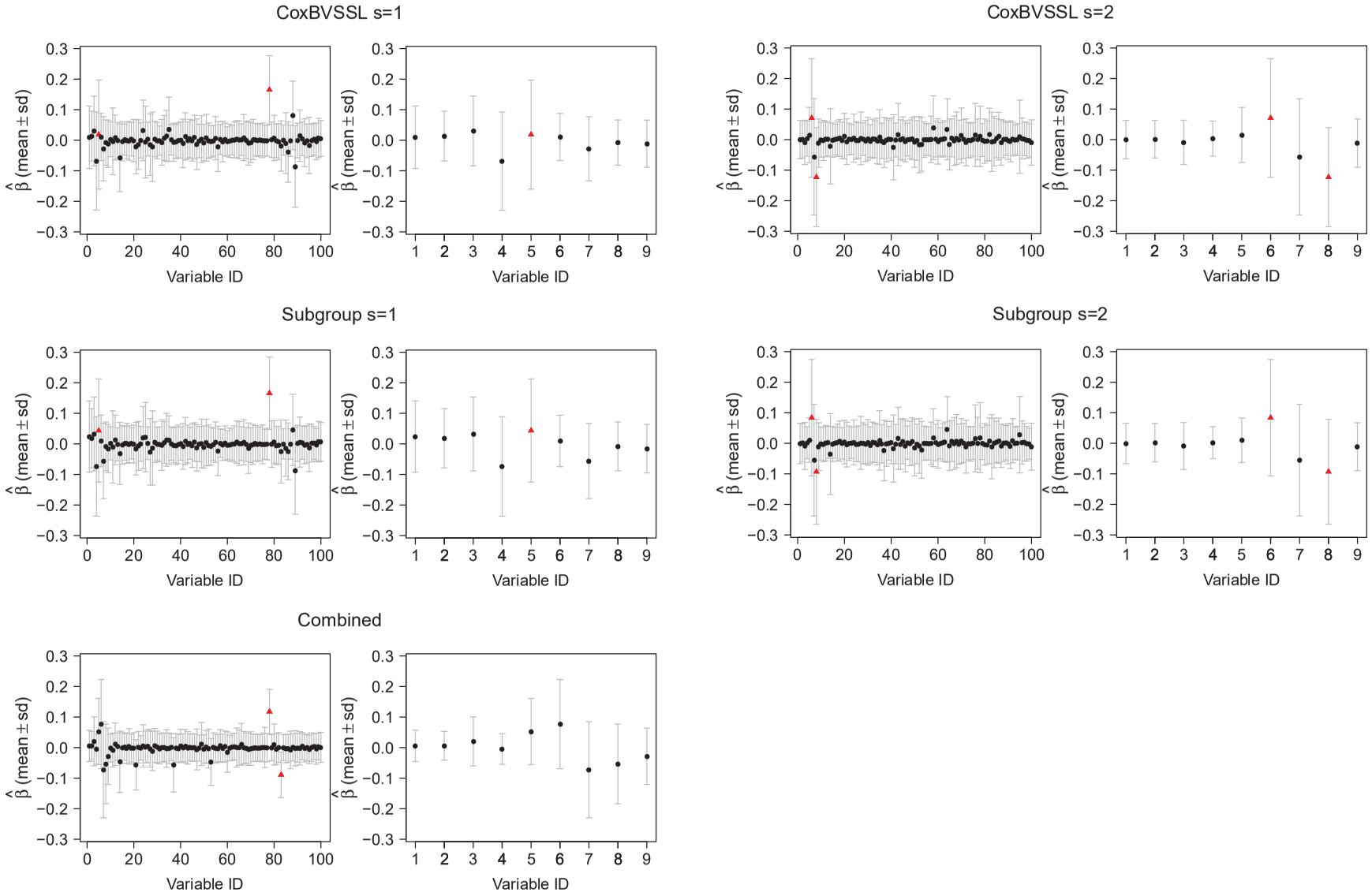


FIGURE B.39: *Posterior mean and standard deviation (sd) of regression coefficients (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(D) $n = 50$ and $p = 100$

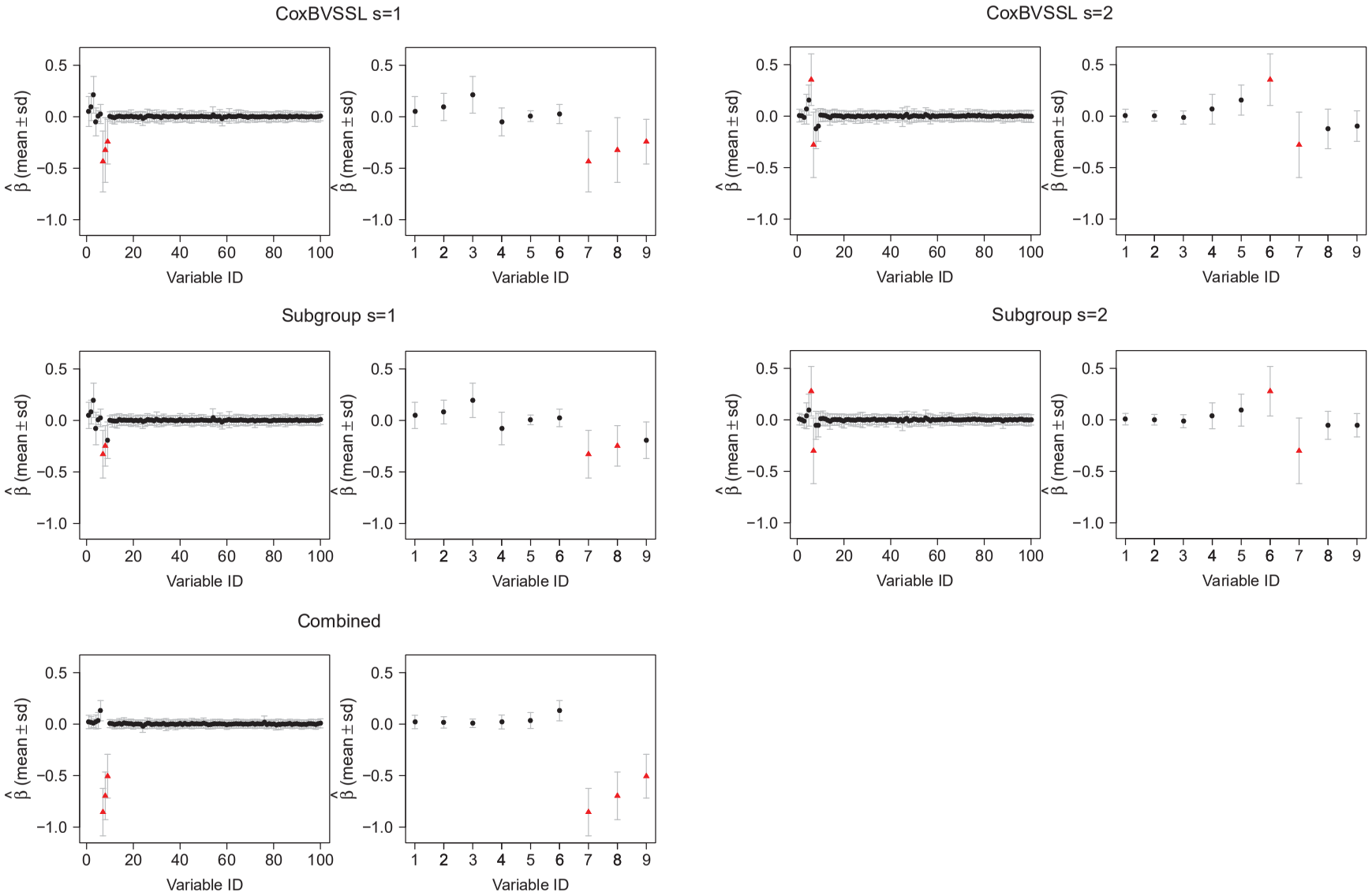


FIGURE B.39: *Posterior mean and standard deviation (sd) of regression coefficients (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(E) $n = 75$ and $p = 100$

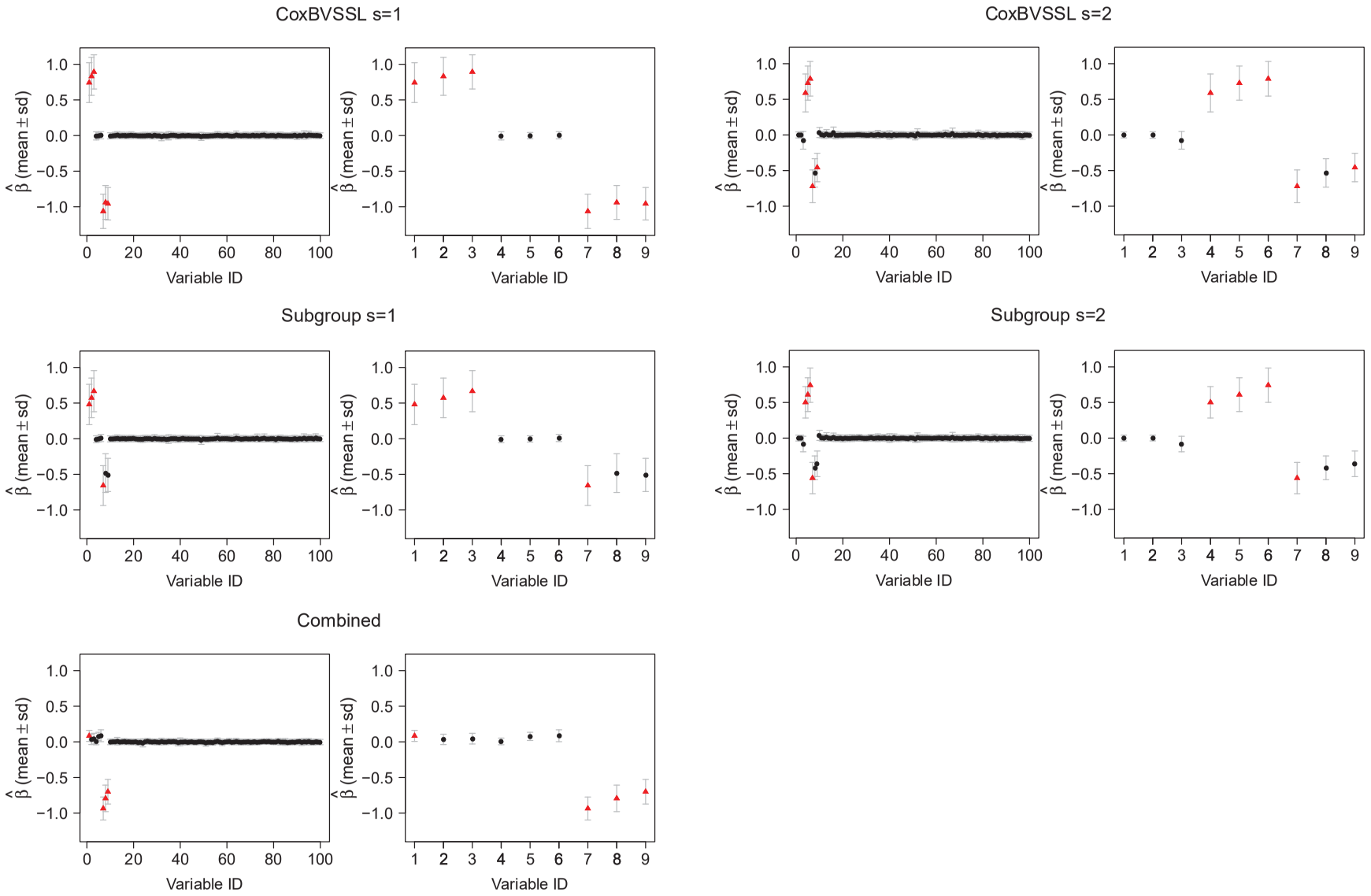


FIGURE B.39: *Posterior mean and standard deviation (sd) of regression coefficients (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(F) $n = 100$ and $p = 100$

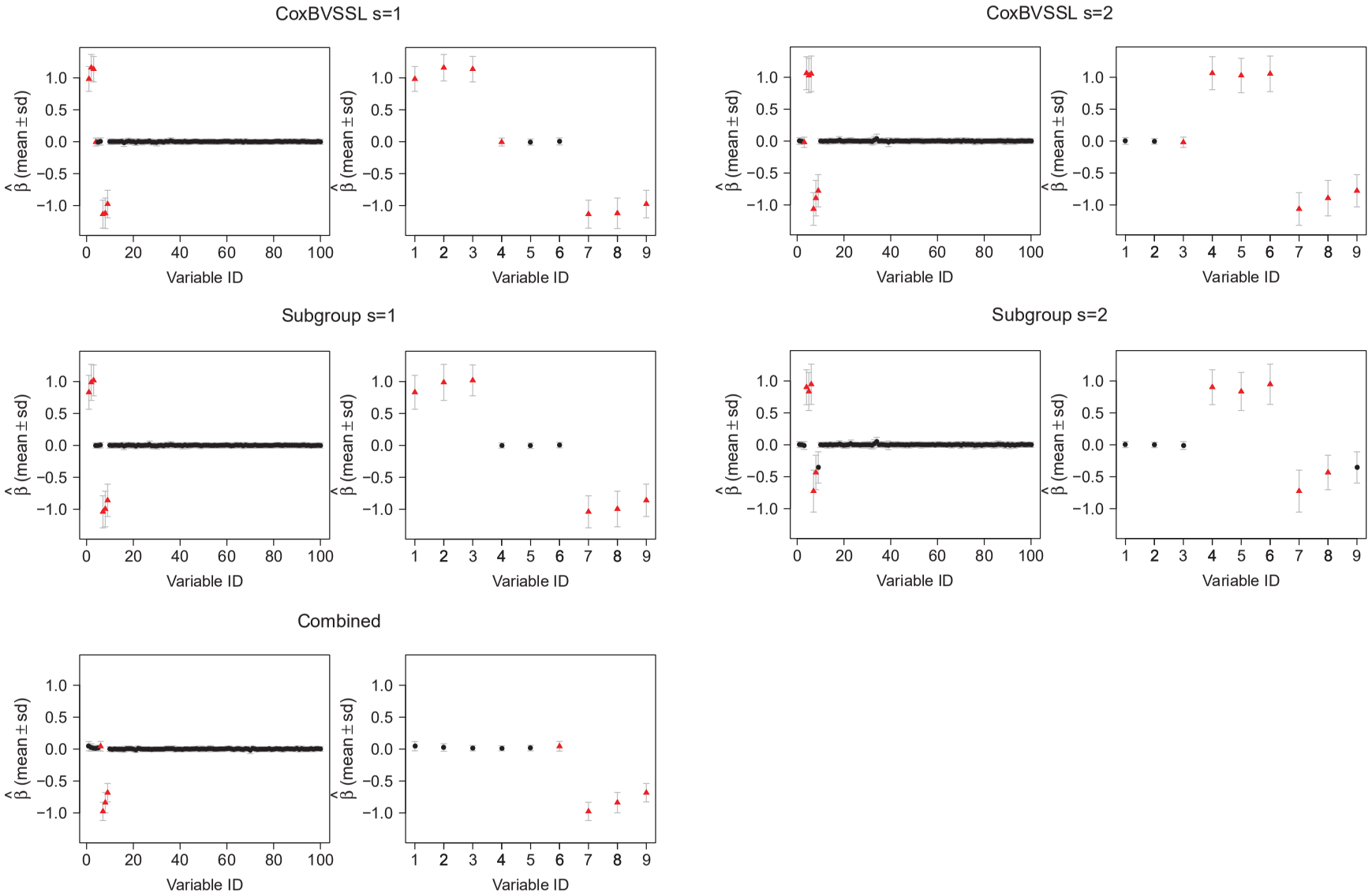


FIGURE B.39: *Posterior mean and standard deviation (sd) of regression coefficients (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

(G) $n = 150$ and $p = 100$

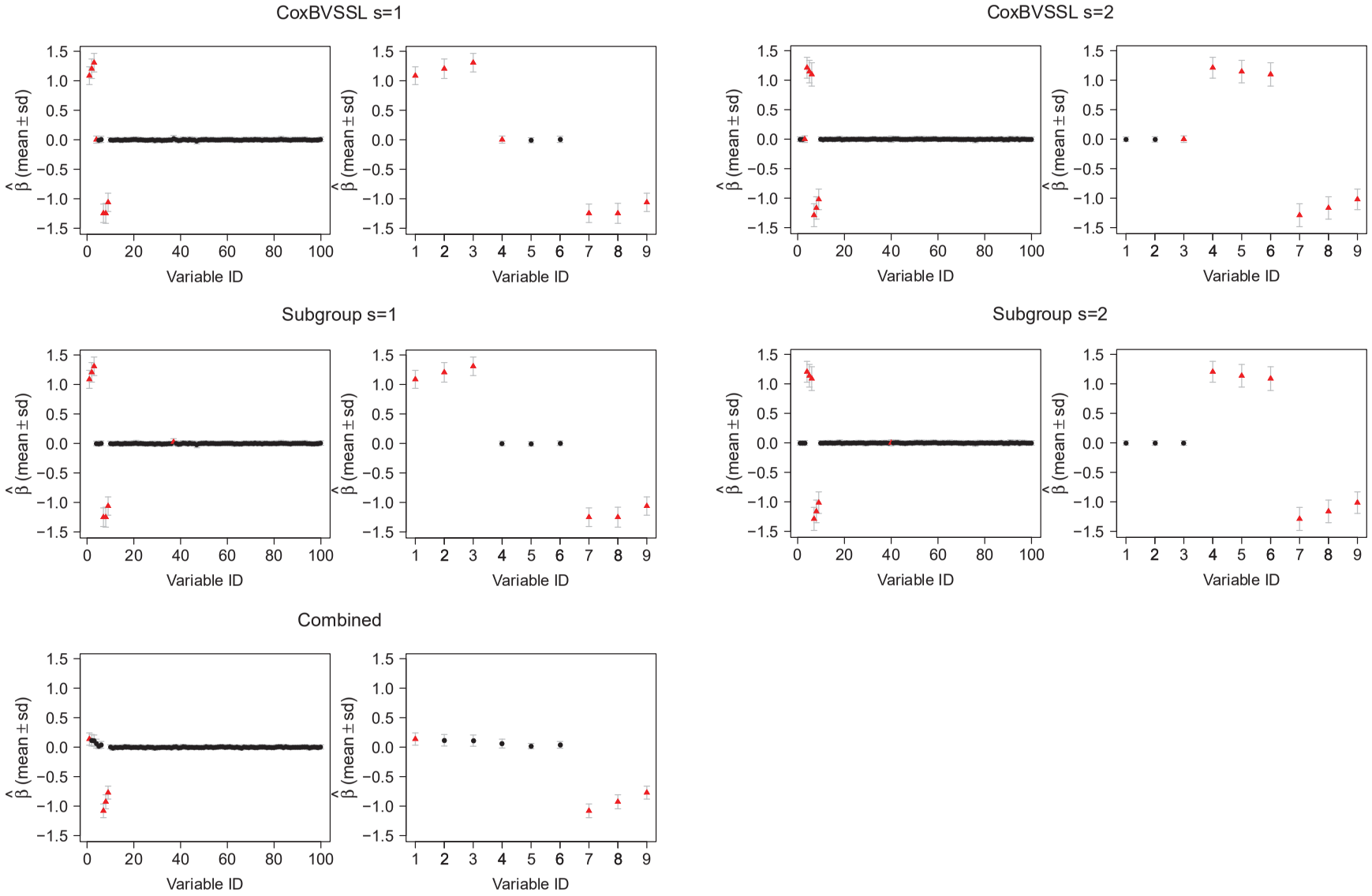


FIGURE B.39: *Posterior mean and standard deviation (sd) of regression coefficients (average across all simulations) of all p variables (left) and close-up of prognostic variables (right) for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. Selected variables are highlighted as red triangles (cont.).*

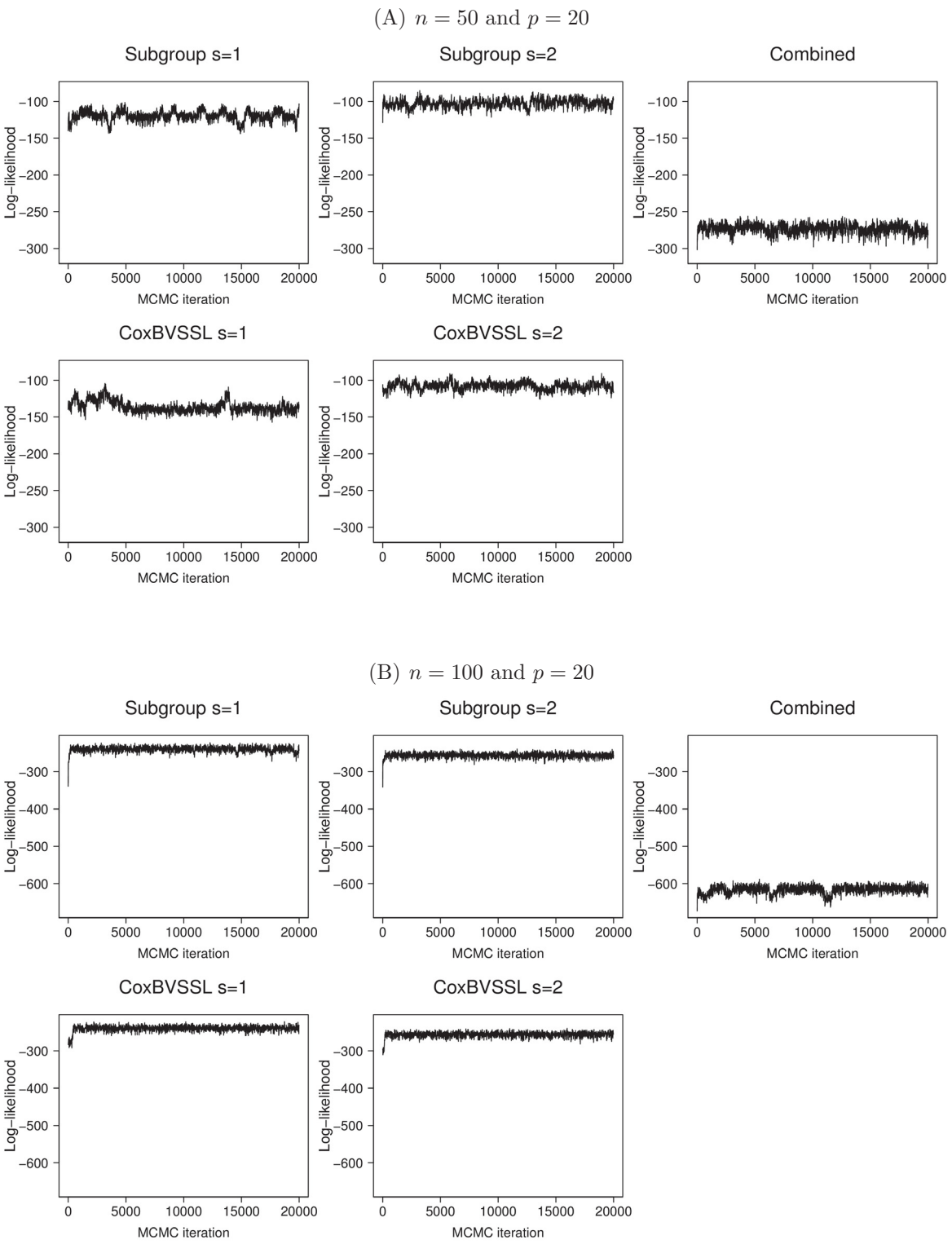
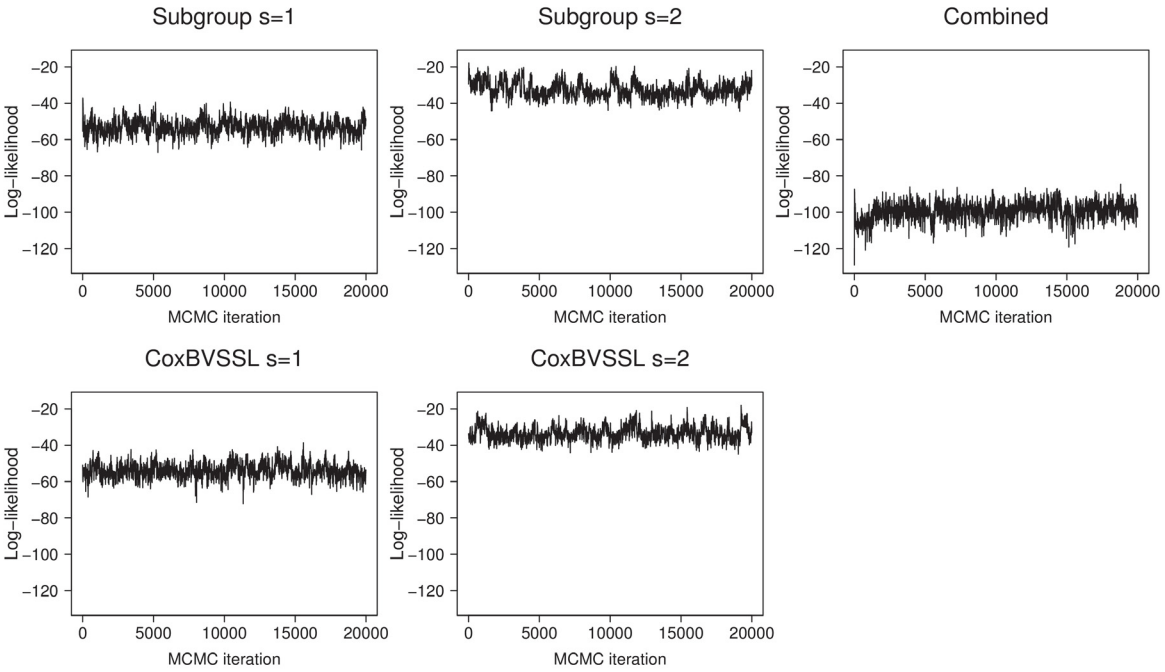


FIGURE B.40: *Trace plots of the log-likelihood from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p.*

(C) $n = 25$ and $p = 100$



(D) $n = 50$ and $p = 100$

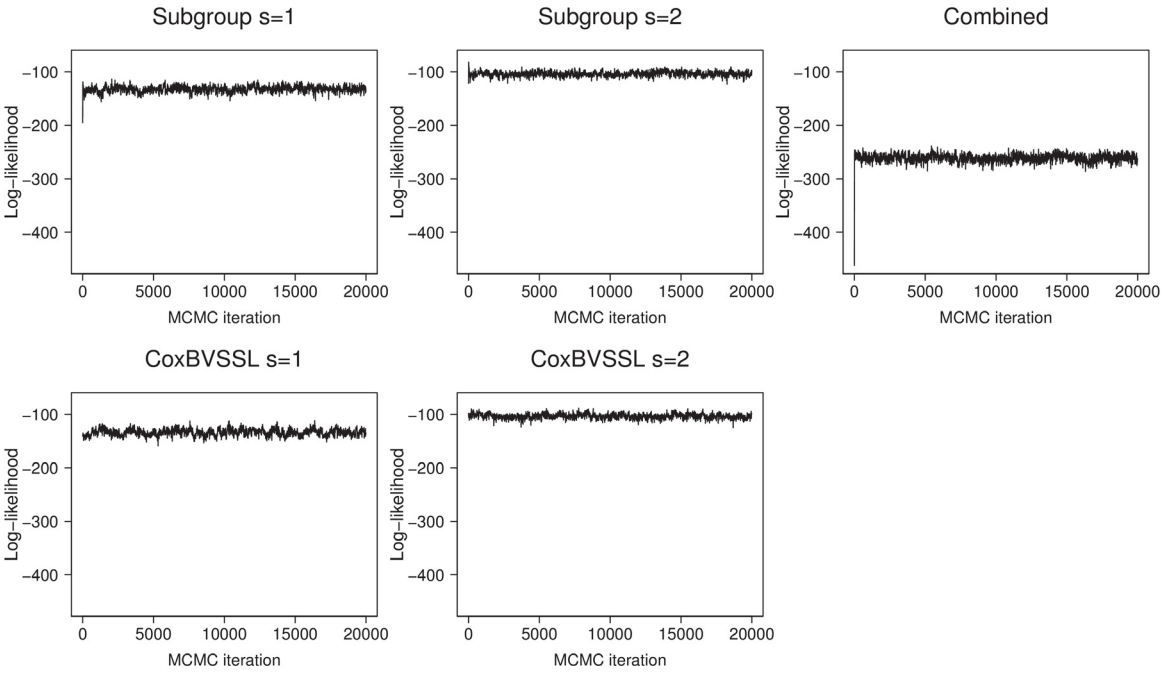
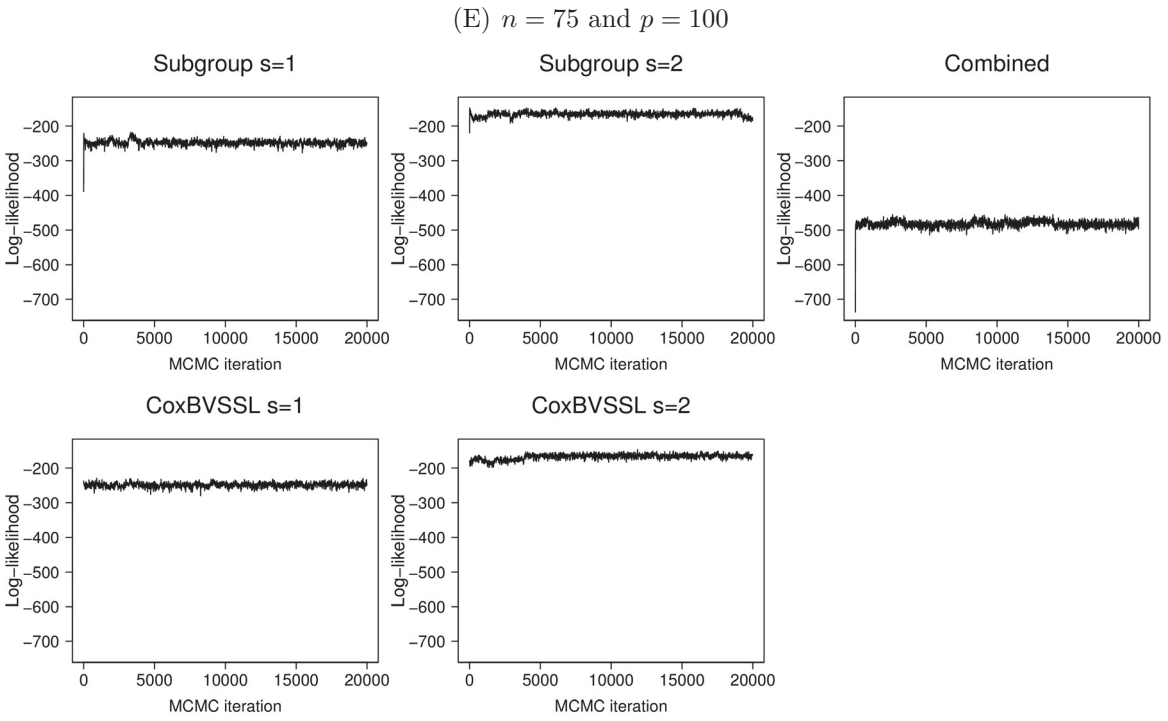


FIGURE B.40: *Trace plots of the log-likelihood from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p (cont.).*



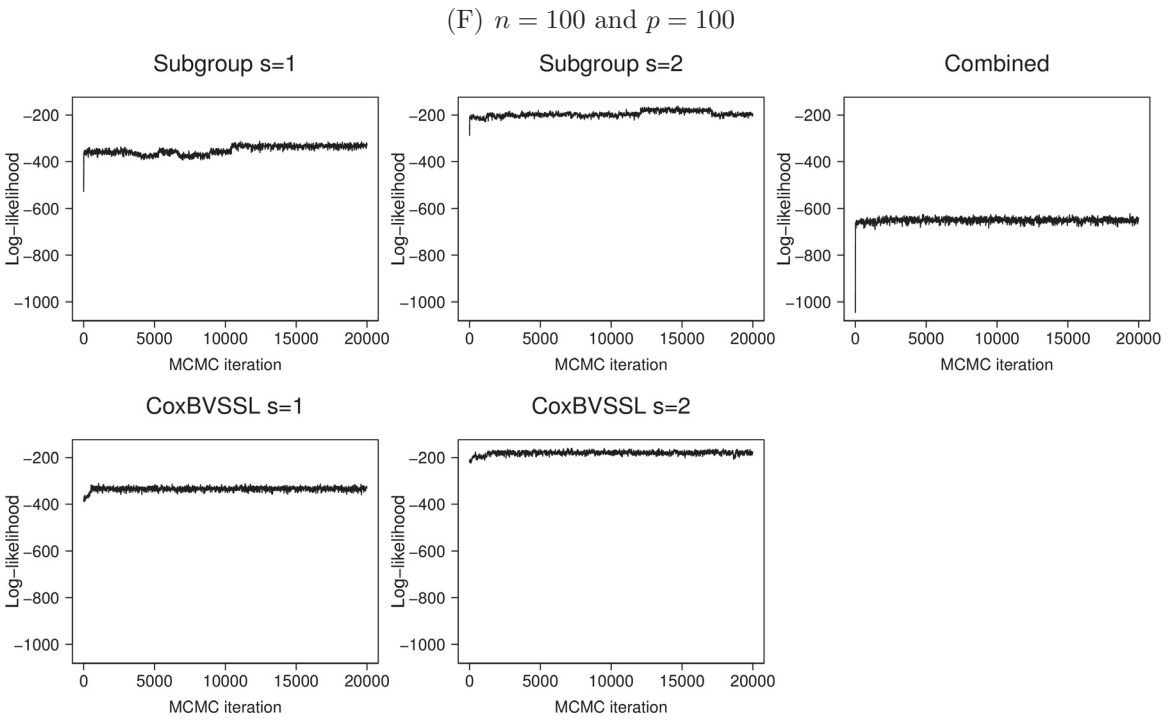


FIGURE B.40: *Trace plots of the log-likelihood from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p (cont.).*

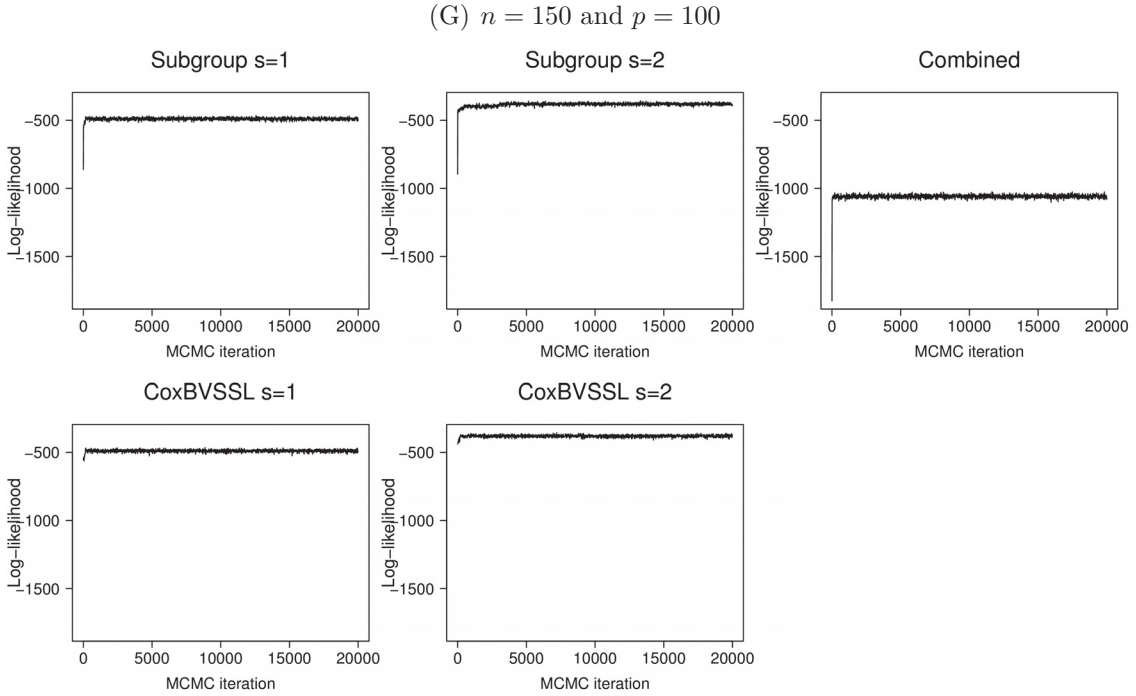


FIGURE B.40: *Trace plots of the log-likelihood from the first simulation for three different Cox models (subgroup, combined, and CoxBVSSL), subgroups $s = 1, 2$ and varying n and p (cont.).*

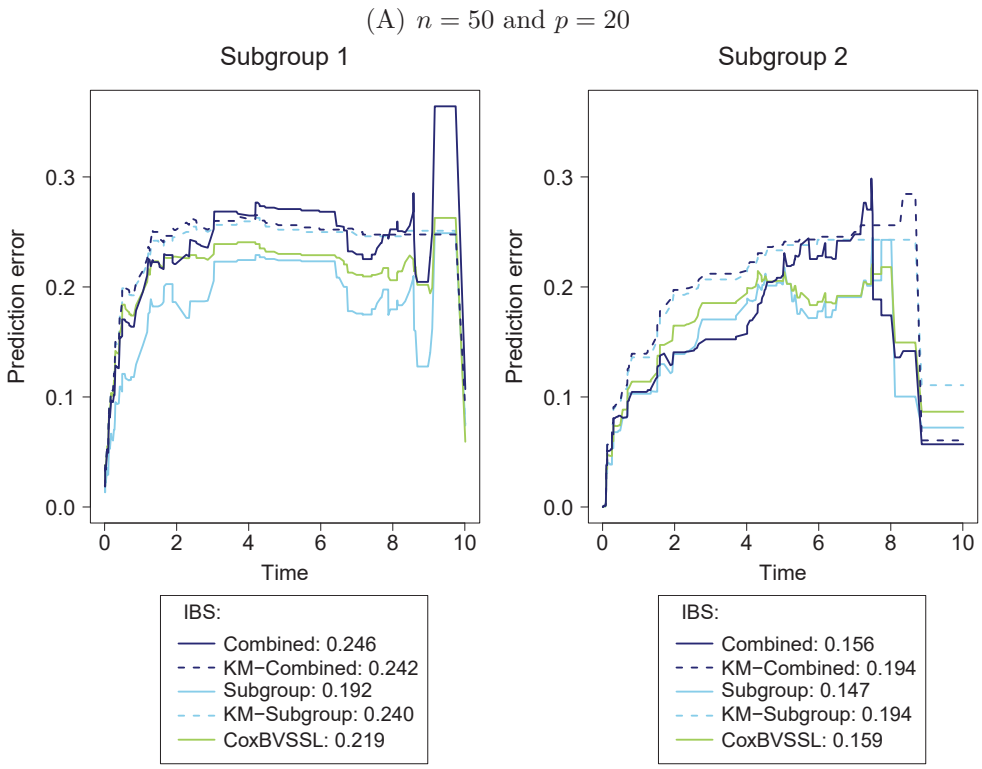


FIGURE B.41: *Prediction error curves and integrated Brier scores (IBS) from the first simulation for varying n and p. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (KM-Subgroup) or combined (KM-Combined) training data.*

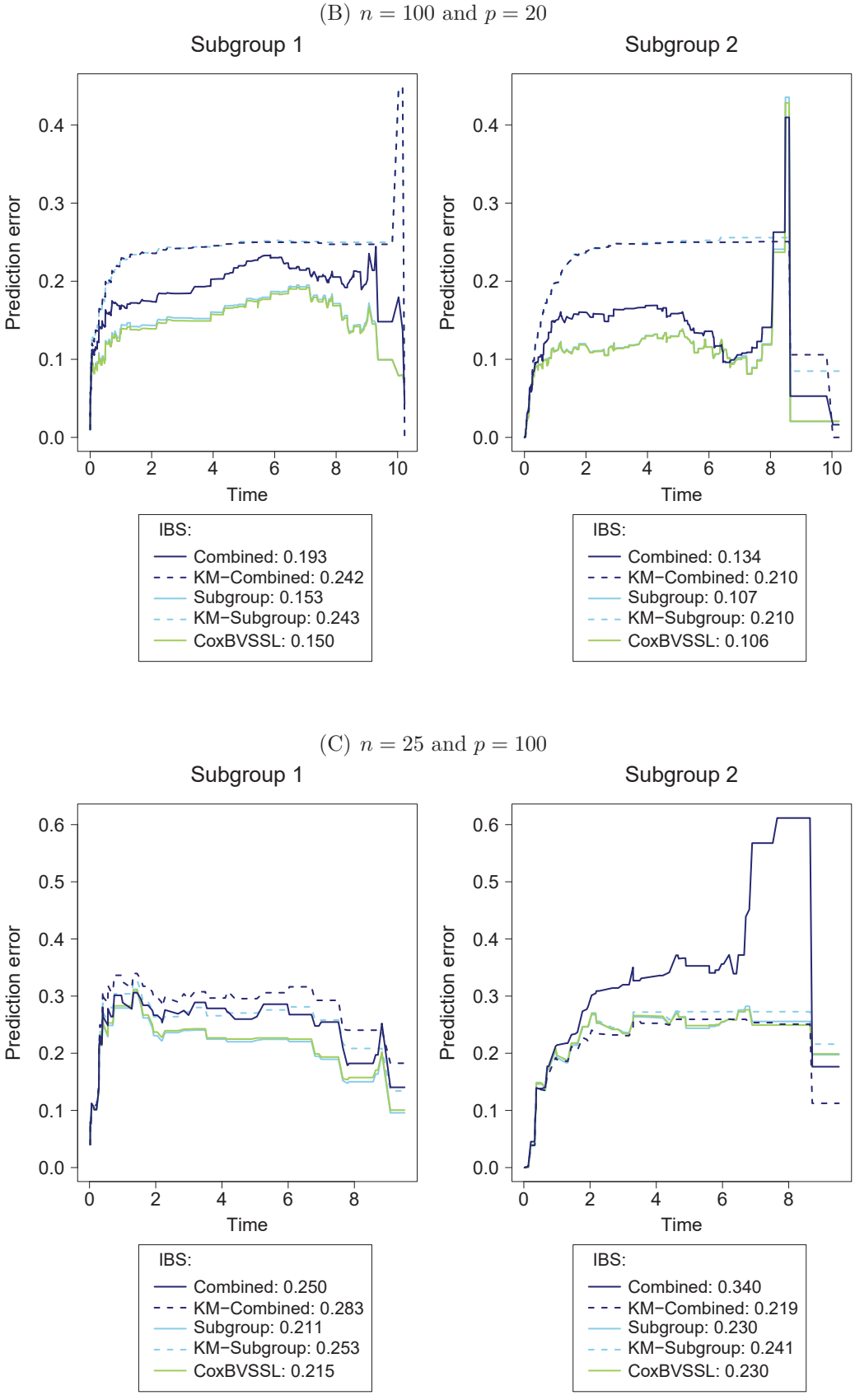


FIGURE B.41: *Prediction error curves and integrated Brier scores (IBS) from the first simulation for varying n and p. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (KM-Subgroup) or combined (KM-Combined) training data (cont.).*

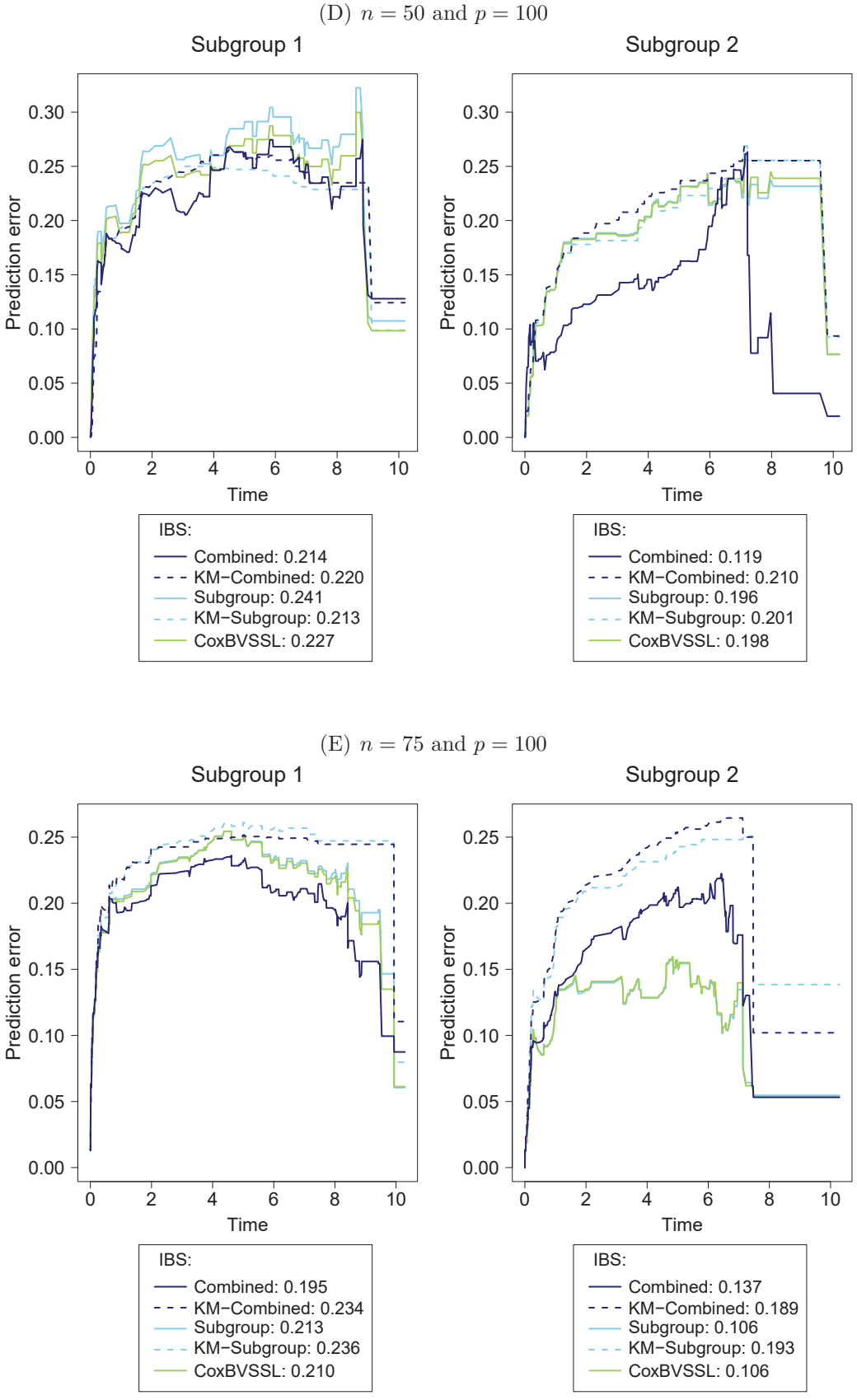
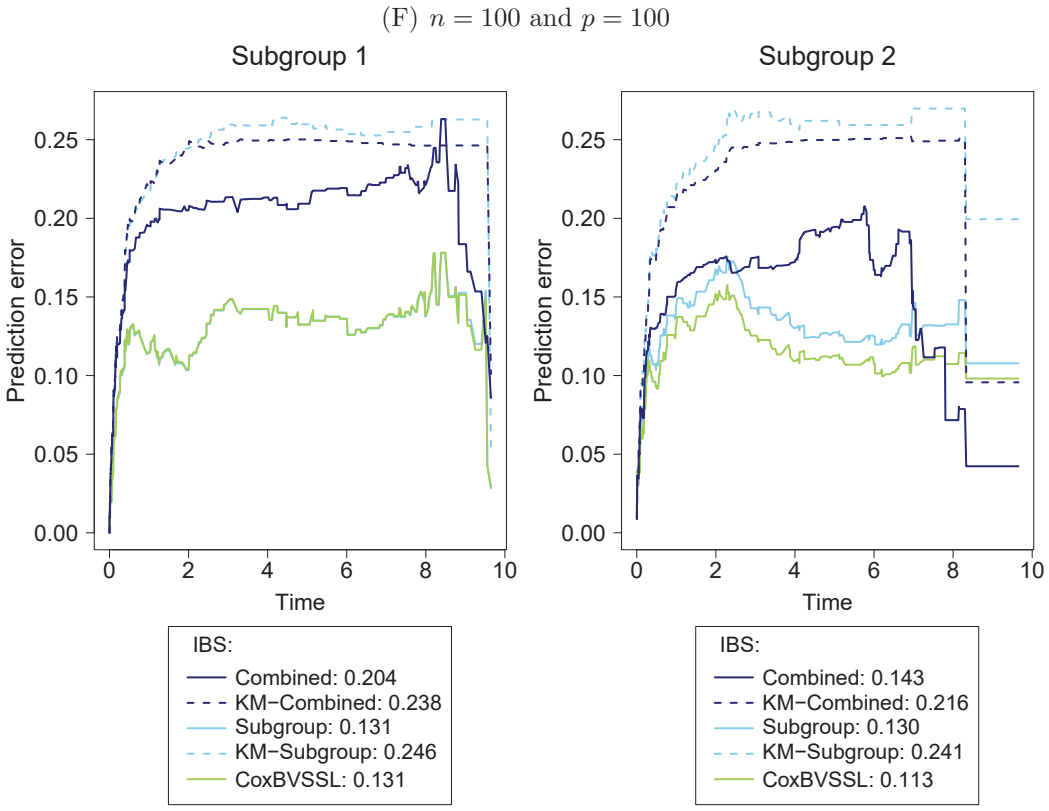


FIGURE B.41: *Prediction error curves and integrated Brier scores (IBS) from the first simulation for varying n and p. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (KM-Subgroup) or combined (KM-Combined) training data (cont.).*



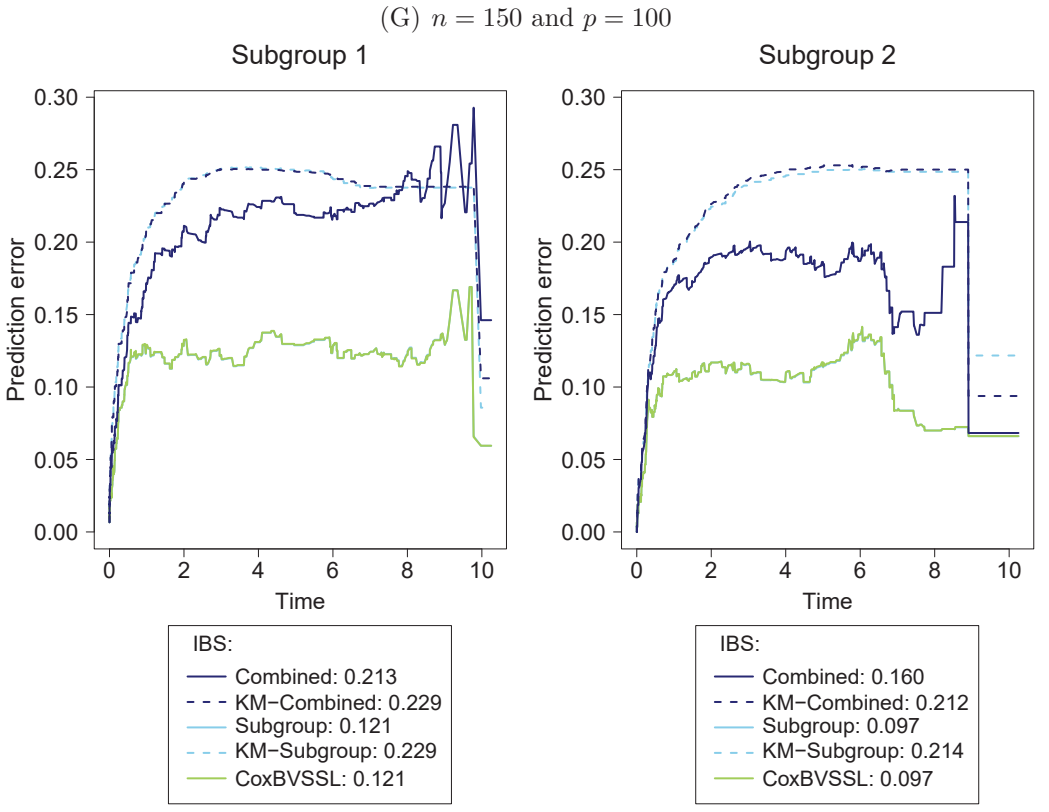
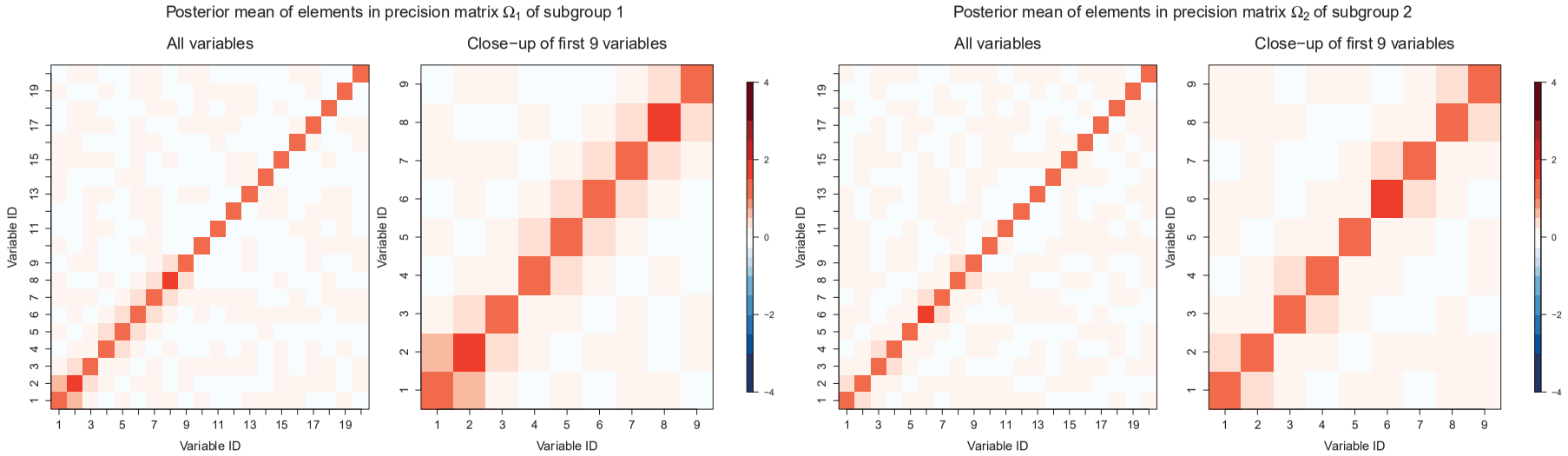


FIGURE B.41: *Prediction error curves and integrated Brier scores (IBS) from the first simulation for varying n and p. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (KM-Subgroup) or combined (KM-Combined) training data (cont.).*

(A) $n = 50$ and $p = 20$



(B) $n = 100$ and $p = 20$

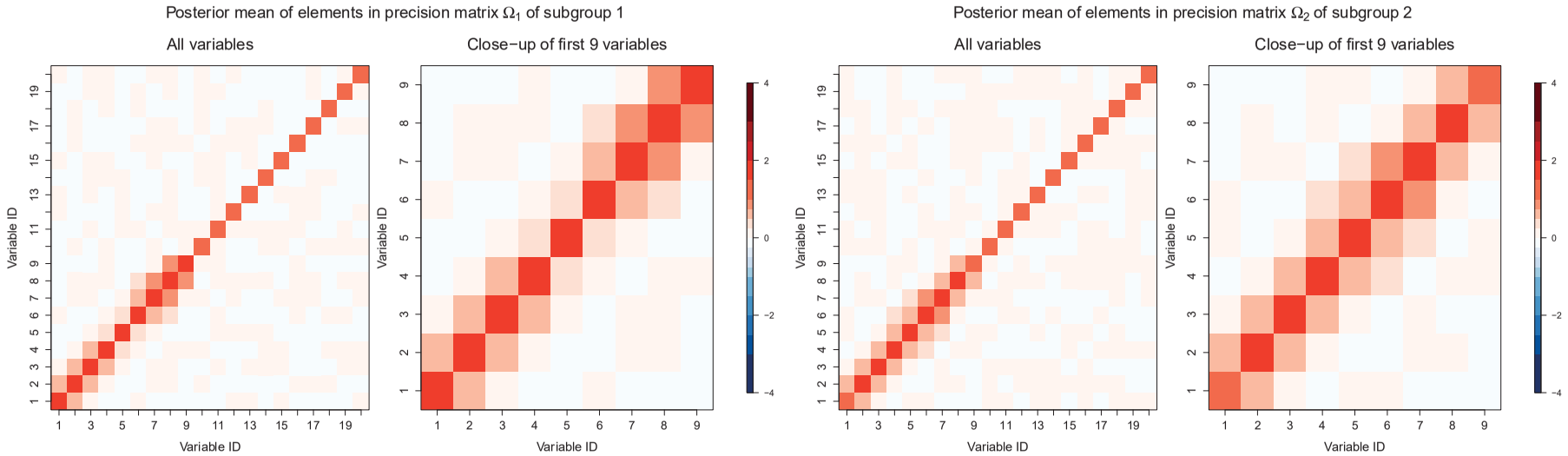


FIGURE B.42: *Posterior mean (average across all simulations) of elements in precision matrix for subgroup 1 (left) and 2 (right). The prior of the diagonal entries of the precision matrix is exponential with parameter* $\frac{1}{2}$*, and the prior of the off-diagonal entries is a mixture of two normal distributions with zero mean and variance* $\nu_0^2 = 0.1^2$ *for non-selected edges and variance* $\nu_1^2 = 5^2$ *for selected edges.*

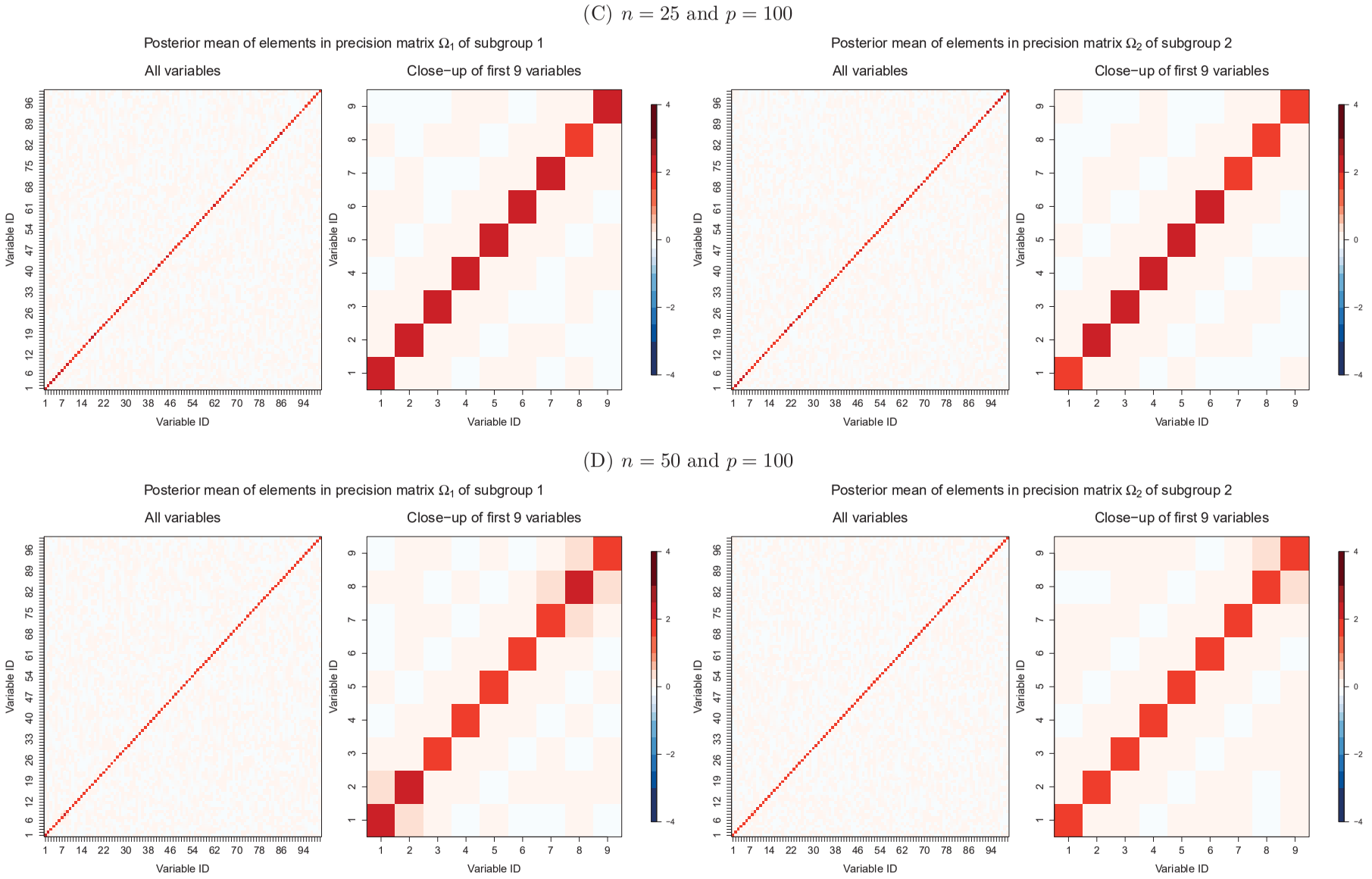
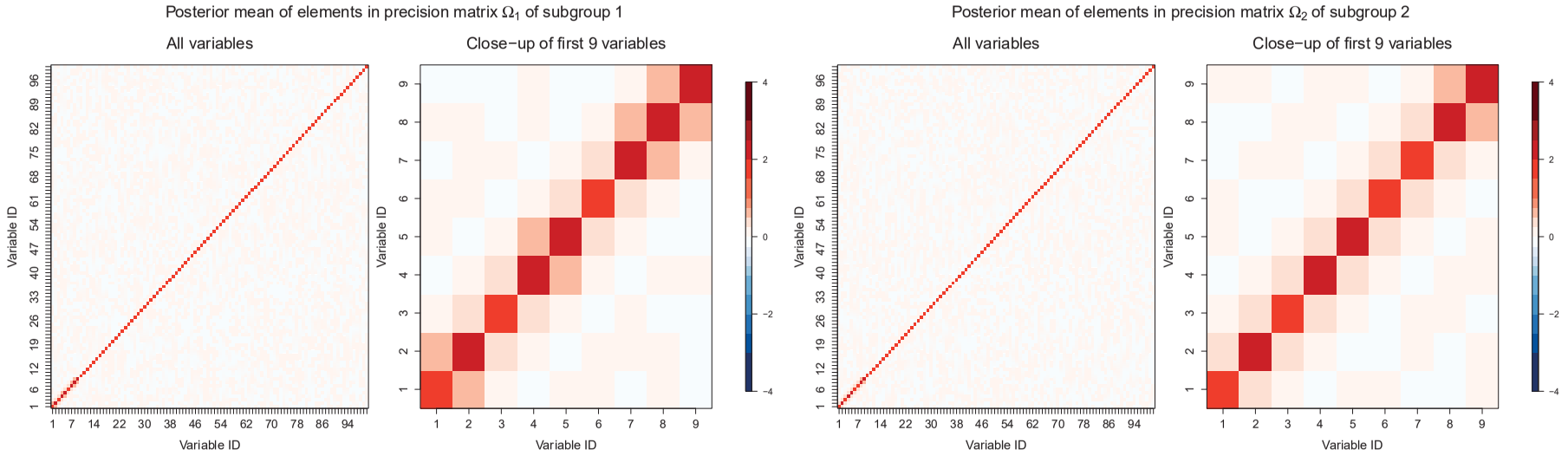


FIGURE B.42: *Posterior mean (average across all simulations) of elements in precision matrix for subgroup 1 (left) and 2 (right). The prior of the diagonal entries of the precision matrix is exponential with parameter* $\frac{1}{2}$*, and the prior of the off-diagonal entries is a mixture of two normal distributions with zero mean and variance* $\nu_0^2 = 0.1^2$ *for non-selected edges and variance* $\nu_1^2 = 5^2$ *for selected edges. (cont.).*

(E) $n = 75$ and $p = 100$



(F) $n = 100$ and $p = 100$

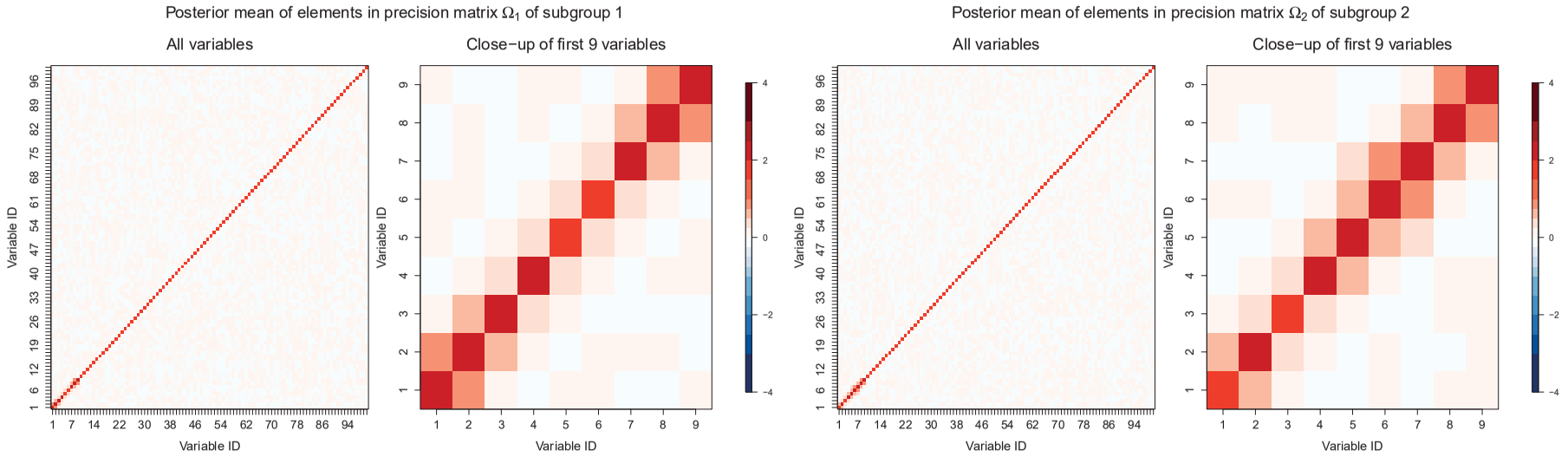


FIGURE B.42: *Posterior mean (average across all simulations) of elements in precision matrix for subgroup 1 (left) and 2 (right). The prior of the diagonal entries of the precision matrix is exponential with parameter* $\frac{1}{2}$*, and the prior of the off-diagonal entries is a mixture of two normal distributions with zero mean and variance* $\nu_0^2 = 0.1^2$ *for non-selected edges and variance* $\nu_1^2 = 5^2$ *for selected edges. (cont.).*

(G) $n = 150$ and $p = 100$

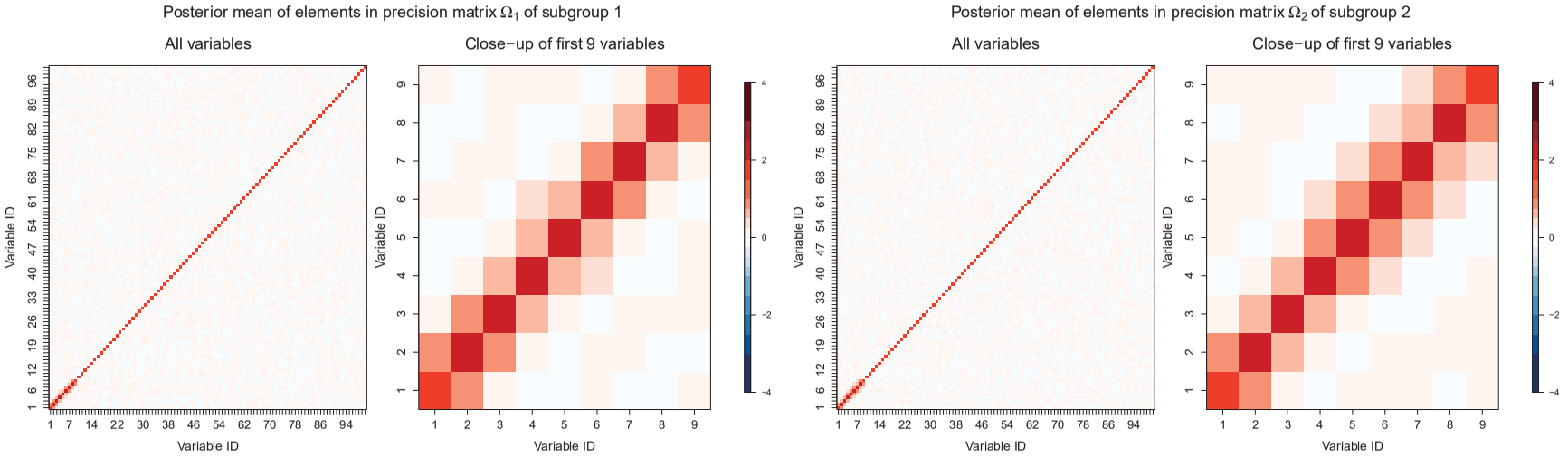


FIGURE B.42: *Posterior mean (average across all simulations) of elements in precision matrix for subgroup 1 (left) and 2 (right). The prior of the diagonal entries of the precision matrix is exponential with parameter* $\frac{1}{2}$*, and the prior of the off-diagonal entries is a mixture of two normal distributions with zero mean and variance* $\nu_0^2 = 0.1^2$ *for non-selected edges and variance* $\nu_1^2 = 5^2$ *for selected edges. (cont.).*

(A) $n = 50$ and $p = 20$

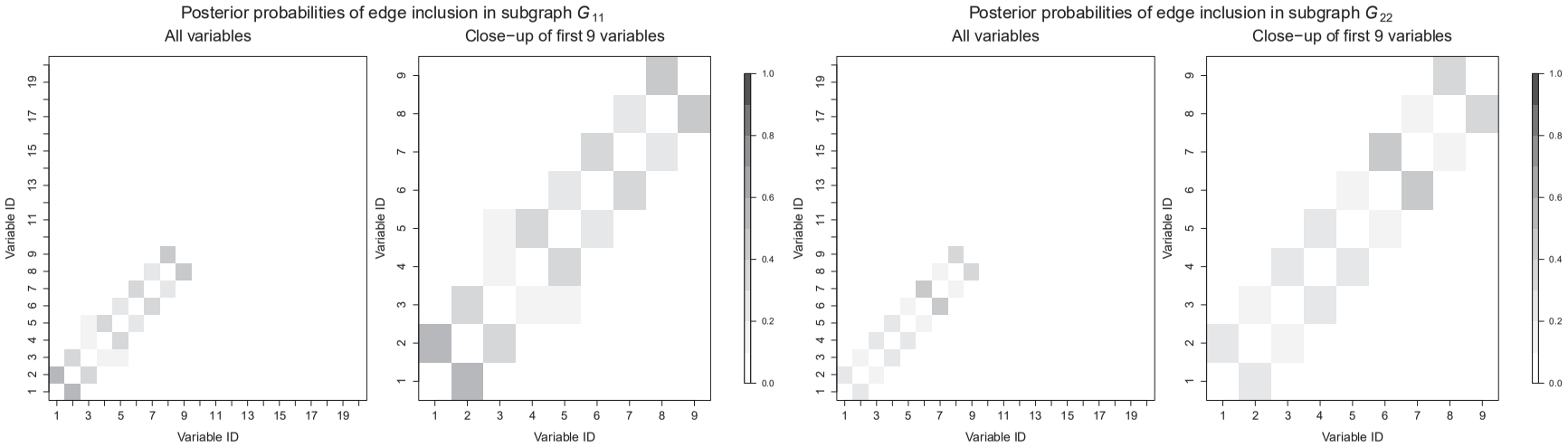
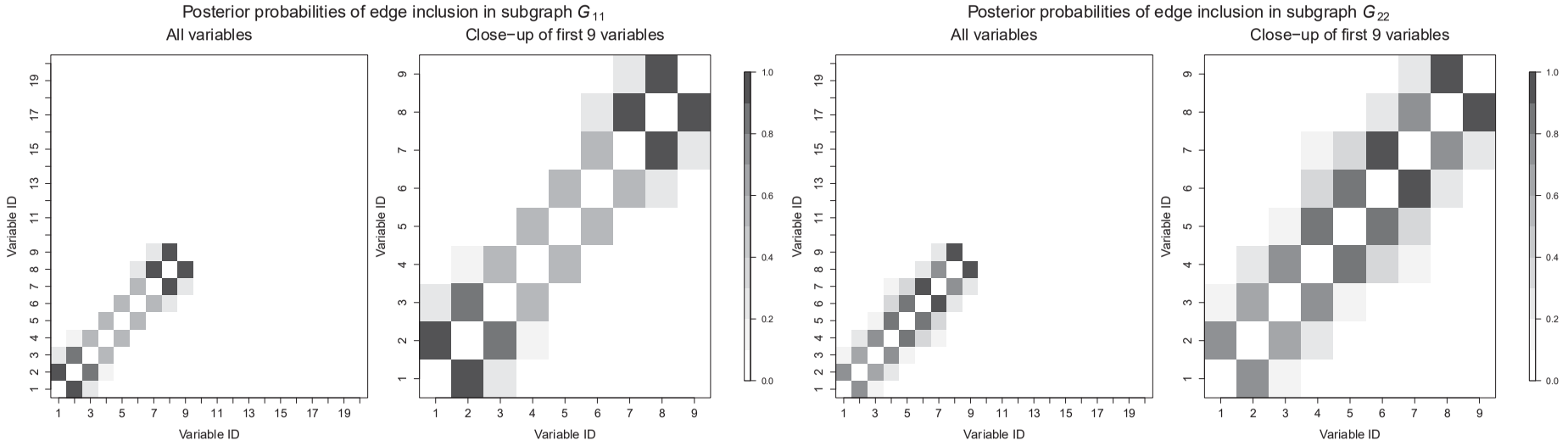


FIGURE B.43: *Mean posterior probabilities of edge inclusion (average across all simulations) in subgraph* $\boldsymbol{G}_{ss}$ *for subgroup* $s = 1$ *(left) and* $s = 2$ *(right). Edges in the graph are assumed independent Bernoulli a priori with parameter* $\pi = 2/(p-1)$*.*

(B) $n = 100$ and $p = 20$



(C) $n = 25$ and $p = 100$

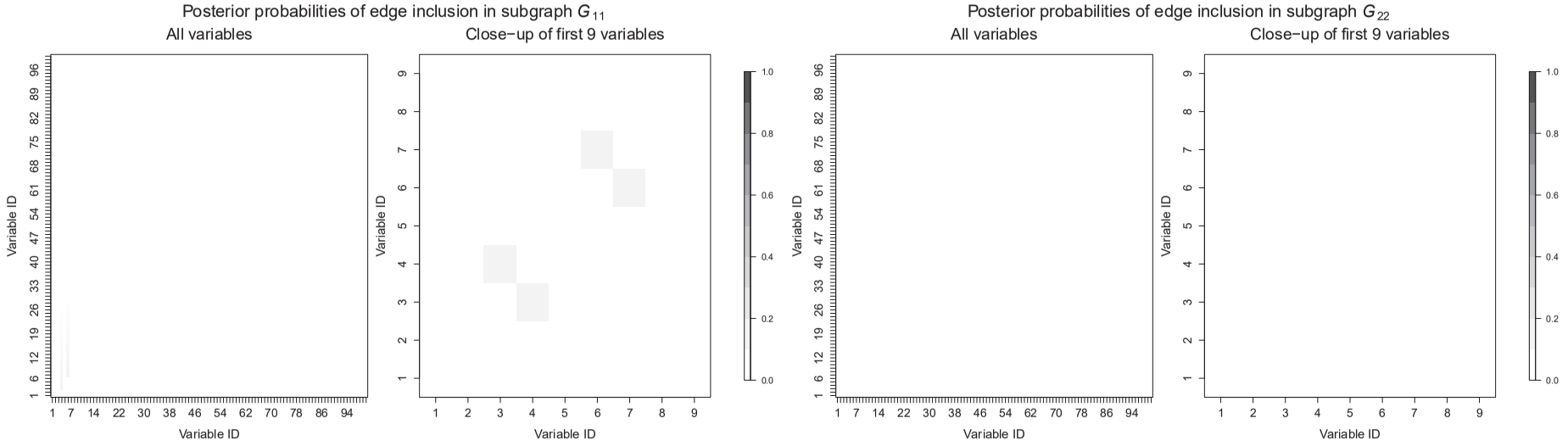
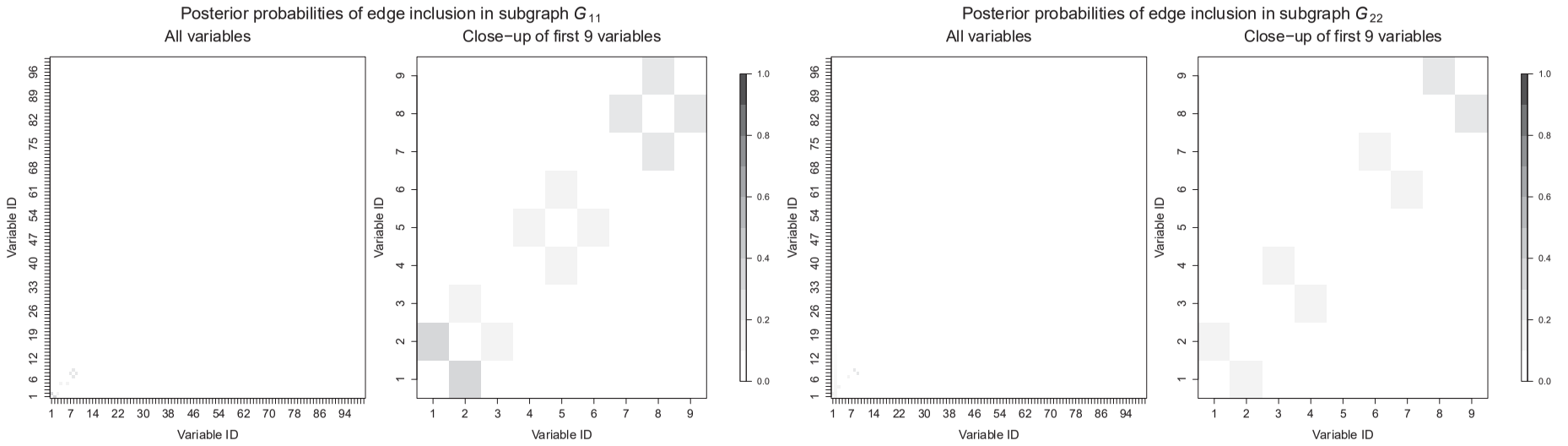


FIGURE B.43: *Mean posterior probabilities of edge inclusion (average across all simulations) in subgraph $\boldsymbol{G}_{ss}$ for subgroup $s = 1$ (left) and $s = 2$ (right). Edges in the graph are assumed independent Bernoulli a priori with parameter $\pi = 2/(p-1)$ (cont.).*

(D) $n = 50$ and $p = 100$



(E) $n = 75$ and $p = 100$

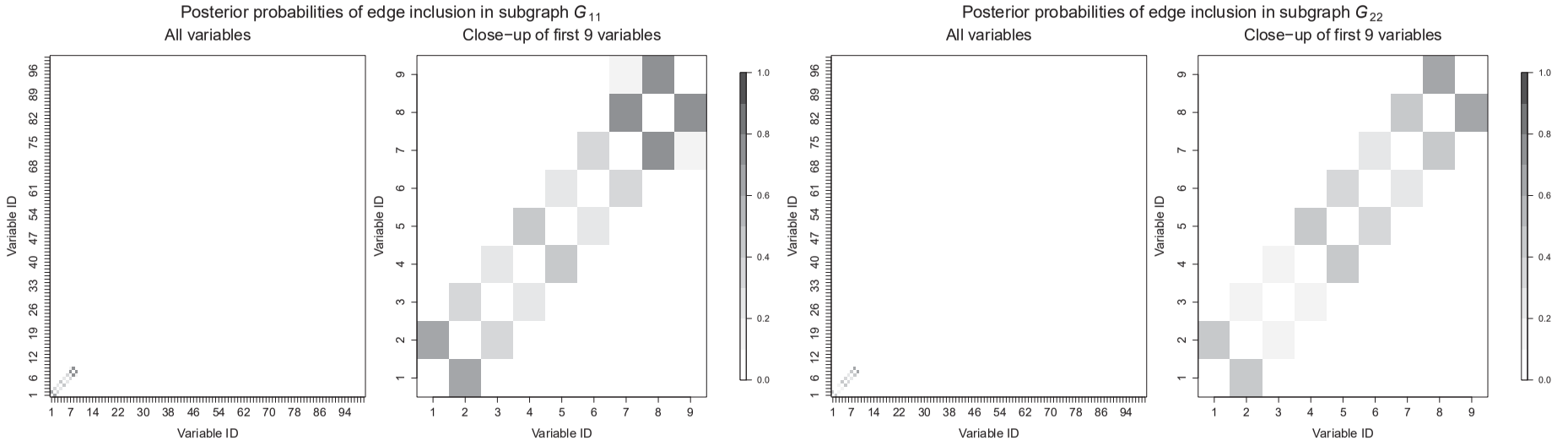
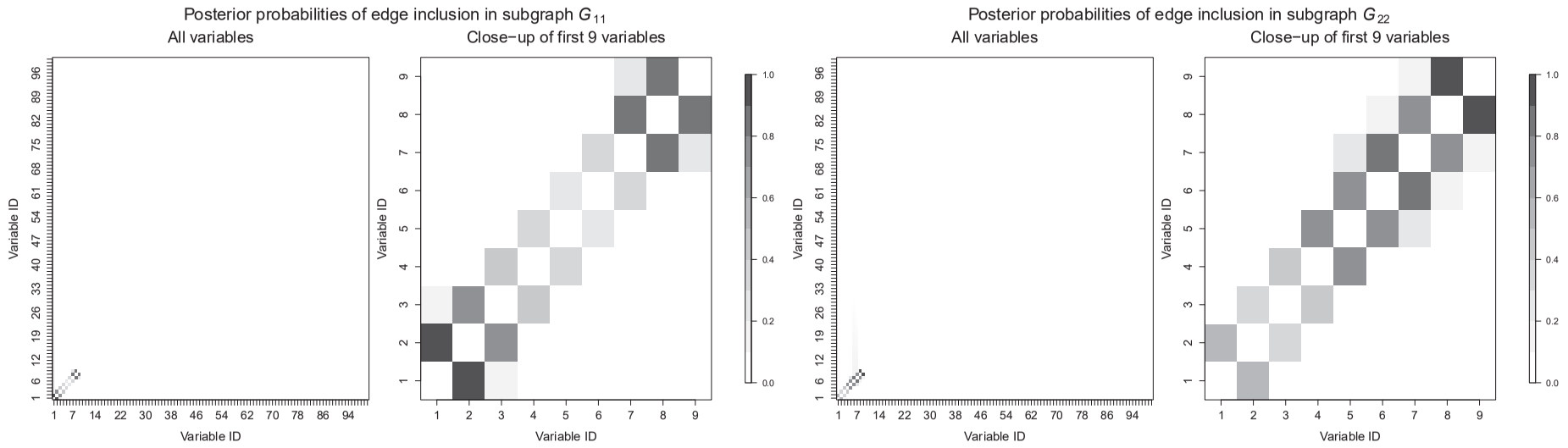


FIGURE B.43: *Mean posterior probabilities of edge inclusion (average across all simulations) in subgraph $\boldsymbol{G}_{ss}$ for subgroup $s = 1$ (left) and $s = 2$ (right). Edges in the graph are assumed independent Bernoulli a priori with parameter $\pi = 2/(p-1)$ (cont.).*

(F) $n = 100$ and $p = 100$



(G) $n = 150$ and $p = 100$

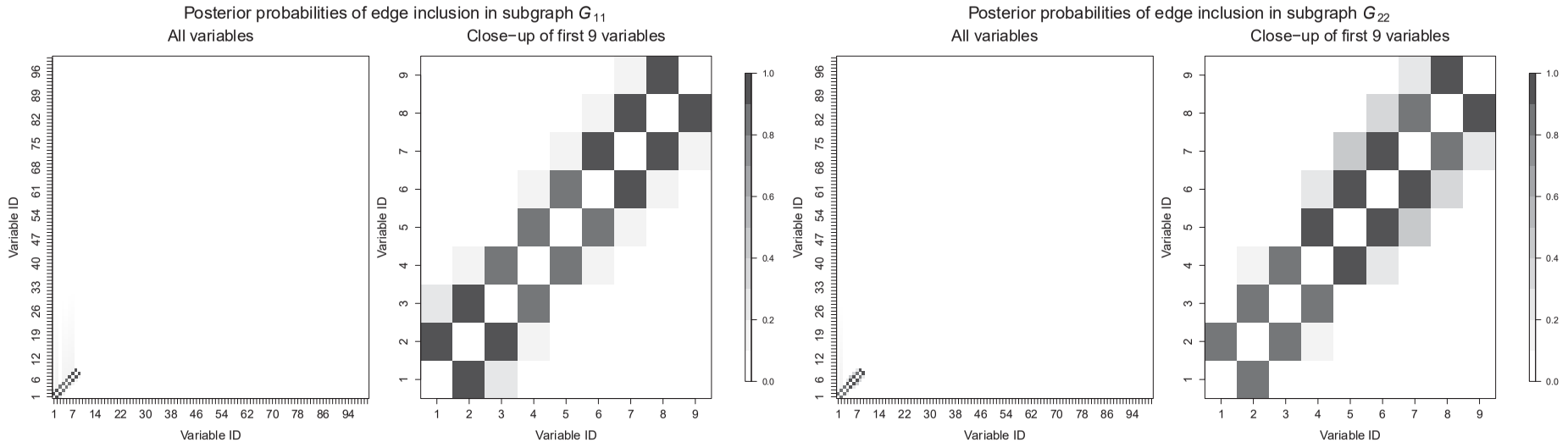


FIGURE B.43: *Mean posterior probabilities of edge inclusion (average across all simulations) in subgraph $\boldsymbol{G}_{ss}$ for subgroup $s = 1$ (left) and $s = 2$ (right). Edges in the graph are assumed independent Bernoulli a priori with parameter $\pi = 2/(p-1)$ (cont.).*

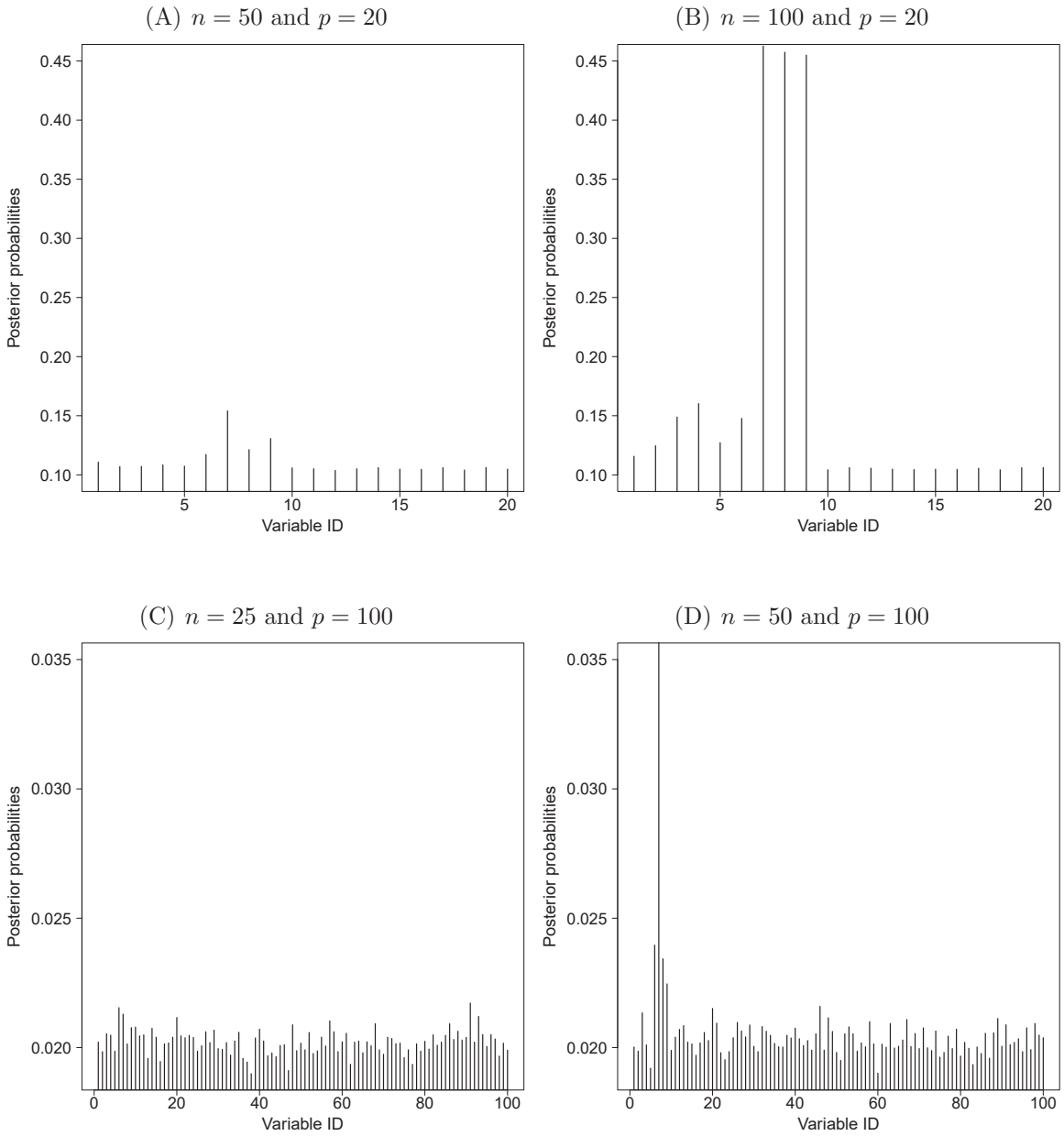


FIGURE B.44: *Posterior probabilities of edge inclusion for diagonal elements in subgraph* $\boldsymbol{G}_{12}$ *(average across all simulations). Edges in the graph are assumed independent Bernoulli a priori with parameter* $\pi = 2/(p-1)$.

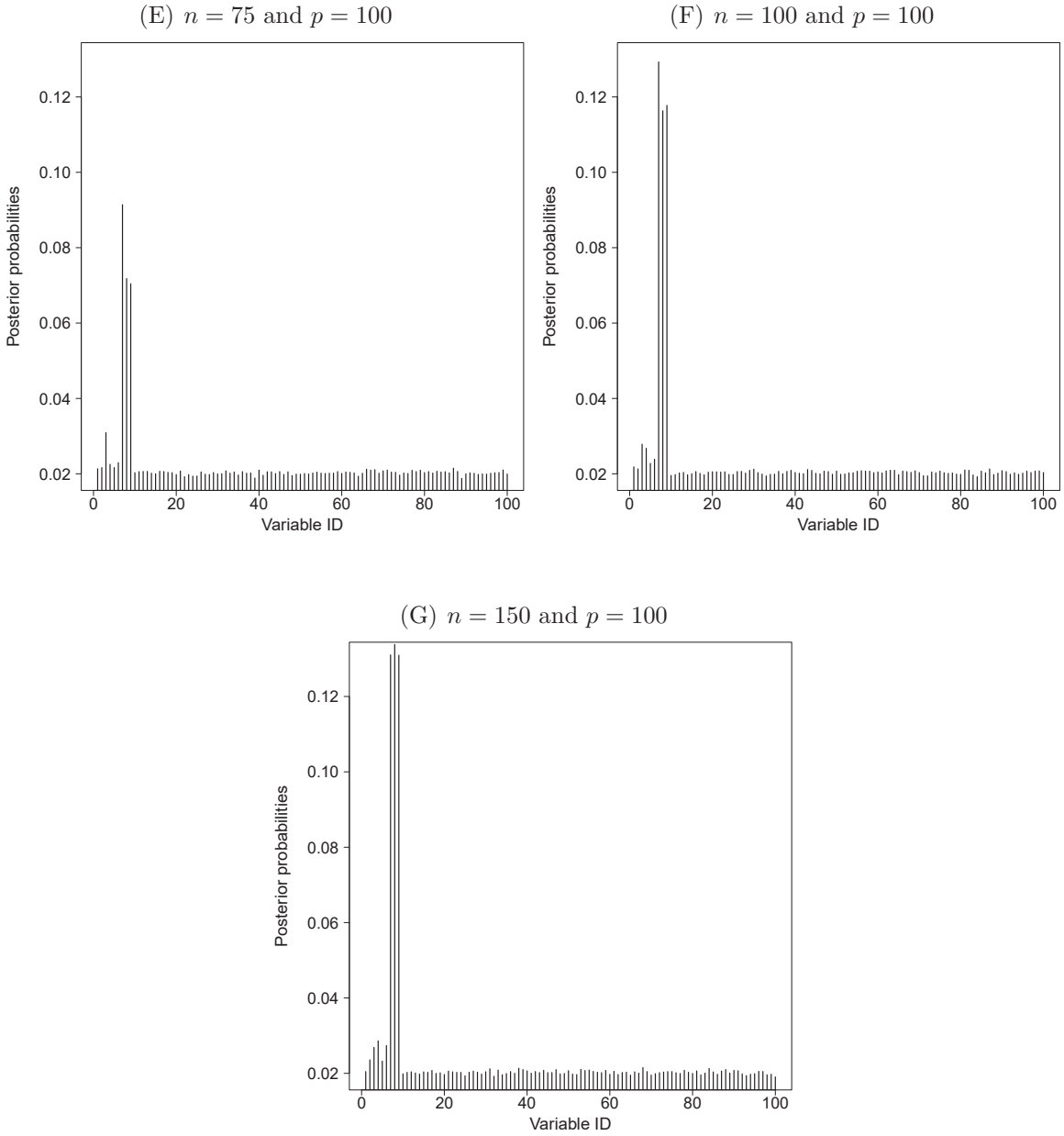


FIGURE B.44: *Posterior probabilities of edge inclusion for diagonal elements in subgraph* $\boldsymbol{G}_{12}$ *(average across all simulations). Edges in the graph are assumed independent Bernoulli a priori with parameter* $\pi = 2/(p-1)$ *(cont.).*

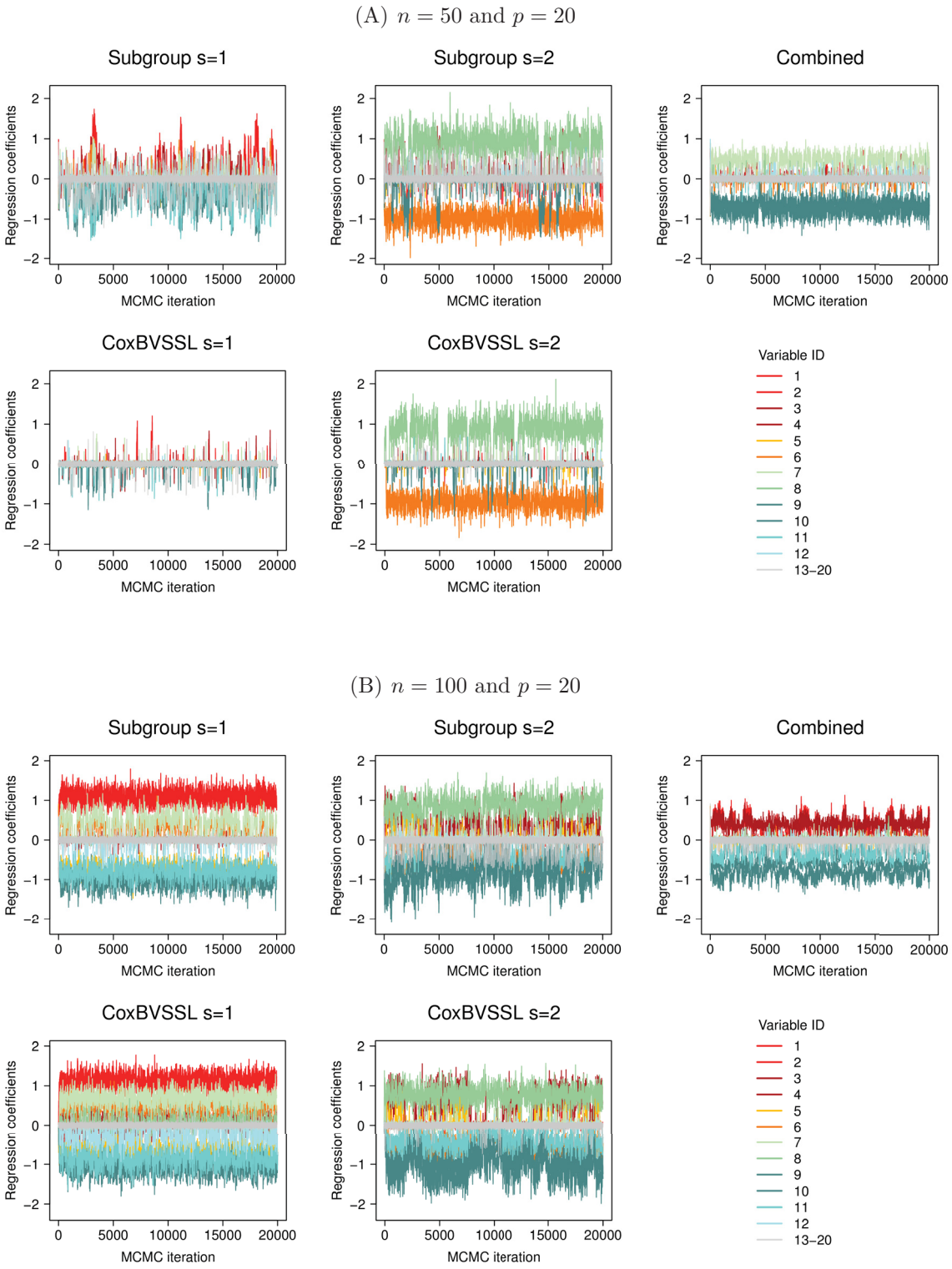


FIGURE B.45: *Trace plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. (A) $n = 50$, $p = 20$; (B) $n = 100$, $p = 20$; (C) $n = 50$, $p = 100$; (D) $n = 75$, $p = 100$; (E) $n = 100$, $p = 100$; (F) $n = 200$, $p = 100$.*

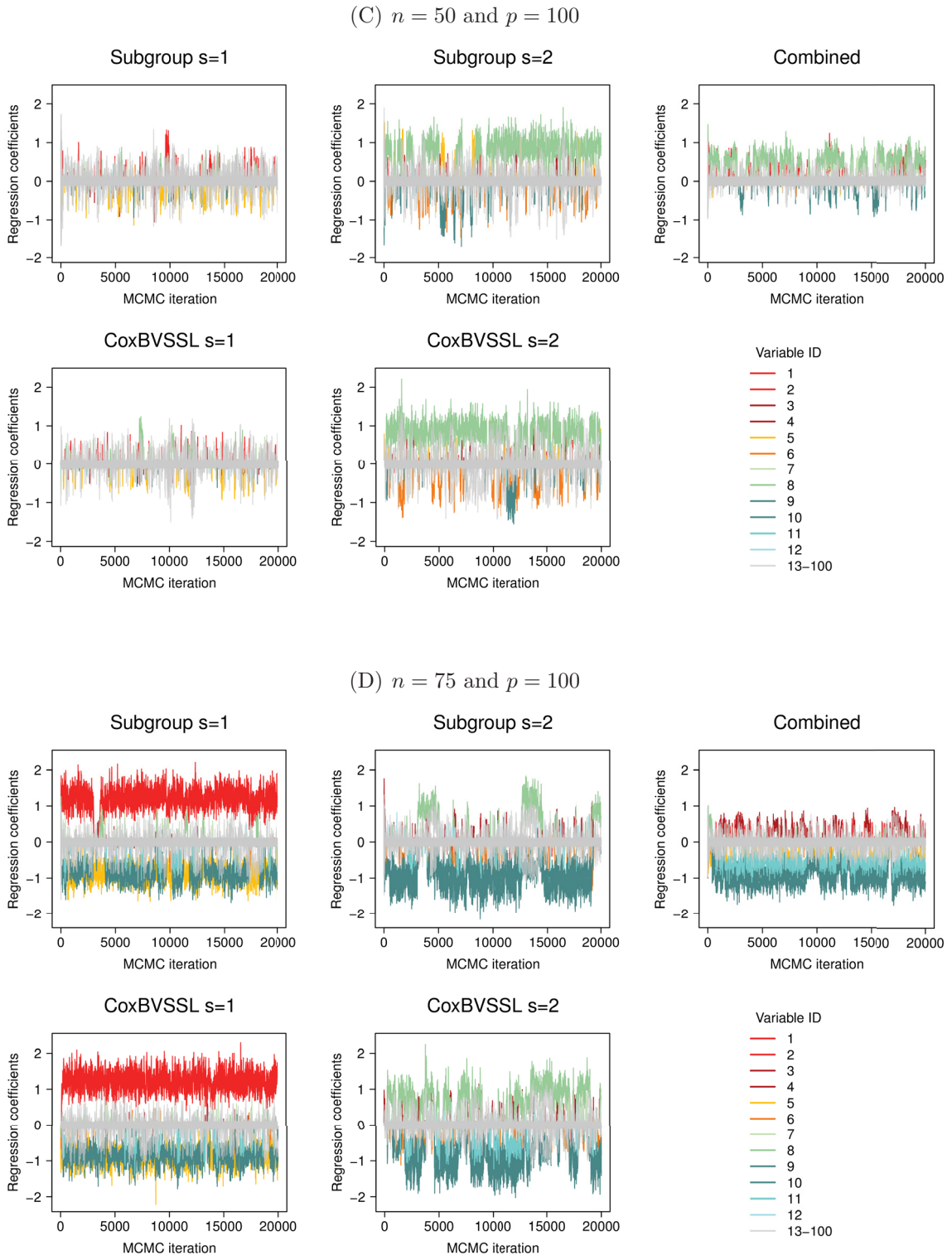
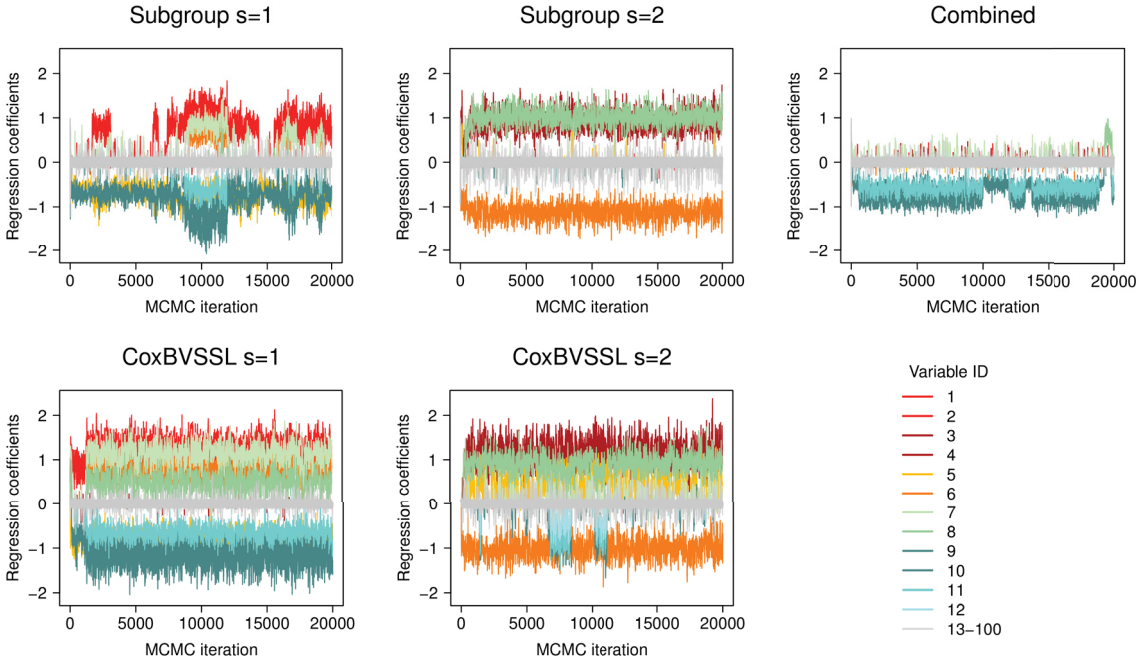


FIGURE B.45: *Trace plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. (A) $n = 50$, $p = 20$; (B) $n = 100$, $p = 20$; (C) $n = 50$, $p = 100$; (D) $n = 75$, $p = 100$; (E) $n = 100$, $p = 100$; (F) $n = 200$, $p = 100$ (cont.).*

(E) $n = 100$ and $p = 100$



(F) $n = 200$ and $p = 100$

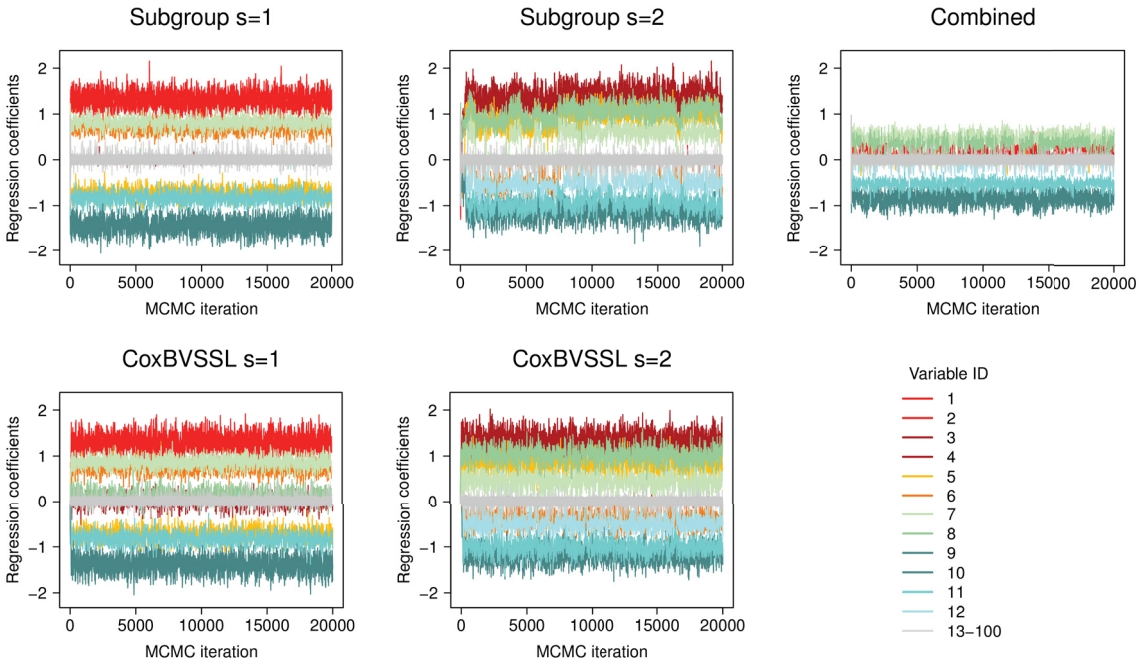
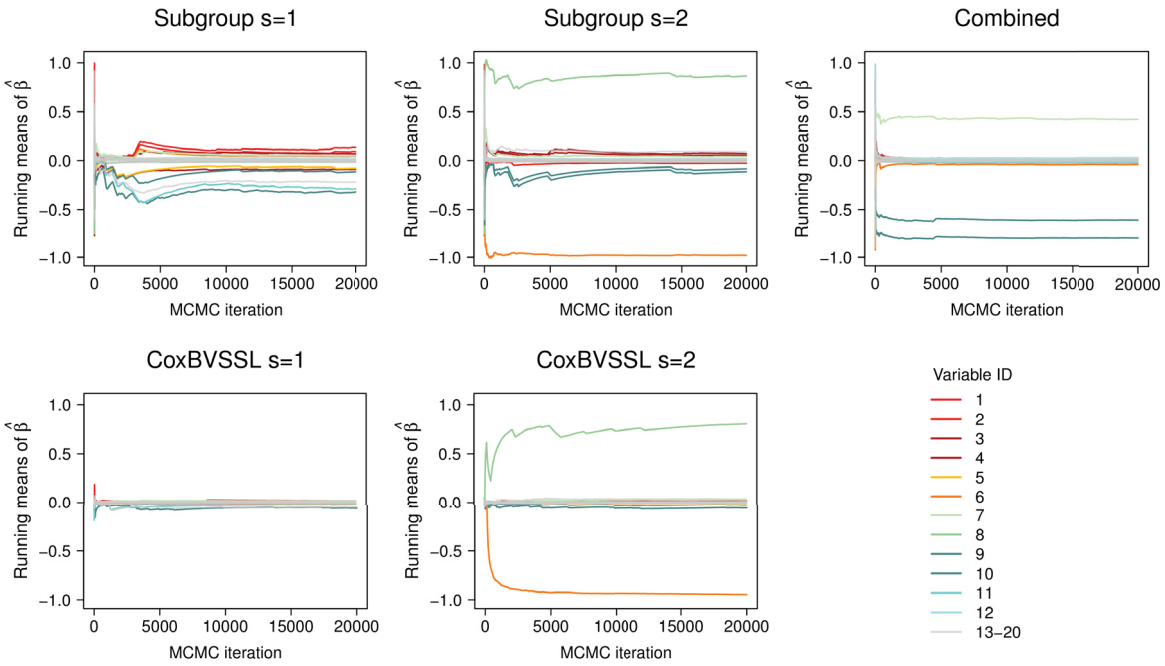


FIGURE B.45: *Trace plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, CoxBVSSL), subgroups $s = 1, 2$ and varying n and p. (A) $n = 50$, $p = 20$; (B) $n = 100$, $p = 20$; (C) $n = 50$, $p = 100$; (D) $n = 75$, $p = 100$; (E) $n = 100$, $p = 100$; (F) $n = 200$, $p = 100$ (cont.).*

(A) $n = 50$ and $p = 20$



(B) $n = 100$ and $p = 20$

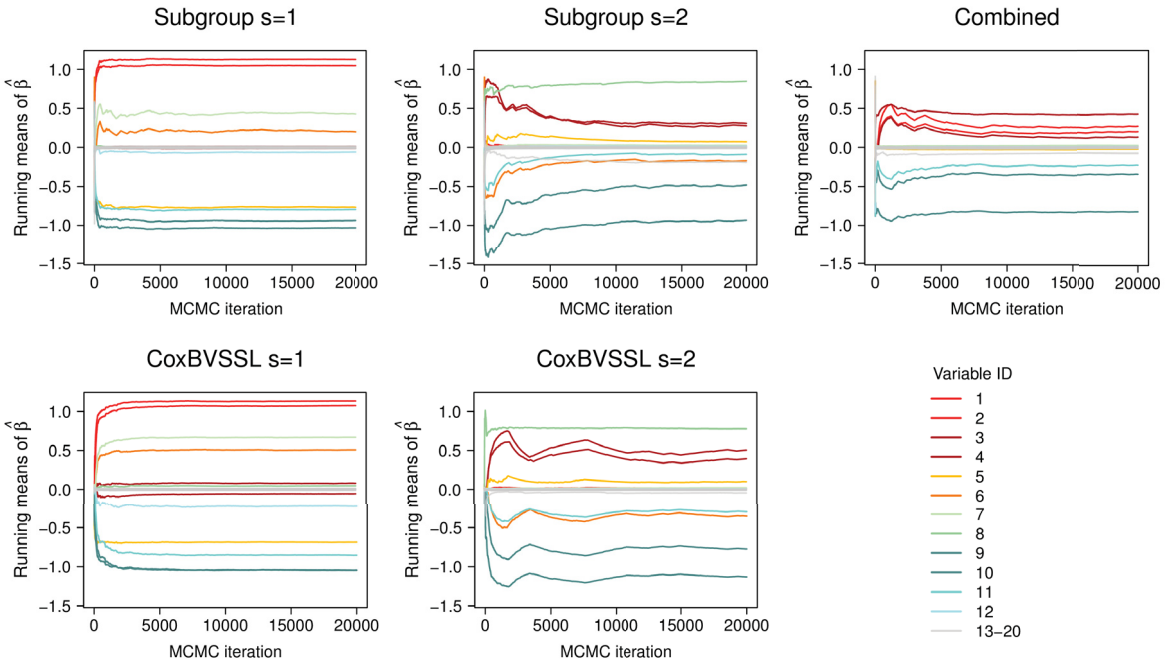


FIGURE B.46: *Running mean plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, CoxBVSSL), subgroups $s = 1, 2$ and varying n and p.*

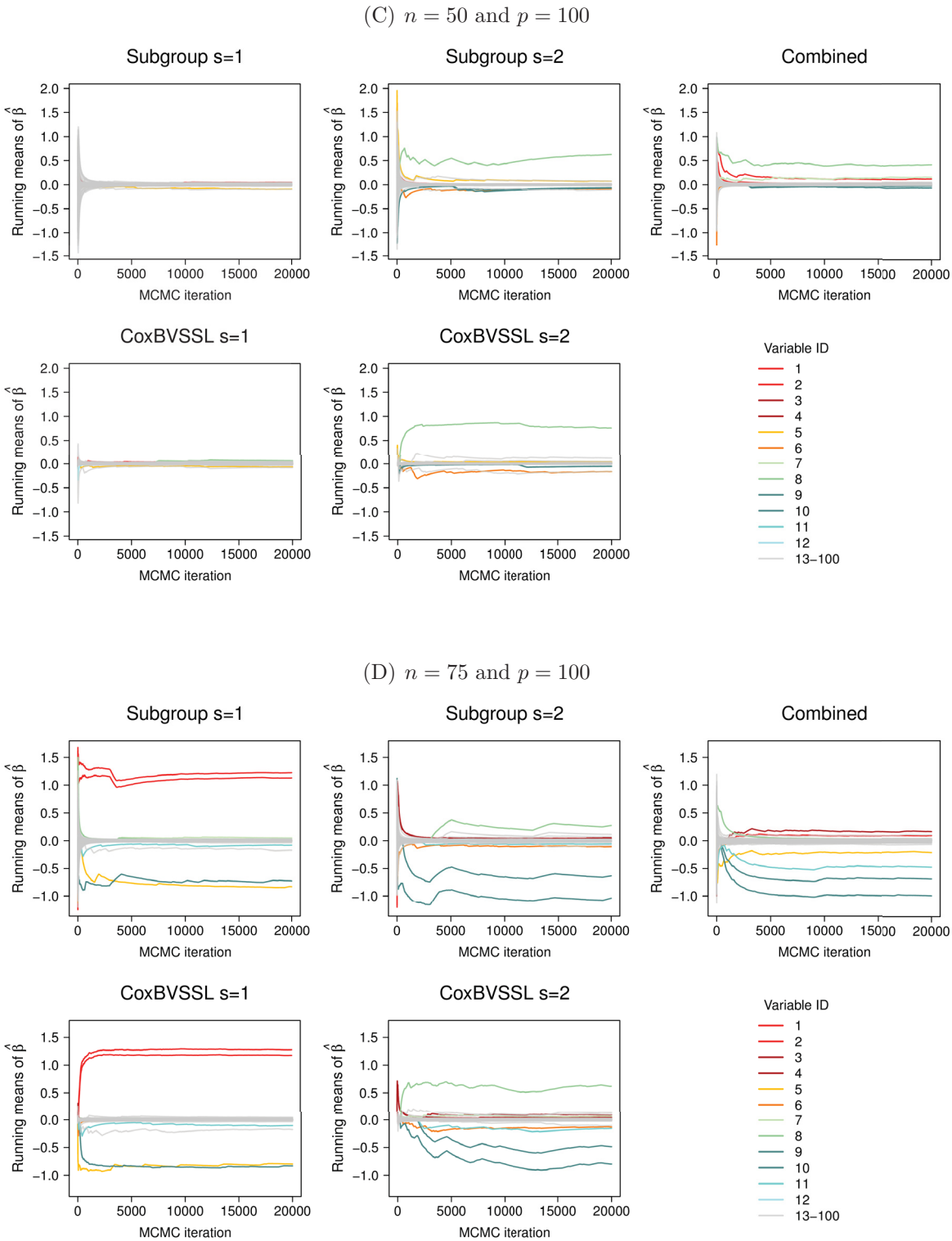


FIGURE B.46: *Running mean plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, CoxBVSSL), subgroups $s = 1, 2$ and varying n and p (cont.).*

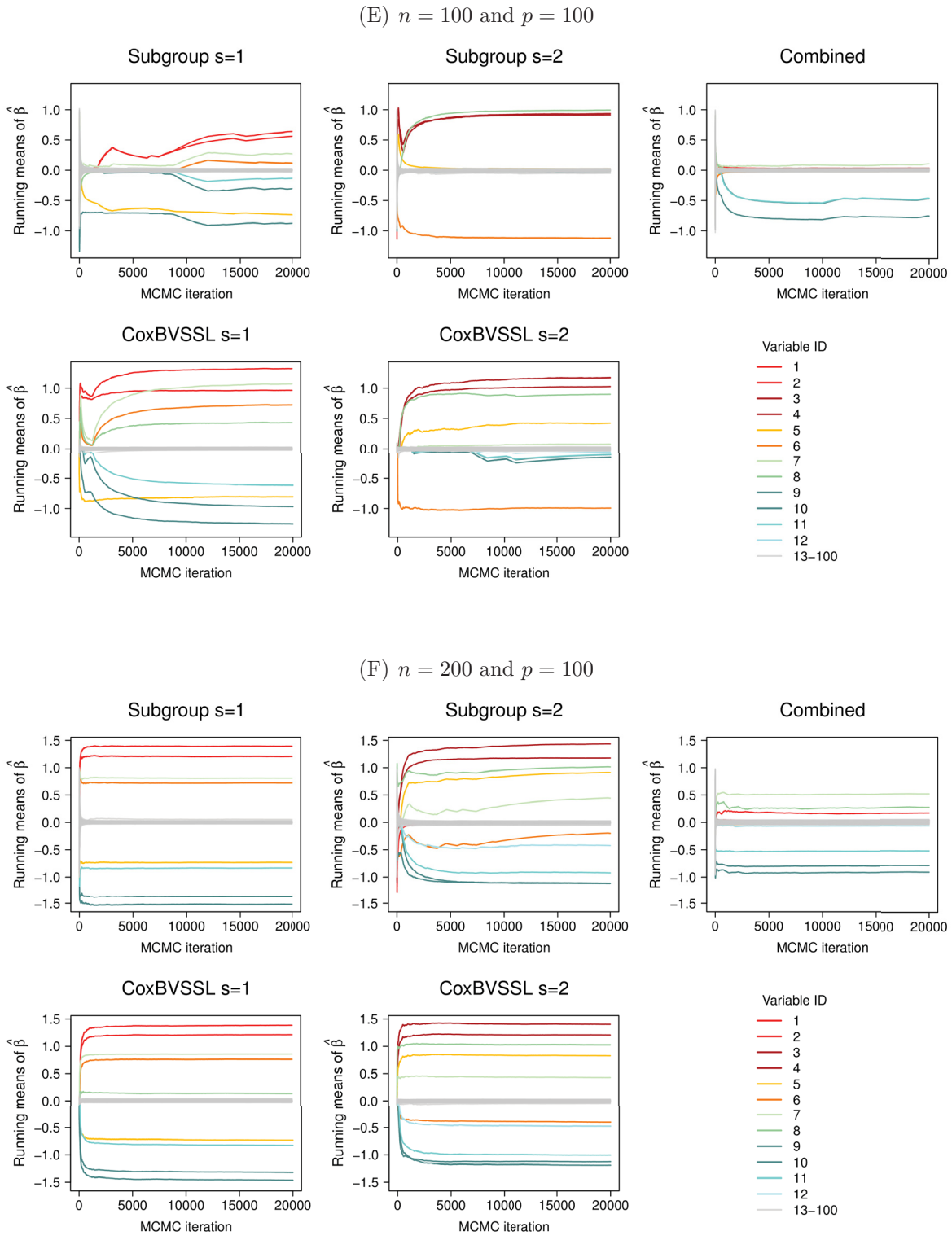
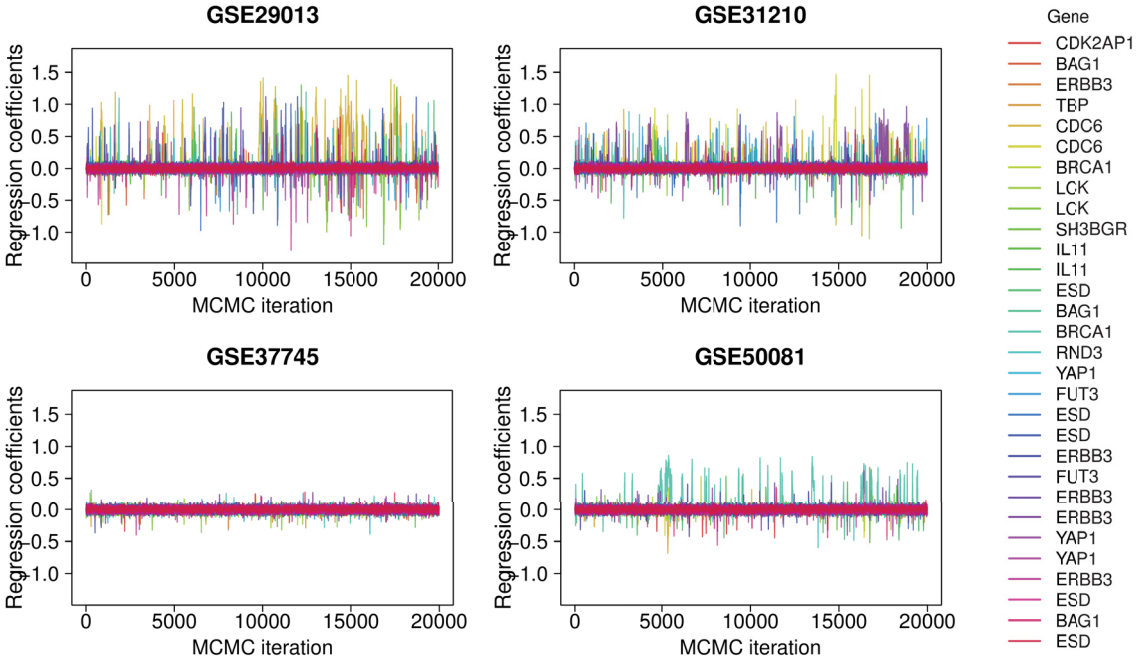


FIGURE B.46: *Running mean plots of the estimated regression coefficients from the first simulation for three different Cox models (subgroup, combined, CoxBVSSL), subgroups $s = 1, 2$ and varying n and p (cont.).*

(A) CoxBVSSL model



(B) Subgroup model

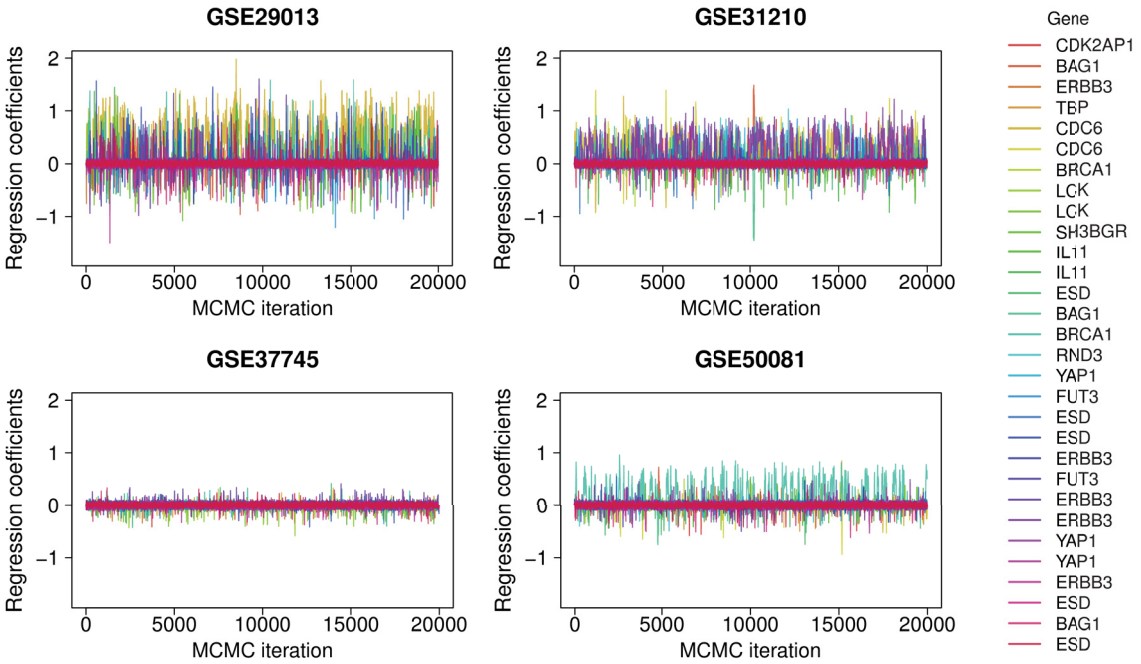
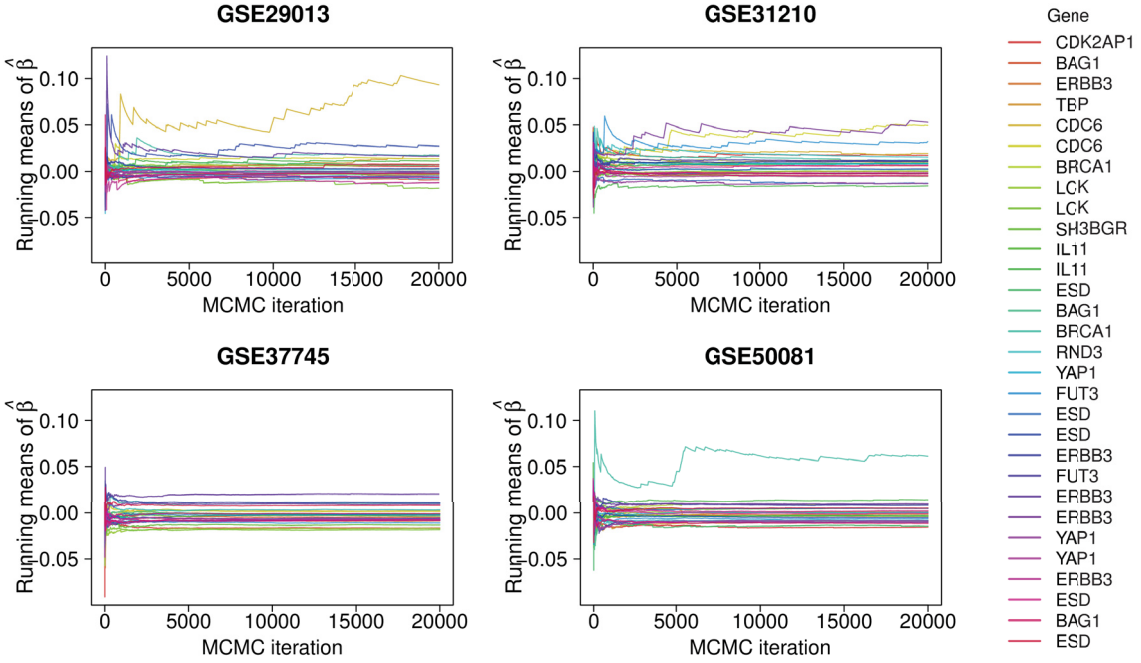


FIGURE B.47: *Trace plots of the estimated regression coefficients of the 30 Kratz genes for (A) the CoxBVSSL model, and (B) the subgroup model, in the first training data set.*

(A) CoxBVSSL model



(B) Subgroup model

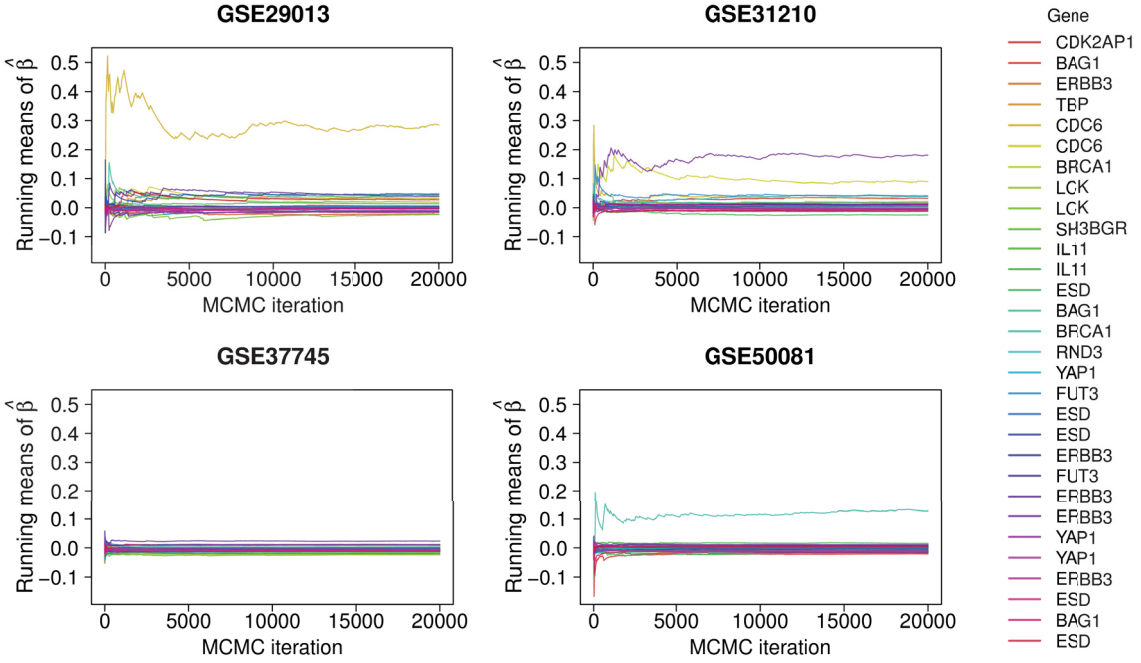


FIGURE B.48: *Running mean plots of the estimated regression coefficients of the 30 Kratz genes for (A) the CoxBVSSL model, and (B) the subgroup model, in the first training data set.*

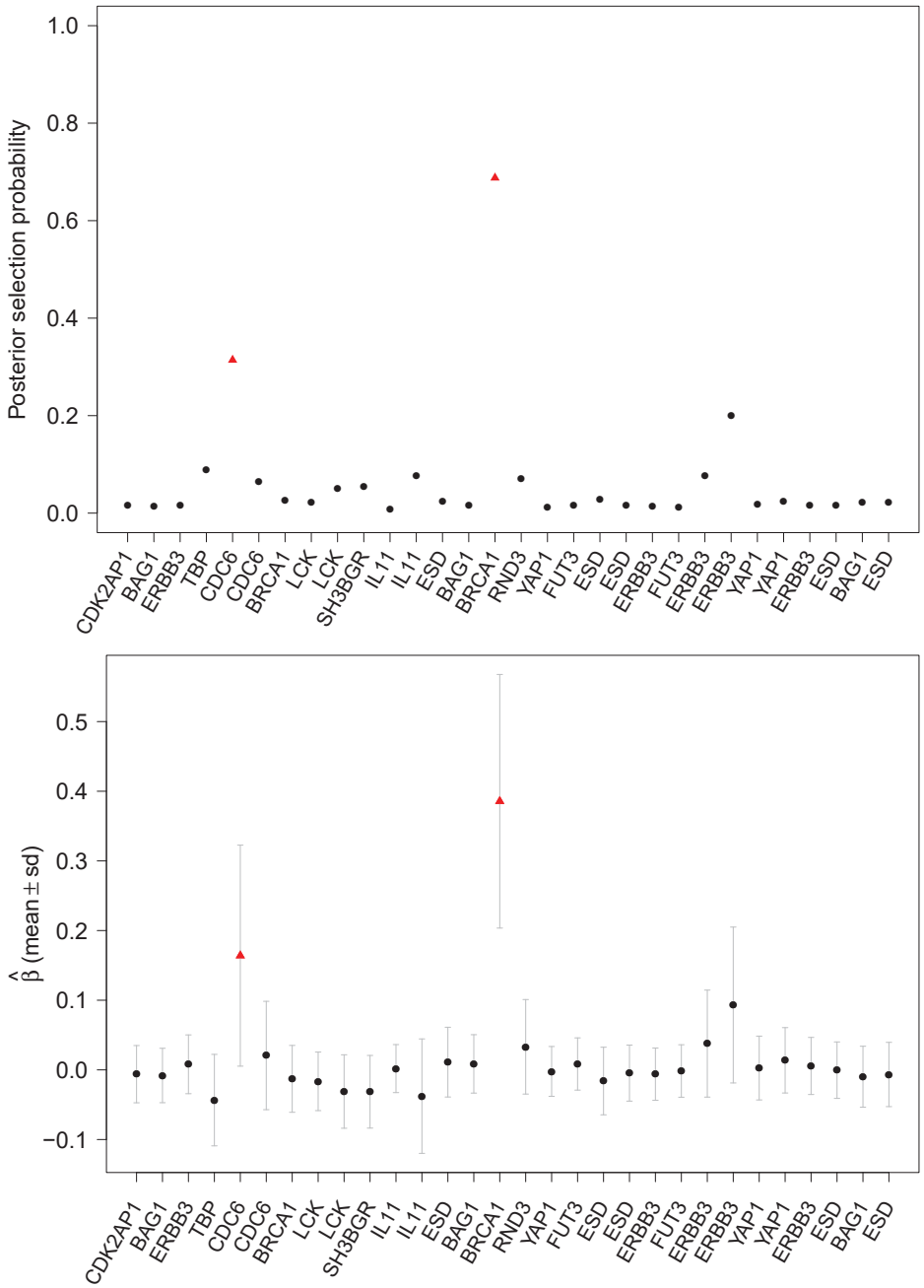
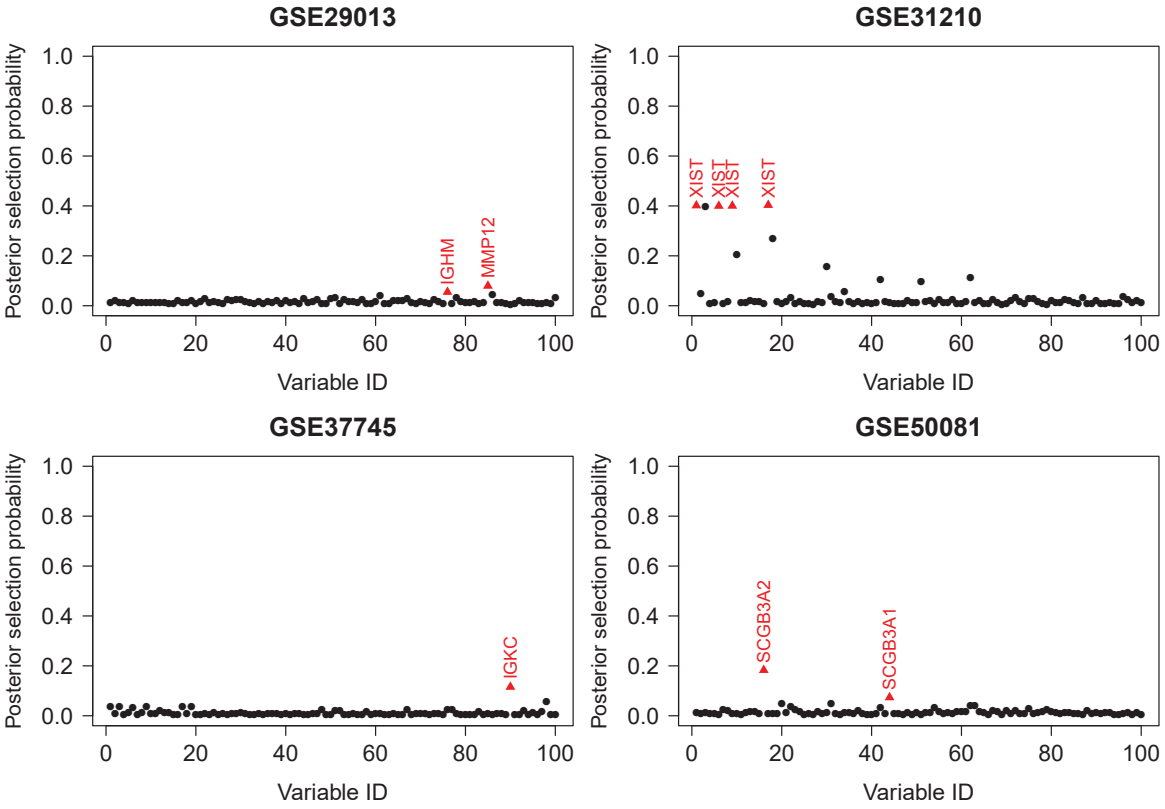


FIGURE B.49: *Mean posterior inclusion probabilities (top) and posterior mean/standard deviation (bottom) of the regression coefficients of the 30 Kratz genes (average across all training data sets) for the combined model. Selected variables are highlighted as red triangles.*

(A) CoxBVSSL model



(B) Subgroup model

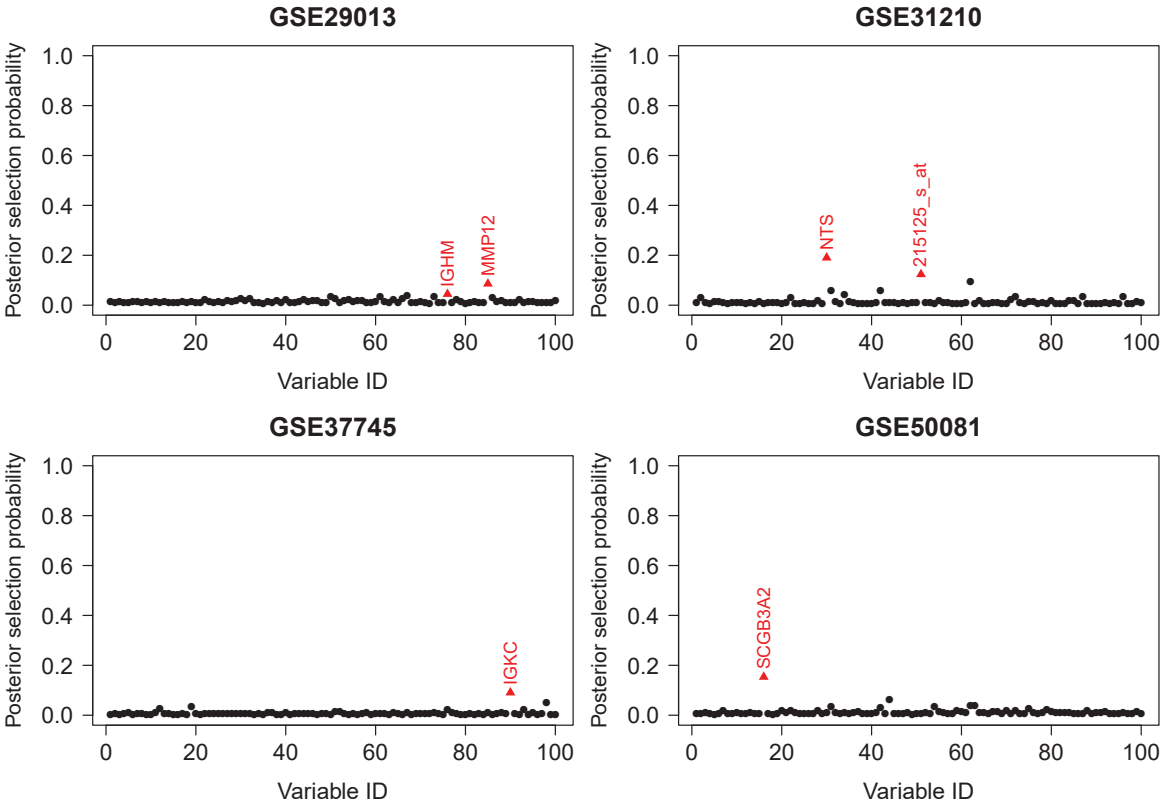
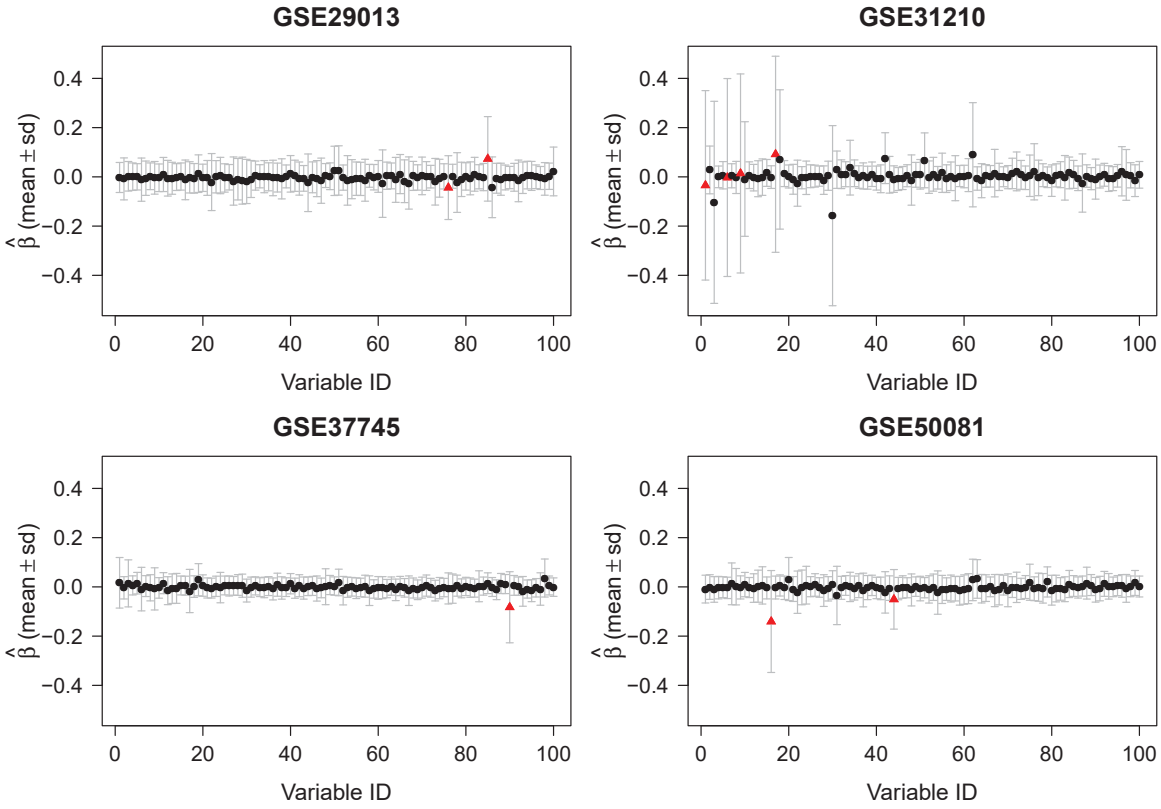


FIGURE B.50: *Mean posterior inclusion probabilities of the top-100-variance genes (average across all training data sets) for (A) the CoxBVSSL model, and (B) the subgroup model. Selected variables are highlighted as red triangles.*

(A) CoxBVSSL model



(B) Subgroup model

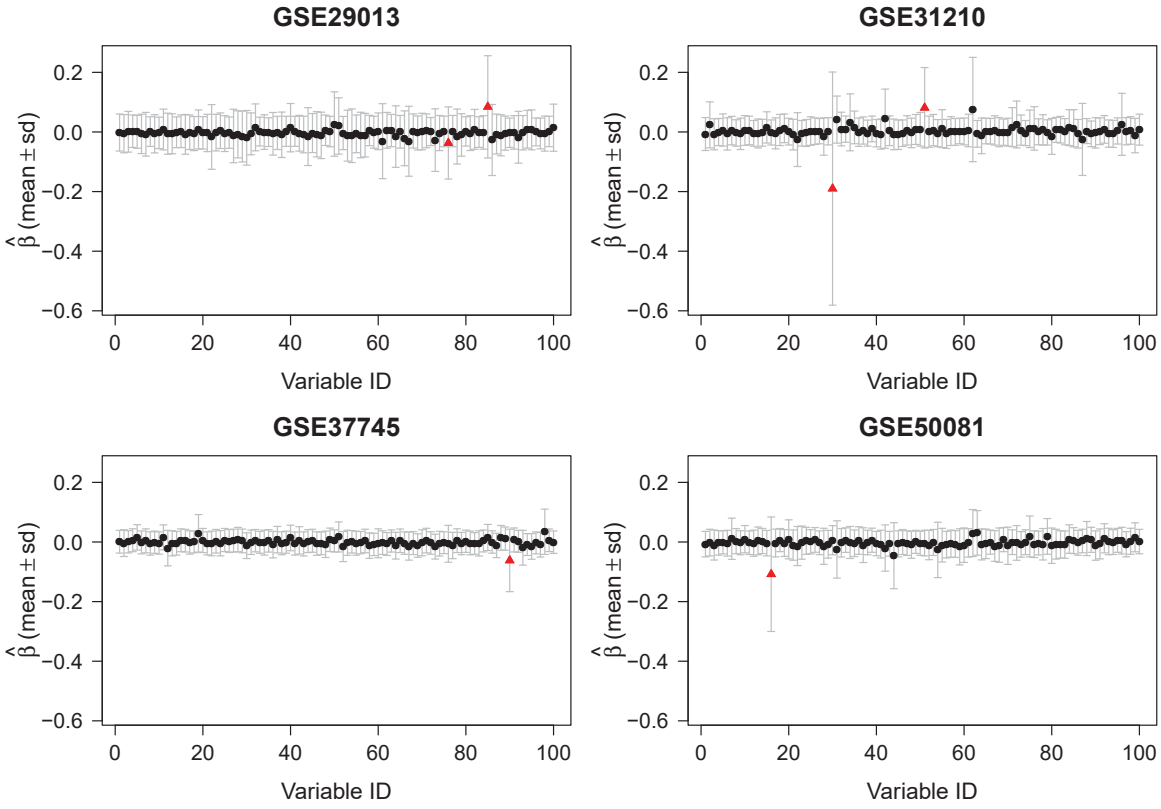


FIGURE B.51: *Posterior mean and standard deviation (sd) of the regression coefficients of the top-100-variance genes (average across all training data sets) for (A) the CoxBVSSL model, and (B) the subgroup model. Selected variables are highlighted as red triangles.*

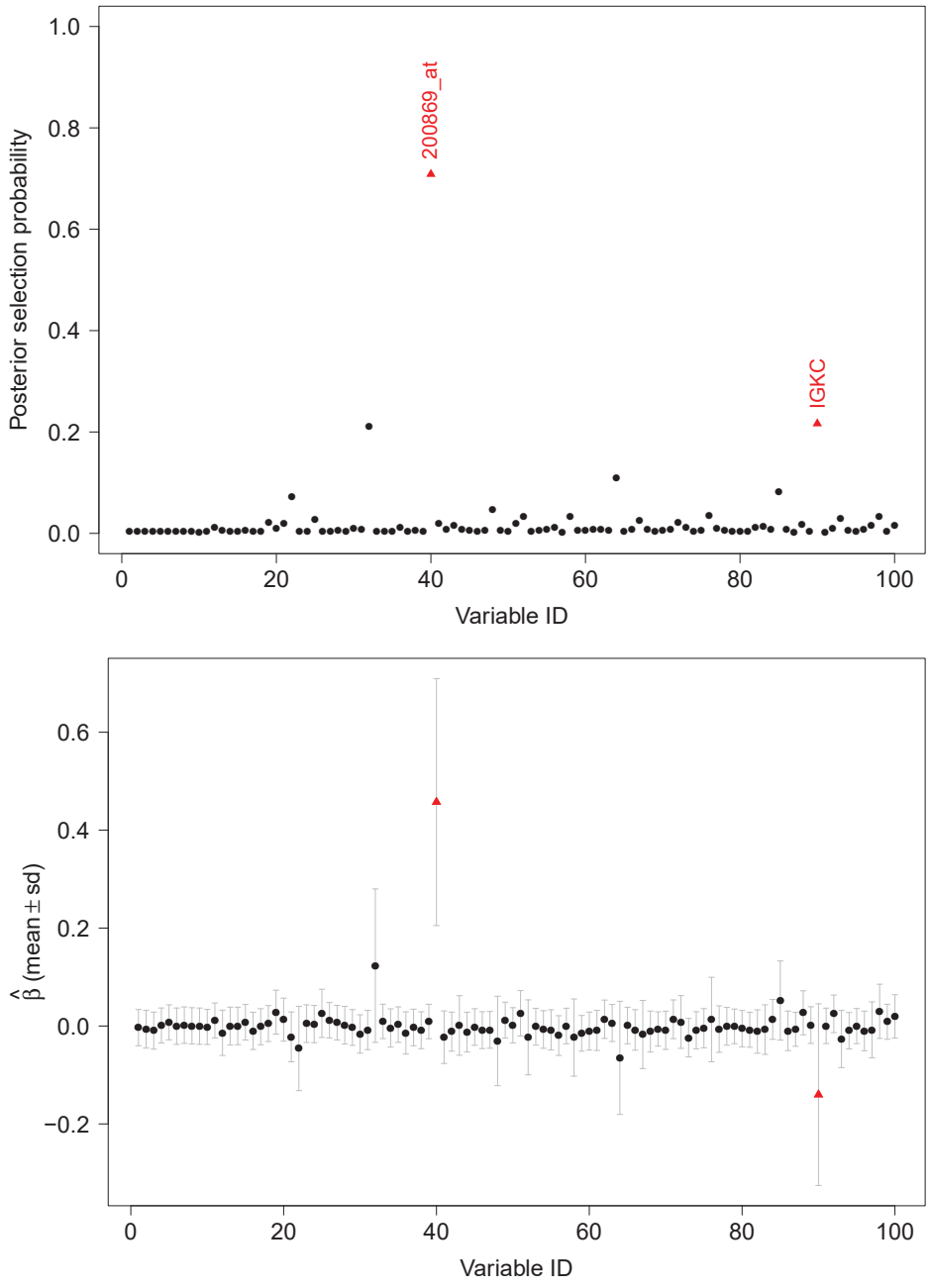
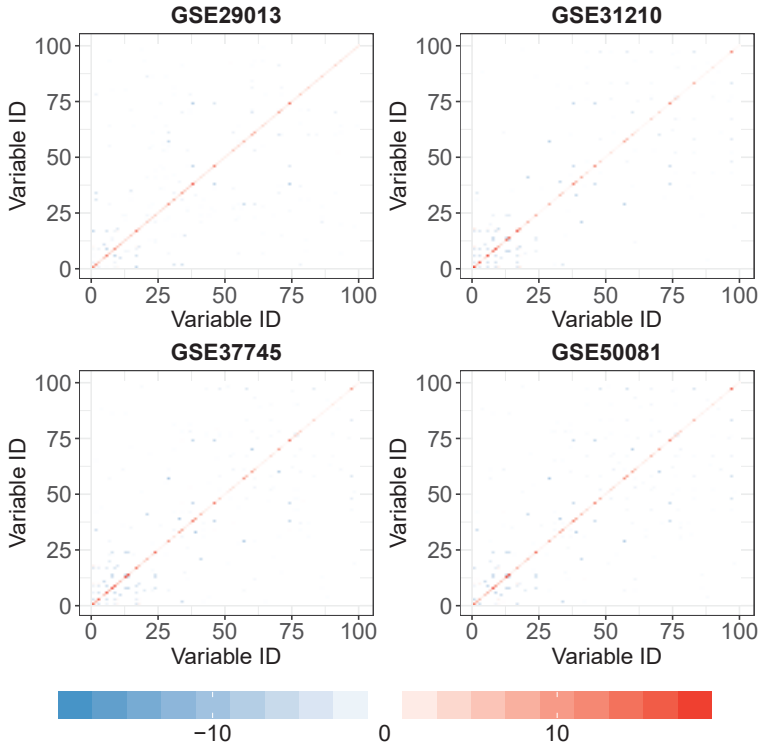


FIGURE B.52: *Mean posterior inclusion probabilities (top) and posterior mean/standard deviation (bottom) of the regression coefficients of the top-100-variance genes (average across all training data sets) for the combined model. Selected variables are highlighted as red triangles.*

(A) Posterior mean of elements in precision matrix $\boldsymbol{\Omega}_s$



(B) Mean posterior probabilities of edge inclusion in subgraph $\boldsymbol{G}_{ss}$

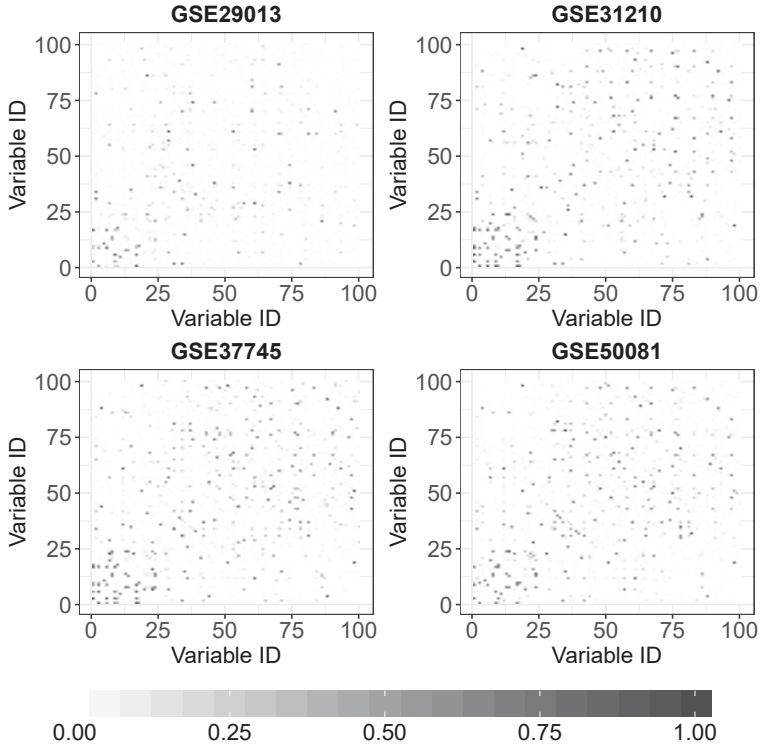


FIGURE B.53: *Mean posterior estimates of precision matrix $\boldsymbol{\Omega}_s$ and subgraph $\boldsymbol{G}_{ss}$ for all subgroups s and the top-100-variance genes (average across all training data sets). Edges in the graph are assumed independent Bernoulli a priori with parameter $\pi = 2/(p-1)$, $p = 100$. The prior of the diagonal entries of the precision matrix is exponential with parameter $\frac{1}{2}$, and the prior of the off-diagonal entries is a mixture of two normal distributions with zero mean and variance $\nu_0^2 = 0.1^2$ for non-selected edges and variance $\nu_1^2 = 5^2$ for selected edges.*

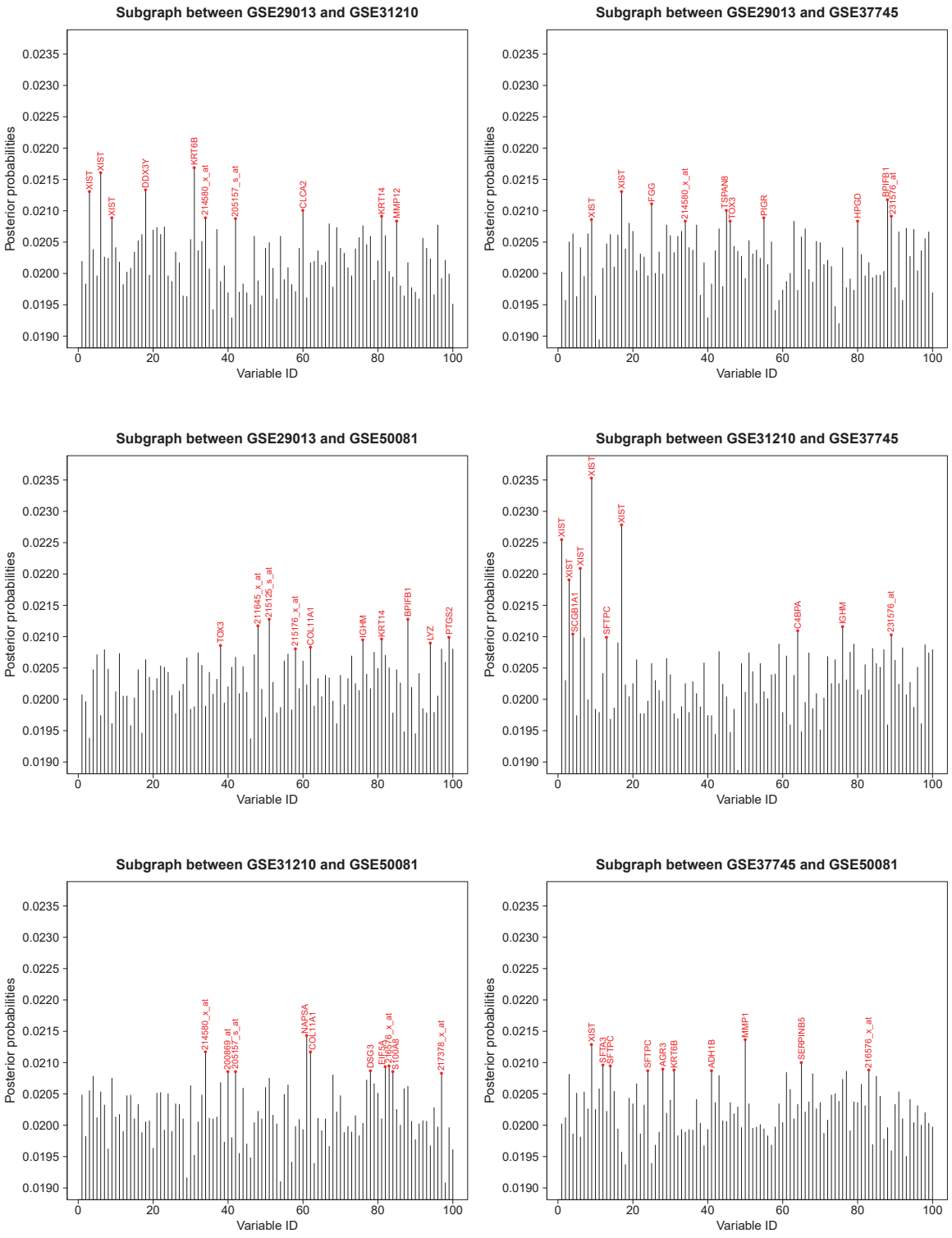
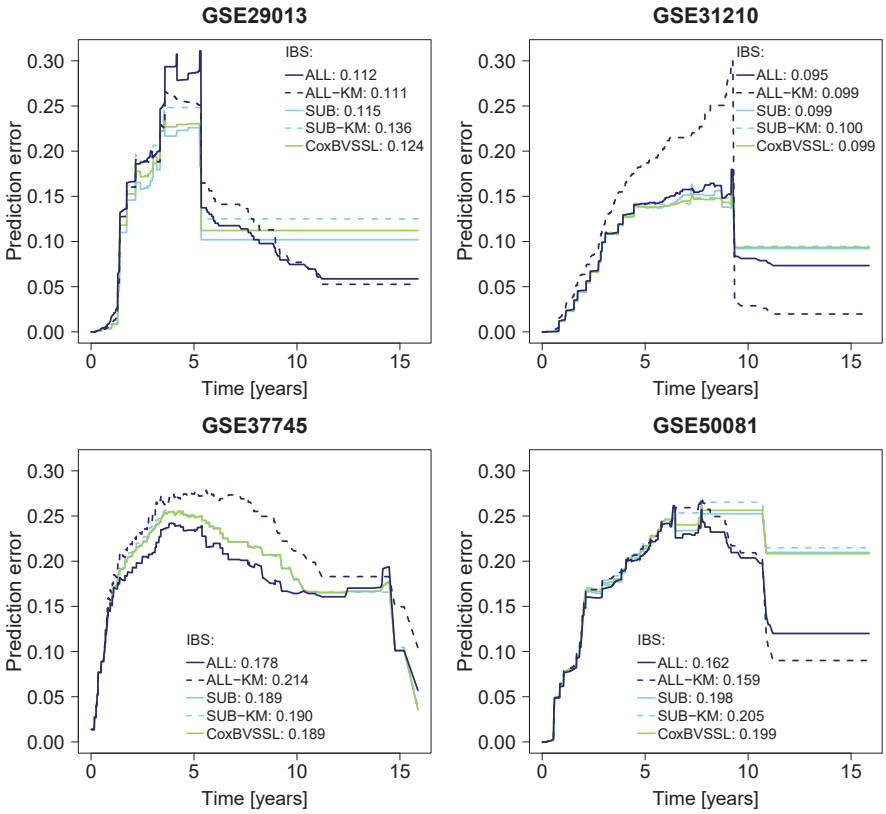


FIGURE B.54: *Posterior probabilities of edge inclusion (PPI) for diagonal elements in subgraphs $\boldsymbol{G}_{rs,r<s}$ between subgroups r and s for the top-100-variance genes (average across all training data sets). Edges in the graph are assumed independent Bernoulli a priori with parameter $\pi = 2/(p-1)$, $p = 100$. The ten genes with highest PPI are highlighted in red.*

(A) 30 Kratz genes



(B) Top-100-variance genes

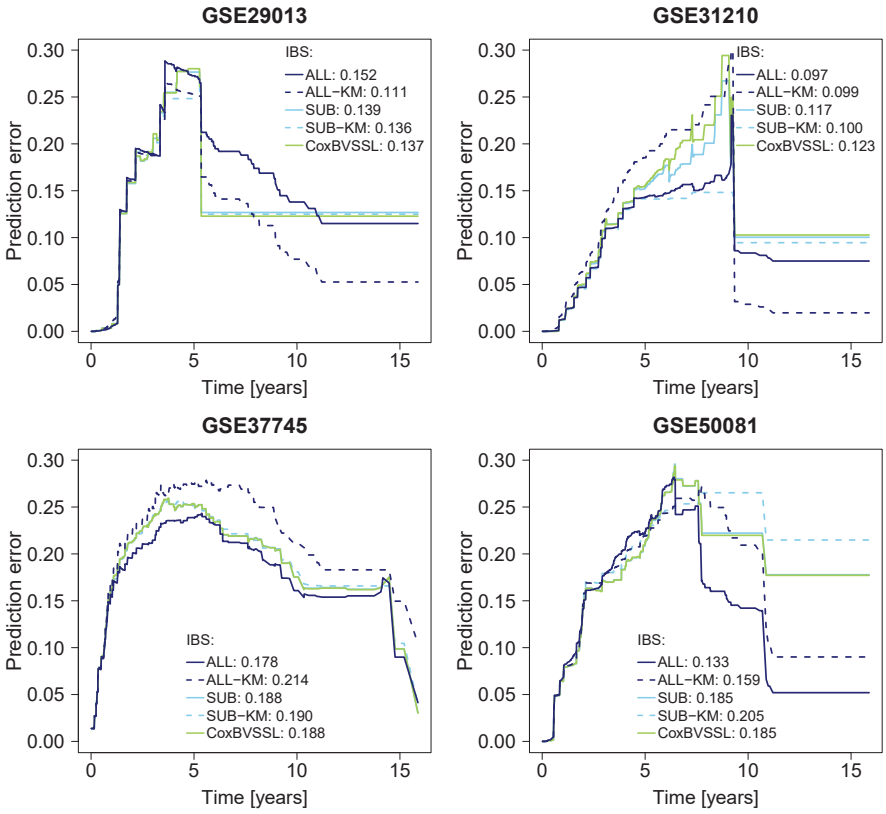


FIGURE B.55: *Prediction error curves and integrated Brier scores (IBS) based on the first test data set. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (SUB-KM) or combined (ALL-KM) training data. ALL: combined model; SUB: subgroup model.*

Appendix C

Tables

Variable	Values	GSE29013	GSE31210	GSE37745	GSE50081
Sample size		55	226	194	160
Age (years)	min.	32	30	39	40
	mean	64	60	64	68
	max.	76	76	84	87
Sex	male	38	105	105	88
	female	17	121	89	72
pTNM stage	I	24	168	128	112
	II-IV	31	58	66	48
Histology	SQC	25	0	64	35
	ADC	30	226	106	115
	other NSCLC	0	0	24	10
Smoking status	never-smoker	2	115	15	24
	current/ former smoker	53	111	179	136
Survival status	censoring	37	191	51	95
	event	18	35	143	65

TABLE C.1: *Summary of clinical pathologic variables of all NSCLC cohorts. Absolute frequencies of variable values.*

				Top five methods				
ϵ	p	n	$\boldsymbol{\Sigma}$	1	2	3	4	5
0	12	50	uncorrelated	rf-4 0.674	rf-1 0.661	rf-3 0.659	rf-2 0.657	ℓ_2-3 0.592
0	12	200	uncorrelated	rf-4 0.637	rf-1 0.619	rf-3 0.617	rf-2 0.616	ℓ_2-3 0.584
0	12	1000	uncorrelated	ℓ_1-4 0.698	ℓ_2-4 0.698	rf-1 0.695	rf-2 0.695	rf-4 0.694
0	100	50	block	ℓ_2-3 0.524	ℓ_1-3 0.510	rf-4 0.500	rf-3 0.499	ℓ_1-4 0.497
0	100	50	blockdiag	rf-3 0.560	rf-4 0.537	ℓ_2-3 0.534	ℓ_1-3 0.516	rf-2 0.496
0	100	50	min01	rf-3 0.519	ℓ_1-2 0.488	ℓ_1-4 0.488	ℓ_1-1 0.482	ℓ_1-3 0.481
0	100	50	shrinkage	rf-4 0.567	rf-1 0.518	rf-2 0.516	rf-3 0.482	ℓ_1-4 0.477
0	100	50	uncorrelated	rf-4 0.566	rf-3 0.532	rf-1 0.506	ℓ_1-1 0.503	rf-2 0.498
0	100	200	block	rf-4 0.617	rf-3 0.599	rf-2 0.574	rf-1 0.573	ℓ_2-3 0.555
0	100	200	blockdiag	rf-3 0.628	ℓ_1-3 0.621	rf-4 0.619	ℓ_2-3 0.592	ℓ_1-1 0.572
0	100	200	min01	rf-4 0.567	rf-3 0.538	ℓ_1-3 0.515	rf-2 0.512	rf-1 0.510
0	100	200	shrinkage	rf-4 0.626	rf-3 0.611	rf-2 0.584	rf-1 0.584	ℓ_2-3 0.573
0	100	200	uncorrelated	rf-4 0.612	rf-3 0.594	rf-1 0.569	rf-2 0.569	ℓ_2-3 0.549
0	100	1000	block	ℓ_1-3 0.620	rf-4 0.620	rf-3 0.620	rf-2 0.596	ℓ_1-4 0.595
0	100	1000	blockdiag	ℓ_1-4 0.658	ℓ_1-3 0.657	rf-3 0.651	rf-4 0.644	rf-2 0.624
0	100	1000	min01	ℓ_1-3 0.634	rf-4 0.631	rf-3 0.629	ℓ_1-4 0.617	rf-2 0.612
0	100	1000	shrinkage	rf-4 0.645	rf-3 0.632	ℓ_1-3 0.630	rf-1 0.622	rf-2 0.622
0	100	1000	uncorrelated	rf-4 0.637	rf-3 0.629	ℓ_1-3 0.625	rf-2 0.616	rf-1 0.615
0	1000	50	block	rf-4 0.633	rf-2 0.569	rf-1 0.556	rf-3 0.556	ℓ_2-1 0.546
0	1000	50	blockdiag	rf-4 0.594	ℓ_2-4 0.505	ℓ_2-3 0.502	rf-2 0.500	rf-1 0.500
0	1000	50	min01	rf-4 0.594	rf-2 0.513	rf-1 0.509	ℓ_1-3 0.499	ℓ_1-4 0.495
0	1000	50	shrinkage	rf-4 0.618	rf-3 0.573	ℓ_2-3 0.560	rf-2 0.520	ℓ_1-3 0.516

0	1000	50	uncorrelated	rf-4	ℓ_1-3	rf-3	ℓ_1-2	ℓ_1-1
				0.532	0.507	0.503	0.502	0.501
0	1000	200	block	rf-4	rf-3	rf-1	ℓ_1-1	ℓ_1-2
				0.568	0.515	0.497	0.494	0.493
0	1000	200	blockdiag	rf-4	rf-3	ℓ_1-3	ℓ_1-1	ℓ_2-3
				0.584	0.538	0.512	0.510	0.509
0	1000	200	min01	rf-4	rf-3	ℓ_1-3	rf-1	rf-2
				0.600	0.547	0.512	0.508	0.506
0	1000	200	shrinkage	rf-4	rf-3	ℓ_2-3	rf-2	rf-1
				0.602	0.578	0.549	0.528	0.526
0	1000	200	uncorrelated	rf-4	rf-3	rf-1	rf-2	ℓ_1-3
				0.615	0.537	0.525	0.525	0.503
0	1000	1000	block	rf-4	ℓ_1-3	ℓ_1-1	rf-3	rf-1
				0.627	0.622	0.612	0.601	0.575
0	1000	1000	blockdiag	ℓ_1-3	rf-4	rf-3	ℓ_1-1	rf-1
				0.644	0.628	0.621	0.603	0.577
0	1000	1000	min01	rf-4	ℓ_1-3	ℓ_1-1	rf-3	rf-2
				0.616	0.615	0.604	0.595	0.561
0	1000	1000	shrinkage	rf-4	ℓ_1-3	rf-3	ℓ_1-1	ℓ_2-3
				0.638	0.628	0.620	0.609	0.577
0	1000	1000	uncorrelated	rf-4	ℓ_1-3	ℓ_1-1	rf-3	rf-1
				0.602	0.590	0.574	0.573	0.549

TABLE C.2: *Top five combinations of methods and parameters for weights estimation in terms of highest mean AUC (mean across all training sets). Results for $\epsilon = 1$ are not shown since all top five methods perform equally well with AUC=1 and therefore, are not distinguishable. rf = random forest, ℓ_1 = lasso, ℓ_2 = ridge; and 1 = no intera. & no cumHR, 2 = no intera. & cumHR, 3 = intera. & no cumHR, 4 = intera. & cumHR.*

					Top five methods				
ϵ	p	n	$\boldsymbol{\Sigma}$	$\tilde{p}$	1	2	3	4	5
0	12	50	uncorrelated	12	w=0.1	w=0.2	w=0.3	w=0.4	w=0.5
				12	4.519	4.708	4.809	4.922	4.985
0	12	200	uncorrelated	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	1.685	2.847	3.394	3.736	3.982
0	12	1000	uncorrelated	12	sub	w=0.1	w=0.2	ℓ_1-4	w=0.3
				12	0.677	3.012	3.613	3.907	3.956
0	100	50	block	12	w=0.4	w=0.3	w=0.5	w=0.2	w=0.6
				12	5.876	5.883	5.902	5.926	5.933
				100	w=0.3	w=0.4	w=0.5	w=0.6	w=0.7
				100	6.645	6.649	6.652	6.675	6.697
0	100	50	blockdiag	12	w=0.2	w=0.3	w=0.1	w=0.4	w=0.5
				12	5.587	5.636	5.663	5.673	5.730
				100	w=0.8	w=0.6	w=0.7	w=0.9	w=0.5
				100	6.322	6.324	6.325	6.335	6.336

0	100	50	min01	12	w=0.4	w=0.3	w=0.5	w=0.6	w=0.7
				12	5.055	5.066	5.073	5.114	5.141
				100	w=0.5	w=0.4	w=0.6	w=0.7	w=0.3
				100	6.090	6.094	6.100	6.108	6.158
0	100	50	shrinkage	12	w=0.3	w=0.5	w=0.4	w=0.2	w=0.6
				12	5.576	5.577	5.585	5.591	5.608
				100	w=0.6	w=0.7	w=0.8	w=0.5	w=0.4
				100	6.478	6.481	6.496	6.496	6.500
0	100	50	uncorrelated	12	w=0.2	w=0.3	w=0.4	w=0.5	w=0.6
				12	5.482	5.484	5.503	5.537	5.548
				100	w=0.4	w=0.3	w=0.5	w=0.6	w=0.7
				100	6.453	6.468	6.473	6.487	6.513
0	100	200	block	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	2.980	3.680	4.032	4.240	4.382
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	4.068	4.578	4.755	4.876	4.966
0	100	200	blockdiag	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	3.036	4.026	4.374	4.590	4.731
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	4.231	5.144	5.424	5.574	5.671
0	100	200	min01	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	3.786	4.277	4.575	4.772	4.890
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	4.925	5.315	5.471	5.544	5.609
0	100	200	shrinkage	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	3.637	4.183	4.388	4.530	4.636
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	4.905	5.324	5.471	5.553	5.614
0	100	200	uncorrelated	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	3.526	4.087	4.337	4.514	4.639
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	4.456	5.103	5.317	5.439	5.520
0	100	1000	block	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	1.188	3.150	3.671	3.974	4.175
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	1.861	3.569	4.031	4.300	4.491
0	100	1000	blockdiag	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	1.075	3.340	3.855	4.139	4.335
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	1.803	3.817	4.292	4.565	4.735
0	100	1000	min01	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	1.230	3.165	3.683	3.979	4.182
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	1.918	3.618	4.080	4.342	4.510

0	100	1000	shrinkage	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	1.427	3.305	3.826	4.118	4.301
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	2.217	3.994	4.415	4.624	4.762
0	100	1000	uncorrelated	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	1.523	3.378	3.859	4.128	4.305
				100	sub	w=0.1	w=0.2	w=0.3	w=0.4
				100	2.370	4.065	4.469	4.694	4.842
0	1000	50	block	12	w=0.8	ℓ_2-3	w=0.7	rf-3	ℓ_2-1
				12	6.527	6.528	6.528	6.535	6.538
				1000	w=0.9	rf-4	rf-2	ℓ_2-4	all
				1000	7.323	7.339	7.343	7.352	7.356
0	1000	50	blockdiag	12	w=0.5	w=0.6	w=0.4	w=0.3	w=0.7
				12	6.547	6.555	6.558	6.561	6.561
				1000	w=0.4	w=0.6	w=0.5	w=0.7	w=0.8
				1000	7.016	7.020	7.023	7.029	7.033
0	1000	50	min01	12	w=0.6	w=0.7	ℓ_2-2	w=0.5	all
				12	6.339	6.347	6.349	6.350	6.351
				1000	w=0.9	w=0.7	w=0.8	ℓ_2-4	rf-2
				1000	7.086	7.088	7.092	7.092	7.098
0	1000	50	shrinkage	12	w=0.3	w=0.4	w=0.2	w=0.5	w=0.6
				12	6.618	6.624	6.632	6.647	6.659
				1000	sub	w=0.5	w=0.6	w=0.4	w=0.7
				1000	7.738	7.861	7.866	7.873	7.885
0	1000	50	uncorrelated	12	w=0.8	all	w=0.6	ℓ_2-3	w=0.7
				12	6.758	6.760	6.765	6.766	6.766
				1000	rf-3	rf-2	ℓ_2-2	all	ℓ_2-3
				1000	7.049	7.051	7.059	7.059	7.061
0	1000	200	block	12	w=0.1	w=0.2	w=0.3	w=0.4	w=0.5
				12	4.835	4.946	5.027	5.090	5.142
				1000	w=0.5	w=0.4	w=0.6	w=0.7	w=0.3
				1000	6.126	6.141	6.149	6.149	6.159
0	1000	200	blockdiag	12	w=0.1	w=0.2	sub	w=0.3	w=0.4
				12	4.685	4.822	4.851	4.936	5.028
				1000	sub	w=0.3	w=0.2	w=0.4	w=0.1
				1000	5.931	6.022	6.023	6.049	6.056
0	1000	200	min01	12	w=0.1	w=0.2	w=0.3	w=0.4	sub
				12	4.854	4.984	5.113	5.213	5.262
				1000	w=0.3	w=0.2	w=0.4	w=0.5	w=0.6
				1000	6.175	6.189	6.205	6.222	6.252
0	1000	200	shrinkage	12	w=0.2	w=0.1	w=0.3	w=0.4	w=0.5
				12	5.022	5.026	5.077	5.140	5.195
				1000	w=0.4	w=0.3	w=0.5	w=0.6	w=0.2
				1000	6.272	6.276	6.307	6.316	6.328

0	1000	200	uncorrelated	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	4.742	4.807	4.934	5.037	5.119
				1000	w=0.4	w=0.5	w=0.3	w=0.6	w=0.7
				1000	5.920	5.925	5.935	5.946	5.969
0	1000	1000	block	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	2.061	3.450	3.910	4.172	4.347
				1000	sub	w=0.1	w=0.2	w=0.3	w=0.4
				1000	3.267	4.365	4.696	4.882	4.984
0	1000	1000	blockdiag	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	2.013	3.728	4.186	4.435	4.600
				1000	sub	w=0.1	w=0.2	w=0.3	w=0.4
				1000	3.354	4.615	4.896	5.046	5.137
0	1000	1000	min01	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	2.315	3.737	4.162	4.423	4.590
				1000	sub	w=0.1	w=0.2	w=0.3	w=0.4
				1000	3.518	4.589	4.913	5.072	5.182
0	1000	1000	shrinkage	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	2.414	3.767	4.194	4.438	4.592
				1000	sub	w=0.1	w=0.2	w=0.3	w=0.4
				1000	3.928	4.888	5.185	5.327	5.408
0	1000	1000	uncorrelated	12	sub	w=0.1	w=0.2	w=0.3	w=0.4
				12	2.267	3.656	4.099	4.349	4.522
				1000	sub	w=0.1	w=0.2	w=0.3	w=0.4
				1000	3.411	4.579	4.914	5.106	5.212
1	12	50	uncorrelated	12	rf-4	rf-2	rf-1	rf-3	ℓ_2-2
				12	3.313	3.315	3.360	3.384	3.521
1	12	200	uncorrelated	12	sub	ℓ_2-3	ℓ_2-1	ℓ_2-2	ℓ_2-4
				12	1.580	2.392	2.419	2.420	2.425
1	12	1000	uncorrelated	12	sub	ℓ_2-3	ℓ_2-1	ℓ_2-2	ℓ_2-4
				12	0.693	2.632	2.638	2.640	2.646
1	100	50	block	12	rf-3	ℓ_1-3	ℓ_1-2	ℓ_1-1	ℓ_1-4
				12	6.149	6.153	6.164	6.169	6.171
				100	rf-3	sub	rf-2	rf-4	rf-1
				100	7.031	7.045	7.053	7.066	7.075
1	100	50	blockdiag	12	ℓ_1-3	ℓ_1-4	ℓ_1-1	ℓ_1-2	rf-4
				12	5.278	5.284	5.286	5.291	5.338
				100	ℓ_2-3	ℓ_2-2	ℓ_2-1	all	rf-3
				100	6.296	6.298	6.310	6.324	6.325
1	100	50	min01	12	all	ℓ_2-3	ℓ_2-1	w=0.9	ℓ_2-2
				12	6.074	6.083	6.095	6.096	6.105
				100	ℓ_2-3	w=0.8	ℓ_2-2	rf-3	w=0.6
				100	7.192	7.196	7.201	7.208	7.212
1	100	50	shrinkage	12	rf-3	ℓ_1-3	rf-2	ℓ_1-2	rf-1
				12	5.863	5.895	5.899	5.899	5.901
				100	ℓ_2-1	ℓ_2-2	w=0.7	w=0.8	w=0.9
				100	6.947	6.969	6.970	6.976	6.978

1	100	50	uncorrelated	12	ℓ_1-2	rf-3	ℓ_2-3	ℓ_1-3	ℓ_1-1
				12	6.294	6.295	6.302	6.310	6.311
				100	w=0.9	all	w=0.6	ℓ_2-2	w=0.8
				100	6.900	6.904	6.906	6.906	6.907
1	100	200	block	12	rf-4	sub	ℓ_1-1	ℓ_1-2	ℓ_1-4
				12	3.247	3.396	3.422	3.432	3.433
				100	sub	rf-4	ℓ_1-2	ℓ_1-1	ℓ_1-4
				100	4.380	4.490	4.668	4.673	4.685
1	100	200	blockdiag	12	rf-4	ℓ_1-2	ℓ_1-4	ℓ_1-1	rf-1
				12	2.475	2.706	2.715	2.715	2.909
				100	rf-4	ℓ_1-1	ℓ_1-2	ℓ_1-4	sub
				100	3.642	3.915	3.923	3.927	3.971
1	100	200	min01	12	rf-4	ℓ_1-2	ℓ_1-1	ℓ_1-4	sub
				12	3.049	3.262	3.268	3.279	3.327
				100	sub	rf-4	ℓ_1-1	ℓ_1-2	ℓ_1-4
				100	4.626	4.737	4.980	4.990	5.020
1	100	200	shrinkage	12	rf-4	ℓ_1-1	ℓ_1-4	ℓ_1-2	ℓ_1-3
				12	3.040	3.179	3.190	3.197	3.299
				100	rf-4	sub	ℓ_1-4	ℓ_1-2	ℓ_1-1
				100	4.815	4.907	4.970	4.988	5.000
1	100	200	uncorrelated	12	rf-4	ℓ_1-1	ℓ_1-2	ℓ_1-4	ℓ_1-3
				12	3.006	3.281	3.313	3.313	3.396
				100	rf-4	sub	ℓ_1-1	ℓ_1-2	ℓ_1-4
				100	4.117	4.356	4.392	4.393	4.403
1	100	1000	block	12	sub	ℓ_2-1	ℓ_2-2	ℓ_1-3	ℓ_2-4
				12	1.267	2.227	2.231	2.245	2.249
				100	sub	ℓ_1-3	ℓ_1-4	ℓ_1-1	ℓ_2-2
				100	1.951	3.265	3.276	3.281	3.283
1	100	1000	blockdiag	12	sub	ℓ_2-4	ℓ_2-1	ℓ_2-2	rf-1
				12	1.314	2.061	2.065	2.069	2.069
				100	sub	rf-2	rf-1	rf-3	ℓ_2-1
				100	2.109	2.877	2.880	2.893	2.921
1	100	1000	min01	12	sub	ℓ_2-1	ℓ_1-3	ℓ_2-2	ℓ_1-1
				12	1.294	2.470	2.471	2.471	2.474
				100	sub	ℓ_1-1	ℓ_1-2	ℓ_1-4	ℓ_1-3
				100	1.980	3.270	3.275	3.279	3.304
1	100	1000	shrinkage	12	sub	ℓ_2-2	ℓ_2-1	ℓ_2-4	ℓ_1-3
				12	1.423	2.335	2.338	2.340	2.350
				100	sub	ℓ_1-2	ℓ_1-3	ℓ_1-1	ℓ_1-4
				100	2.234	3.406	3.407	3.407	3.408
1	100	1000	uncorrelated	12	sub	ℓ_2-4	ℓ_2-1	ℓ_2-2	rf-2
				12	1.615	2.113	2.116	2.118	2.135
				100	sub	rf-2	rf-1	ℓ_2-2	rf-3
				100	2.398	3.052	3.057	3.065	3.072

1	1000	50	block	12	all	w=0.8	w=0.9	w=0.7	w=0.6
				12	6.186	6.206	6.207	6.214	6.236
				1000	w=0.9	w=0.8	all	w=0.7	w=0.6
				1000	6.814	6.825	6.839	6.844	6.880
1	1000	50	blockdiag	12	all	w=0.8	rf-3	w=0.9	w=0.7
				12	6.473	6.482	6.484	6.490	6.490
				1000	w=0.9	w=0.8	all	w=0.7	w=0.6
				1000	7.129	7.150	7.160	7.160	7.161
1	1000	50	min01	12	w=0.8	w=0.7	all	w=0.6	ℓ_2-2
				12	6.630	6.641	6.643	6.643	6.644
				1000	w=0.9	ℓ_2-4	ℓ_2-1	all	w=0.8
				1000	7.377	7.432	7.440	7.446	7.448
1	1000	50	shrinkage	12	rf-3	w=0.9	w=0.8	w=0.7	all
				12	6.665	6.672	6.678	6.678	6.680
				1000	rf-2	ℓ_2-3	w=0.5	rf-3	all
				1000	7.575	7.597	7.598	7.599	7.600
1	1000	50	uncorrelated	12	w=0.9	w=0.8	all	w=0.7	w=0.6
				12	6.432	6.438	6.439	6.446	6.455
				1000	all	w=0.8	w=0.7	w=0.9	ℓ_2-2
				1000	6.827	6.860	6.861	6.862	6.875
1	1000	200	block	12	sub	rf-3	ℓ_1-1	ℓ_2-3	ℓ_1-4
				12	4.874	5.048	5.058	5.059	5.067
				1000	sub	ℓ_1-2	ℓ_1-4	ℓ_1-3	ℓ_1-1
				1000	6.097	6.772	6.791	6.797	6.805
1	1000	200	blockdiag	12	ℓ_1-1	ℓ_1-2	ℓ_1-4	ℓ_1-3	sub
				12	5.297	5.301	5.304	5.311	5.321
				1000	sub	ℓ_1-4	ℓ_1-2	ℓ_1-1	ℓ_1-3
				1000	6.281	6.757	6.767	6.779	6.782
1	1000	200	min01	12	sub	ℓ_1-2	ℓ_1-4	ℓ_1-1	ℓ_1-3
				12	4.643	4.890	4.890	4.893	4.896
				1000	sub	ℓ_1-4	ℓ_1-2	ℓ_1-1	ℓ_1-3
				1000	5.879	6.566	6.568	6.577	6.583
1	1000	200	shrinkage	12	rf-3	rf-4	sub	rf-1	rf-2
				12	5.228	5.269	5.275	5.284	5.286
				1000	sub	w=0.9	w=0.8	w=0.7	w=0.6
				1000	6.769	7.177	7.190	7.195	7.206
1	1000	200	uncorrelated	12	sub	ℓ_1-2	ℓ_1-4	ℓ_1-1	ℓ_1-3
				12	4.910	5.101	5.103	5.104	5.122
				1000	sub	ℓ_1-4	ℓ_1-2	ℓ_1-3	ℓ_1-1
				1000	6.019	6.325	6.334	6.337	6.339
1	1000	1000	block	12	ℓ_1-4	ℓ_1-2	ℓ_1-1	sub	ℓ_1-3
				12	2.203	2.206	2.206	2.243	2.244
				1000	sub	ℓ_1-1	ℓ_1-2	ℓ_1-4	ℓ_1-3
				1000	3.413	3.916	3.924	3.938	3.989

1	1000	1000	blockdiag	12	ℓ_1-1	ℓ_1-2	ℓ_1-4	ℓ_1-3	sub
				12	1.866	1.868	1.868	1.873	2.059
				1000	sub	ℓ_1-2	ℓ_1-1	ℓ_1-4	ℓ_1-3
				1000	3.223	3.317	3.320	3.323	3.327
1	1000	1000	min01	12	sub	ℓ_1-1	ℓ_1-2	ℓ_1-4	ℓ_1-3
				12	1.954	2.116	2.116	2.117	2.130
				1000	sub	ℓ_1-1	ℓ_1-4	ℓ_1-2	ℓ_1-3
				1000	3.252	3.689	3.704	3.723	3.754
1	1000	1000	shrinkage	12	ℓ_1-2	ℓ_1-1	ℓ_1-4	ℓ_1-3	sub
				12	1.970	1.974	1.976	2.075	2.458
				1000	sub	ℓ_1-2	ℓ_1-1	ℓ_1-4	ℓ_1-3
				1000	3.768	4.342	4.351	4.371	4.616
1	1000	1000	uncorrelated	12	ℓ_1-4	ℓ_1-2	ℓ_1-1	ℓ_1-3	sub
				12	1.749	1.751	1.751	1.770	2.093
				1000	ℓ_1-4	ℓ_1-1	ℓ_1-2	ℓ_1-3	sub
				1000	3.137	3.143	3.161	3.185	3.239

TABLE C.3: *Top five combinations of Cox models and parameters for weights estimation in terms of smallest mean Manhattan distance between estimated and true β_j, $j = 1, ..., \tilde{p}$ (mean across all training sets and all subgroups). rf = random forest, ℓ_1 = lasso, ℓ_2 = ridge; and 1 = no intera. & no cumHR, 2 = no intera. & cumHR, 3 = intera. & no cumHR, 4 = intera. & cumHR.*

				Top five methods				
ϵ	p	n	$\boldsymbol{\Sigma}$	1	2	3	4	5
0	12	50	uncorrelated	w=0.1	w=0.2	w=0.3	w=0.4	w=0.5
				0.778	0.777	0.768	0.755	0.748
0	12	200	uncorrelated	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.880	0.872	0.858	0.844	0.832
0	12	1000	uncorrelated	sub	w=0.1	w=0.2	ℓ_1-4	ℓ_2-4
				0.871	0.861	0.848	0.842	0.840
0	100	50	block	ℓ_2-3	w=0.9	w=0.8	rf-4	ℓ_1-1
				0.745	0.745	0.745	0.745	0.745
0	100	50	blockdiag	w=0.4	w=0.6	w=0.5	w=0.7	w=0.8
				0.764	0.763	0.763	0.762	0.762
0	100	50	min01	w=0.5	w=0.8	w=0.7	w=0.6	all
				0.776	0.775	0.775	0.775	0.773
0	100	50	shrinkage	w=0.5	w=0.3	w=0.4	w=0.6	w=0.2
				0.779	0.778	0.778	0.777	0.775
0	100	50	uncorrelated	w=0.4	w=0.5	w=0.3	w=0.6	w=0.7
				0.754	0.751	0.750	0.750	0.745
0	100	200	block	w=0.1	sub	w=0.2	w=0.3	w=0.4
				0.864	0.864	0.859	0.854	0.850
0	100	200	blockdiag	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.855	0.852	0.843	0.832	0.822

0	100	200	min01	w=0.1	w=0.2	sub	w=0.3	w=0.4
				0.821	0.817	0.814	0.812	0.808
0	100	200	shrinkage	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.842	0.841	0.835	0.829	0.825
0	100	200	uncorrelated	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.831	0.827	0.822	0.815	0.809
0	100	1000	block	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.877	0.869	0.858	0.847	0.839
0	100	1000	blockdiag	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.878	0.866	0.852	0.840	0.830
0	100	1000	min01	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.872	0.864	0.853	0.843	0.835
0	100	1000	shrinkage	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.871	0.861	0.851	0.842	0.834
0	100	1000	uncorrelated	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.860	0.851	0.840	0.830	0.822
0	1000	50	block	ℓ_2-3	ℓ_1-3	ℓ_2-1	ℓ_1-2	rf-1
				0.649	0.647	0.647	0.646	0.646
0	1000	50	blockdiag	w=0.7	w=0.5	w=0.6	w=0.9	w=0.8
				0.636	0.632	0.632	0.630	0.629
0	1000	50	min01	w=0.9	ℓ_2-1	ℓ_1-3	rf-1	rf-2
				0.706	0.706	0.705	0.704	0.704
0	1000	50	shrinkage	w=0.4	w=0.5	w=0.3	w=0.6	w=0.7
				0.696	0.694	0.692	0.691	0.689
0	1000	50	uncorrelated	all	w=0.8	w=0.6	w=0.9	rf-4
				0.566	0.566	0.563	0.563	0.562
0	1000	200	block	w=0.2	w=0.3	w=0.4	w=0.1	w=0.5
				0.823	0.822	0.821	0.821	0.819
0	1000	200	blockdiag	w=0.1	w=0.2	w=0.3	w=0.4	w=0.5
				0.838	0.837	0.832	0.825	0.818
0	1000	200	min01	w=0.2	w=0.1	w=0.3	w=0.4	w=0.5
				0.810	0.809	0.807	0.802	0.798
0	1000	200	shrinkage	w=0.2	w=0.1	w=0.3	w=0.4	w=0.5
				0.817	0.815	0.815	0.811	0.807
0	1000	200	uncorrelated	w=0.1	w=0.2	w=0.3	w=0.4	sub
				0.827	0.823	0.818	0.813	0.811
0	1000	1000	block	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.875	0.869	0.858	0.849	0.842
0	1000	1000	blockdiag	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.875	0.865	0.852	0.839	0.829
0	1000	1000	min01	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.860	0.853	0.842	0.831	0.823
0	1000	1000	shrinkage	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.871	0.863	0.853	0.844	0.836
0	1000	1000	uncorrelated	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.867	0.859	0.848	0.837	0.828

1	12	50	uncorrelated	rf-2	rf-4	rf-1	rf-3	ℓ_1-4
				0.823	0.820	0.820	0.816	0.812
1	12	200	uncorrelated	ℓ_1-1	ℓ_1-2	ℓ_1-4	rf-2	rf-1
				0.870	0.869	0.868	0.868	0.868
1	12	1000	uncorrelated	ℓ_1-1	ℓ_1-2	rf-2	rf-1	rf-4
				0.871	0.871	0.871	0.871	0.871
1	100	50	block	ℓ_2-2	ℓ_2-3	ℓ_2-1	w=0.9	w=0.7
				0.692	0.690	0.690	0.689	0.689
1	100	50	blockdiag	ℓ_2-4	ℓ_2-2	ℓ_2-1	ℓ_1-3	ℓ_1-4
				0.760	0.760	0.759	0.758	0.758
1	100	50	min01	w=0.9	all	w=0.8	w=0.7	ℓ_2-1
				0.666	0.665	0.664	0.663	0.660
1	100	50	shrinkage	w=0.4	w=0.3	w=0.5	ℓ_2-3	rf-3
				0.749	0.749	0.746	0.745	0.745
1	100	50	uncorrelated	ℓ_2-3	rf-3	ℓ_2-2	ℓ_2-4	ℓ_1-3
				0.708	0.708	0.704	0.704	0.704
1	100	200	block	sub	rf-4	ℓ_1-1	ℓ_1-4	ℓ_1-2
				0.843	0.838	0.837	0.837	0.836
1	100	200	blockdiag	sub	rf-4	ℓ_1-1	ℓ_1-2	ℓ_1-4
				0.862	0.856	0.851	0.850	0.850
1	100	200	min01	sub	rf-4	ℓ_1-1	ℓ_1-2	ℓ_1-4
				0.838	0.837	0.833	0.833	0.833
1	100	200	shrinkage	sub	rf-4	ℓ_1-4	ℓ_1-2	ℓ_1-1
				0.825	0.825	0.822	0.821	0.821
1	100	200	uncorrelated	sub	rf-4	ℓ_1-1	ℓ_1-2	ℓ_1-4
				0.831	0.823	0.816	0.815	0.815
1	100	1000	block	ℓ_1-2	ℓ_1-1	ℓ_1-4	sub	ℓ_1-3
				0.876	0.876	0.875	0.875	0.875
1	100	1000	blockdiag	ℓ_1-2	ℓ_1-4	ℓ_1-1	sub	ℓ_1-3
				0.873	0.873	0.873	0.872	0.872
1	100	1000	min01	ℓ_1-1	ℓ_1-2	sub	ℓ_1-4	ℓ_1-3
				0.870	0.870	0.869	0.869	0.869
1	100	1000	shrinkage	sub	ℓ_1-2	ℓ_1-1	ℓ_1-4	ℓ_1-3
				0.877	0.877	0.877	0.877	0.876
1	100	1000	uncorrelated	ℓ_1-2	ℓ_1-1	sub	ℓ_1-4	rf-4
				0.862	0.862	0.862	0.862	0.862
1	1000	50	block	all	w=0.7	w=0.9	w=0.8	rf-3
				0.716	0.715	0.714	0.714	0.713
1	1000	50	blockdiag	w=0.7	w=0.6	w=0.8	all	w=0.5
				0.665	0.665	0.664	0.663	0.662
1	1000	50	min01	all	w=0.9	w=0.8	w=0.7	ℓ_2-1
				0.622	0.618	0.616	0.615	0.611
1	1000	50	shrinkage	rf-3	rf-1	rf-2	rf-4	all
				0.688	0.687	0.686	0.686	0.684
1	1000	50	uncorrelated	all	w=0.8	w=0.9	w=0.7	w=0.6
				0.669	0.668	0.667	0.664	0.664

1	1000	200	block	sub	ℓ_2-3	ℓ_2-1	ℓ_2-2	rf-3
				0.811	0.789	0.787	0.787	0.787
1	1000	200	blockdiag	sub	ℓ_1-4	ℓ_1-2	ℓ_1-1	ℓ_1-3
				0.786	0.767	0.767	0.767	0.766
1	1000	200	min01	sub	ℓ_1-2	ℓ_1-1	ℓ_1-3	ℓ_1-4
				0.822	0.786	0.786	0.786	0.786
1	1000	200	shrinkage	sub	rf-3	ℓ_2-3	rf-1	rf-2
				0.804	0.784	0.783	0.782	0.782
1	1000	200	uncorrelated	sub	ℓ_1-4	ℓ_1-2	ℓ_1-1	ℓ_1-3
				0.809	0.779	0.779	0.779	0.778
1	1000	1000	block	sub	ℓ_1-1	ℓ_1-2	ℓ_1-4	ℓ_1-3
				0.869	0.865	0.865	0.865	0.864
1	1000	1000	blockdiag	sub	ℓ_1-2	ℓ_1-4	ℓ_1-1	ℓ_1-3
				0.873	0.871	0.871	0.871	0.871
1	1000	1000	min01	sub	ℓ_1-1	ℓ_1-2	ℓ_1-4	ℓ_1-3
				0.872	0.866	0.866	0.866	0.865
1	1000	1000	shrinkage	sub	ℓ_1-2	ℓ_1-1	ℓ_1-4	ℓ_1-3
				0.867	0.861	0.861	0.861	0.860
1	1000	1000	uncorrelated	sub	ℓ_1-1	ℓ_1-4	ℓ_1-2	ℓ_1-3
				0.875	0.870	0.870	0.870	0.870

TABLE C.4: *Top five combinations of Cox models and parameters for weights estimation in terms of highest mean C-index (mean across all test sets and subgroups). rf = random forest, ℓ_1 = lasso, ℓ_2 = ridge; and 1 = no intera. & no cumHR, 2 = no intera. & cumHR, 3 = intera. & no cumHR, 4 = intera. & cumHR.*

				Top five methods				
ϵ	p	n	$\boldsymbol{\Sigma}$	1	2	3	4	5
0	12	50	uncorrelated	w=0.3	w=0.2	w=0.4	w=0.5	w=0.6
				0.091	0.091	0.092	0.094	0.095
0	12	200	uncorrelated	sub	w=0.2	w=0.3	w=0.1	w=0.4
				0.079	0.087	0.089	0.090	0.091
0	12	1000	uncorrelated	sub	w=0.1	w=0.2	ℓ_1-4	ℓ_2-4
				0.080	0.092	0.093	0.095	0.095
0	100	50	block	w=0.6	w=0.5	w=0.8	all	w=0.7
				0.099	0.099	0.099	0.099	0.100
0	100	50	blockdiag	w=0.8	w=0.6	w=0.7	ℓ_1-2	w=0.4
				0.125	0.125	0.125	0.126	0.126
0	100	50	min01	w=0.5	w=0.4	w=0.7	w=0.6	w=0.8
				0.115	0.115	0.115	0.115	0.115
0	100	50	shrinkage	w=0.7	w=0.5	w=0.6	w=0.3	w=0.4
				0.095	0.095	0.095	0.095	0.095
0	100	50	uncorrelated	ℓ_1-1	w=0.9	ℓ_2-1	ℓ_2-4	w=0.7
				0.105	0.105	0.105	0.105	0.105
0	100	200	block	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.081	0.090	0.091	0.093	0.095

0	100	200	blockdiag	sub	w=0.2	w=0.1	w=0.3	w=0.4
				0.080	0.088	0.089	0.090	0.093
0	100	200	min01	w=0.1	w=0.2	w=0.3	w=0.4	sub
				0.092	0.093	0.095	0.097	0.098
0	100	200	shrinkage	sub	w=0.2	w=0.1	w=0.3	w=0.4
				0.086	0.098	0.099	0.099	0.100
0	100	200	uncorrelated	sub	w=0.2	w=0.3	w=0.1	w=0.4
				0.065	0.077	0.078	0.078	0.080
0	100	1000	block	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.057	0.081	0.081	0.085	0.088
0	100	1000	blockdiag	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.067	0.082	0.082	0.085	0.088
0	100	1000	min01	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.073	0.083	0.084	0.087	0.089
0	100	1000	shrinkage	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.067	0.083	0.084	0.086	0.089
0	100	1000	uncorrelated	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.077	0.082	0.084	0.087	0.090
0	1000	50	block	ℓ_2-3	ℓ_2-1	ℓ_1-2	ℓ_2-4	rf-3
				0.139	0.139	0.139	0.139	0.140
0	1000	50	blockdiag	w=0.6	w=0.5	w=0.7	w=0.4	w=0.3
				0.129	0.130	0.130	0.130	0.131
0	1000	50	min01	all	ℓ_2-2	w=0.8	w=0.7	w=0.9
				0.112	0.112	0.112	0.112	0.112
0	1000	50	shrinkage	w=0.4	w=0.3	w=0.5	w=0.7	w=0.6
				0.167	0.167	0.168	0.169	0.169
0	1000	50	uncorrelated	rf-2	all	ℓ_2-3	w=0.9	ℓ_2-1
				0.142	0.142	0.142	0.142	0.143
0	1000	200	block	w=0.6	w=0.5	w=0.7	w=0.4	w=0.8
				0.066	0.066	0.067	0.067	0.067
0	1000	200	blockdiag	w=0.1	w=0.2	w=0.3	w=0.4	w=0.5
				0.092	0.092	0.094	0.096	0.098
0	1000	200	min01	w=0.3	w=0.4	w=0.2	w=0.5	w=0.6
				0.107	0.107	0.107	0.108	0.109
0	1000	200	shrinkage	w=0.4	w=0.5	w=0.6	w=0.3	w=0.7
				0.078	0.078	0.078	0.079	0.079
0	1000	200	uncorrelated	w=0.1	w=0.2	w=0.3	w=0.4	w=0.5
				0.096	0.097	0.098	0.099	0.101
0	1000	1000	block	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.078	0.085	0.085	0.087	0.089
0	1000	1000	blockdiag	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.072	0.080	0.082	0.085	0.088
0	1000	1000	min01	sub	w=0.2	w=0.1	w=0.3	w=0.4
				0.065	0.076	0.077	0.078	0.079
0	1000	1000	shrinkage	sub	w=0.1	w=0.2	w=0.3	w=0.4
				0.069	0.080	0.083	0.086	0.089

0	1000	1000	uncorrelated	sub 0.080	w=0.1 0.088	w=0.2 0.090	w=0.3 0.094	w=0.4 0.097
1	12	50	uncorrelated	ℓ_1-2 0.093	ℓ_1-4 0.094	ℓ_1-1 0.094	ℓ_1-3 0.095	ℓ_2-2 0.095
1	12	200	uncorrelated	ℓ_1-1 0.074	ℓ_1-2 0.074	ℓ_1-4 0.075	rf-2 0.075	rf-1 0.075
1	12	1000	uncorrelated	sub 0.083	ℓ_1-2 0.087	ℓ_1-1 0.087	ℓ_1-4 0.088	ℓ_1-3 0.088
1	100	50	block	ℓ_2-3 0.136	ℓ_2-1 0.136	ℓ_2-2 0.136	w=0.9 0.137	w=0.8 0.137
1	100	50	blockdiag	ℓ_2-1 0.119	ℓ_2-3 0.119	ℓ_2-4 0.119	ℓ_2-2 0.119	ℓ_1-4 0.119
1	100	50	min01	w=0.9 0.149	ℓ_2-1 0.150	ℓ_2-2 0.150	all 0.150	w=0.8 0.151
1	100	50	shrinkage	w=0.5 0.112	w=0.4 0.112	w=0.6 0.112	w=0.7 0.113	ℓ_2-3 0.113
1	100	50	uncorrelated	rf-3 0.113	ℓ_2-4 0.113	ℓ_2-1 0.113	ℓ_2-3 0.113	ℓ_2-2 0.113
1	100	200	block	sub 0.081	ℓ_2-2 0.083	ℓ_2-4 0.084	ℓ_2-1 0.084	w=0.5 0.084
1	100	200	blockdiag	sub 0.075	rf-4 0.088	ℓ_1-1 0.090	ℓ_1-4 0.090	rf-2 0.090
1	100	200	min01	w=0.5 0.090	w=0.4 0.090	w=0.6 0.090	sub 0.090	w=0.7 0.090
1	100	200	shrinkage	rf-4 0.075	rf-2 0.077	rf-1 0.077	ℓ_1-4 0.077	rf-3 0.077
1	100	200	uncorrelated	sub 0.098	rf-4 0.116	rf-3 0.116	ℓ_1-3 0.116	ℓ_2-3 0.117
1	100	1000	block	sub 0.069	w=0.1 0.079	w=0.2 0.080	w=0.3 0.081	w=0.4 0.082
1	100	1000	blockdiag	sub 0.068	ℓ_1-2 0.079	ℓ_1-1 0.079	ℓ_1-4 0.079	ℓ_1-3 0.080
1	100	1000	min01	sub 0.060	ℓ_1-2 0.072	ℓ_1-1 0.072	ℓ_1-4 0.072	rf-4 0.073
1	100	1000	shrinkage	sub 0.074	ℓ_1-2 0.091	ℓ_1-1 0.091	ℓ_1-4 0.091	ℓ_1-3 0.092
1	100	1000	uncorrelated	sub 0.068	ℓ_1-2 0.083	ℓ_1-1 0.083	ℓ_1-4 0.083	rf-4 0.083
1	1000	50	block	all 0.130	w=0.9 0.131	w=0.8 0.131	w=0.7 0.132	ℓ_2-3 0.132
1	1000	50	blockdiag	all 0.109	ℓ_2-3 0.109	w=0.9 0.109	rf-3 0.109	w=0.8 0.109
1	1000	50	min01	w=0.7 0.175	all 0.176	ℓ_2-1 0.176	ℓ_2-2 0.176	w=0.8 0.176
1	1000	50	shrinkage	rf-3 0.108	w=0.9 0.108	ℓ_2-1 0.109	all 0.109	ℓ_2-2 0.109

1	1000	50	uncorrelated	all	w=0.9	w=0.8	w=0.7	ℓ_2-3
				0.162	0.162	0.163	0.163	0.164
1	1000	200	block	sub	ℓ_2-3	all	w=0.9	w=0.8
				0.090	0.092	0.092	0.092	0.092
1	1000	200	blockdiag	sub	ℓ_2-1	ℓ_2-2	rf-3	ℓ_2-4
				0.109	0.123	0.123	0.123	0.123
1	1000	200	min01	sub	w=0.6	w=0.8	w=0.7	ℓ_2-1
				0.097	0.104	0.104	0.104	0.104
1	1000	200	shrinkage	all	w=0.9	w=0.8	w=0.7	ℓ_2-1
				0.094	0.094	0.095	0.095	0.095
1	1000	200	uncorrelated	sub	ℓ_1-4	ℓ_1-3	ℓ_1-1	ℓ_1-2
				0.096	0.114	0.115	0.115	0.115
1	1000	1000	block	sub	rf-4	rf-2	rf-1	ℓ_2-2
				0.070	0.086	0.087	0.087	0.087
1	1000	1000	blockdiag	sub	ℓ_1-1	ℓ_1-2	ℓ_1-4	ℓ_1-3
				0.058	0.075	0.075	0.075	0.076
1	1000	1000	min01	sub	rf-1	rf-2	rf-3	ℓ_2-4
				0.065	0.084	0.084	0.084	0.084
1	1000	1000	shrinkage	sub	rf-1	rf-2	ℓ_2-4	rf-3
				0.078	0.087	0.087	0.087	0.088
1	1000	1000	uncorrelated	sub	rf-1	rf-2	rf-3	ℓ_2-4
				0.077	0.101	0.101	0.101	0.101

TABLE C.5: *Top five combinations of Cox models and parameters for weights estimation in terms of smallest mean integrated Brier score (IBS) (mean across all test sets and subgroups). rf = random forest, ℓ_1 = lasso, ℓ_2 = ridge; and 1 = no intera. & no cumHR, 2 = no intera. & cumHR, 3 = intera. & no cumHR, 4 = intera. & cumHR.*

Probe set	Gene	s	w0.1	w0.2	w0.3	w0.4	w0.5	w0.6	w0.7	w0.8	w0.9	all	sub	lasso	ridge	rf
1554883_a_at	ERCC8	3	0	0	0	0	1	1	0	0	0	0	0	0	0	0
1562507_at		1	0	1	1	1	1	1	0	0	0	0	0	0	0	0
1568574_x_at	SPP1	1,2,3,4	0	1	1	2	2	2	3	4	4	0	0	0	0	0
210491_at		2	1	1	1	0	0	0	0	0	0	0	0	0	0	0
214435_x_at	RALA	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
221288_at	GPR22	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0
221626_at	ZNF506	1,2,3,4	0	0	1	1	2	2	3	3	4	0	0	0	0	0
228662_at	SOCS7	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0
229118_at	PRRG3	3	0	0	1	1	1	1	1	0	0	0	0	0	0	0
229384_at		2	0	0	0	1	0	1	1	1	0	0	0	0	0	0

TABLE C.6: *Cox models including all genes and mandatory clinical covariates. For each subgroup s genes with a mean inclusion frequency larger than 0.4 in any model are selected. Values correspond to the number of selections of each gene in each Cox model across all four subgroups. Subgroups: $s = 1$: GSE29013; $s = 2$: GSE31210; $s = 3$: GSE37745; $s = 4$: GSE50081.*

Probe set	Gene	s	w0.1	w0.2	w0.3	w0.4	w0.5	w0.6	w0.7	w0.8	w0.9	all	sub	lasso	ridge	rf
1556854_at	ATP11A-AS1	1,2,3,4	0	2	3	4	4	4	4	4	4	4	0	0	0	0
1560683_at	NBEAP1	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1562048_at	LINC02085	2	0	0	0	0	0	0	0	0	0	0	0	0	1	1
1562507_at		1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
1568574_x_at	SPP1	1,2,3,4	0	2	2	3	4	4	4	4	4	4	0	0	0	0
203844_at	VHL	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
206943_at	TGFBR1	2	0	1	1	0	0	0	0	0	0	0	1	0	0	0
210491_at		2	0	1	1	0	0	0	0	0	0	0	0	0	0	0
211909_x_at	PTGER3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	1
214435_x_at	RALA	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
217761_at	ADI1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
218573_at	MAGEH1	1,2,3,4	0	1	1	1	1	2	2	2	4	4	0	0	0	0
220756_s_at	SLC52A1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0

221288_at	GPR22	2	1	1	1	1	0	0	0	0	0	0	0	0	0
221626_at	ZNF506	1,2,3,4	0	2	2	3	3	4	4	4	4	4	0	0	0	0
224714_at	NIFK	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0
227755_at	ERN1	1,2,3,4	0	0	1	1	1	3	3	3	4	4	0	0	0	0
230012_at	LINC00324	2	0	0	0	1	1	1	1	0	0	0	0	0	0	0
231489_x_at	CREG2	3	0	0	1	1	0	0	0	0	0	0	0	0	0	0
231938_at	SGO1	3	0	1	1	1	1	1	0	0	0	0	0	0	0	0
235559_at		4	1	1	1	0	0	0	0	0	0	0	0	0	0	0
236629_at	IBA57	2	0	0	1	1	1	1	1	1	1	0	0	0	0	0
237390_at		2	0	0	0	0	0	0	0	0	0	0	0	1	1	1
237853_x_at		3	0	0	0	1	1	0	0	0	0	0	0	0	0	0
239658_at		2	0	0	0	0	0	0	0	0	0	0	1	0	0	0
241616_at		2	0	1	0	0	0	0	0	0	0	0	0	0	0	0

TABLE C.7: *Cox models including all genes. For each subgroup s genes with a mean inclusion frequency larger than 0.4 in any model are selected. Values correspond to the number of selections of each gene in each Cox model across all four subgroups. Subgroups: $s = 1$: GSE29013; $s = 2$: GSE31210; $s = 3$: GSE37745; $s = 4$: GSE50081.*

Probe set	Gene	*s*	w0.1	w0.2	w0.3	w0.4	w0.5	w0.6	w0.7	w0.8	w0.9	all	sub	lasso	ridge	rf
1554316_at	PGLS	2	0	1	1	1	0	0	0	0	0	0	0	0	0	0
1566342_at	SOD2	1,2,3,4	0	1	2	3	4	4	4	4	4	0	0	0	0	0
1568574_x_at	SPP1	1,2,3,4	1	2	2	4	4	4	4	4	4	0	0	0	0	0
210397_at	DEFB1	1,2,3,4	0	0	0	1	2	3	3	3	4	0	0	0	0	0
210663_s_at	KYNU	1,3	0	0	1	1	1	1	1	2	0	0	0	0	0	0
210794_s_at	MEG3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1
211095_at	NF1	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0
211851_x_at	BRCA1	1,4	0	0	0	1	1	1	1	1	2	0	0	0	0	0
213201_s_at	TNNT1	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0
214435_x_at	RALA	1,2,4	1	1	1	1	1	1	2	3	3	0	0	0	0	1
219927_at	FCF1	1,2,3,4	0	0	0	1	2	3	4	4	4	0	0	0	0	0
220655_at	TNIP3	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0

221234_s_at	BACH2	2	0	1	1	1	0	0	0	0	0	0	0	0	0
227755_at	ERN1	1,2,3,4	0	1	2	3	3	3	4	4	4	0	0	0	0
228722_at	PRMT2	1,2,4	0	0	1	2	2	3	2	3	3	0	0	0	0
229997_at	VANGL1	2	0	1	1	1	1	1	0	0	0	0	0	0	0
231926_at	EPS15L1	1,2,3,4	0	0	1	3	3	3	4	4	4	0	0	0	0
234474_x_at	IL6ST	1	1	0	0	0	0	0	0	0	0	0	0	0	0
238117_at	PPOX	3	0	0	0	0	0	1	0	0	0	0	0	0	0
238483_at	SSBP2	2	1	1	0	0	0	0	0	0	0	0	0	0	0

TABLE C.8: *Cox models including the prognostic genes and mandatory clinical covariates. For each subgroup s genes with a mean inclusion frequency larger than 0.4 in any model are selected. Values correspond to the number of selections of each gene in each Cox model across all four subgroups. Subgroups: $s = 1$: GSE29013; $s = 2$: GSE31210; $s = 3$: GSE37745; $s = 4$: GSE50081.*

Probe set	Gene	s	w0.1	w0.2	w0.3	w0.4	w0.5	w0.6	w0.7	w0.8	w0.9	all	sub	lasso	ridge	rf
1554316_at	PGLS	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0
1557836_at	ELMOD2	2	1	0	0	0	0	0	0	0	0	0	0	0	1	1
1563881_at	MAGI1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1566342_at	SOD2	1,2,3,4	0	1	3	3	4	4	4	4	4	4	0	0	0	0
1568574_x_at	SPP1	1,2,3,4	2	2	4	4	4	4	4	4	4	4	0	0	0	0
203249_at	EZH1	2	0	0	0	1	1	1	0	0	0	0	0	0	0	0
204498_s_at	ADCY9	2	1	1	1	0	0	0	0	0	0	0	0	1	1	1
206018_at	FOXG1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
206529_x_at	SLC26A4	2	0	0	0	0	0	0	0	0	0	0	0	0	1	1
209315_at	HBS1L	2	0	1	1	1	1	1	0	0	0	0	0	0	0	0
209828_s_at	IL16	1,2,3,4	0	0	1	2	2	2	2	2	4	4	0	0	0	0
210397_at	DEFB1	3	0	1	1	1	1	1	1	0	1	0	0	0	0	0
210662_at	KYNU	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
211095_at	NF1	2	0	0	0	0	0	0	0	0	0	0	1	1	1	1
211851_x_at	BRCA1	1,2,3,4	0	0	2	2	2	2	4	4	4	4	0	0	0	0
213201_s_at	TNNT1	4	1	1	1	1	1	1	0	0	0	0	0	0	0	0

214435_x_at	RALA	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
214557_at	PTTG2	4	0	0	1	1	0	0	0	0	0	0	0	0	0	0
215384_s_at	MAP1A	2	1	0	0	0	0	0	0	0	0	0	0	1	1	1
215717_s_at	FBN2	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0
219927_at	FCF1	1,2,3,4	0	0	0	1	2	3	4	4	4	4	0	0	0	0
220655_at	TNIP3	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0
222043_at	CLU	4	0	1	1	0	0	0	0	0	0	0	0	0	0	0
226145_s_at	FRAS1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0
227755_at	ERN1	1,2,3,4	1	2	3	4	4	4	4	4	4	4	0	0	0	0
228722_at	PRMT2	1,2,3,4	0	0	1	3	3	3	3	3	3	4	0	0	0	0
229895_s_at	NCK1	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0
229997_at	VANGL1	1,2,3,4	1	1	2	2	2	4	4	4	4	4	0	0	0	0
231926_at	EPS15L1	2	0	0	1	1	1	1	0	1	0	0	0	0	0	0
234474_x_at	IL6ST	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
235404_at	ARID5B	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0
238483_at	SSBP2	2	1	1	1	1	0	0	0	0	0	0	0	1	1	1
239219_at	AURKB	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0
56829_at	TRAPPC9	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0

TABLE C.9: *Cox models including the prognostic genes. For each subgroup s genes with a mean inclusion frequency larger than 0.4 in any model are selected. Values correspond to the number of selections of each gene in each Cox model across all four subgroups. Subgroups: $s = 1$: GSE29013; $s = 2$: GSE31210; $s = 3$: GSE37745; $s = 4$: GSE50081.*

Probe set	Gene	*s*	w0.1	w0.2	w0.3	w0.4	w0.5	w0.6	w0.7	w0.8	w0.9	all	sub	lasso	ridge	rf
1555758_a_at	CDKN3	1,2,3,4	0	1	2	2	2	2	3	4	4	0	0	0	0	0
202376_at	SERPINA3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
202619_s_at	PLOD2	2	0	1	1	1	1	1	1	0	0	0	0	0	0	0
203559_s_at	AOC1	1,2,3,4	1	3	3	4	4	4	4	4	4	4	0	0	1	0
203764_at	DLGAP5	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
204260_at	CHGB	2	1	1	1	1	1	1	1	0	0	0	0	0	1	0
204532_x_at		2	1	1	1	0	0	0	0	0	0	0	0	0	0	0

205476_at	CCL20	2	1	1	1	0	0	0	0	0	0	0	0	0	0
206224_at	CST1	3	0	0	0	0	0	0	0	0	0	0	1	0	0
206262_at	ADH1C	2	1	1	1	0	0	0	0	0	0	0	0	1	0
208960_s_at	KLF6	4	0	0	0	0	0	1	0	0	0	0	0	0	0
209875_s_at	SPP1	1,2,3,4	2	2	2	3	3	4	4	4	4	4	0	0	1	0
210397_at	DEFB1	1,2,3,4	1	2	3	3	4	4	4	4	4	4	0	0	0	0
212531_at	LCN2	1,2	1	1	1	1	1	1	1	1	2	0	0	0	1	0
215780_s_at		1,2,3,4	1	1	2	2	2	3	4	4	4	0	0	0	0	0
217388_s_at	KYNU	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
217678_at	SLC7A11	3	1	1	1	1	1	1	0	0	0	0	0	0	1	0
218990_s_at	SPRR3	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0
219148_at	PBK	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0
223020_at	CLPTM1L	4	0	0	0	0	0	0	0	0	0	0	1	0	0	0
226145_s_at	FRAS1	1,2,3,4	1	1	2	3	4	4	4	4	4	0	0	0	0	0
228824_s_at	PTGR1	3	0	0	0	1	1	1	0	0	0	0	0	0	0	0
230193_at	WDR66	4	0	1	0	0	0	0	0	0	0	0	0	1	1	0
235811_at		1,2,3,4	1	2	2	4	4	4	4	4	4	0	0	0	1	0
238103_at	SMIM31	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0
242299_at		4	0	0	0	0	0	0	0	0	0	0	1	0	0	0
243623_at	SLC1A7	1,2,3,4	0	0	0	0	0	0	0	0	0	4	0	0	0	0
40665_at	FMO3	1,3	0	0	0	0	1	1	1	1	2	0	0	0	0	0

TABLE C.10: *Cox models including the top-1000-variance genes and mandatory clinical covariates. For each subgroup s genes with a mean inclusion frequency larger than 0.4 in any model are selected. Values correspond to the number of selections of each gene in each Cox model across all four subgroups. Subgroups: $s = 1$: GSE29013; $s = 2$: GSE31210; $s = 3$: GSE37745; $s = 4$: GSE50081.*

Probe set	Gene	*s*	w0.1	w0.2	w0.3	w0.4	w0.5	w0.6	w0.7	w0.8	w0.9	all	sub	lasso	ridge	rf
1554997_a_at	PTGS2	4	1	1	0	0	0	0	0	0	0	0	0	0	0	1
1555758_a_at	CDKN3	1,2,3,4	0	2	2	2	2	2	2	1	2	4	0	0	0	0
202357_s_at	CFB	2	0	0	0	0	0	0	0	0	0	0	1	0	0	1

202619_s_at	PLOD2	1,2,3,4	1	1	1	2	2	3	4	4	4	4	0	0	1	0
202935_s_at	SOX9	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
203159_at	GLS	1,2,3,4	0	1	1	1	1	4	4	4	4	4	0	0	0	0
203559_s_at	AOC1	1,2,3,4	2	3	3	4	4	4	4	4	4	4	0	0	2	1
203764_at	DLGAP5	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
204260_at	CHGB	2,4	1	2	1	2	2	2	2	2	1	0	0	0	1	0
204532_x_at		2	1	1	1	0	0	0	0	0	0	0	0	0	0	0
204580_at	MMP12	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
204753_s_at	HLF	1,2,3,4	0	0	0	1	2	3	3	4	4	4	0	0	0	0
205239_at	AREG	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1
205430_at	BMP5	2	1	1	0	0	0	0	0	0	0	0	0	0	1	1
205433_at	BCHE	1,2,3,4	0	1	1	2	4	4	4	4	4	4	0	0	0	0
205624_at	CPA3	4	1	1	0	0	0	0	0	0	0	0	1	1	0	0
205626_s_at	CALB1	3	0	0	1	0	0	0	0	0	0	0	0	1	0	0
206224_at	CST1	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0
206262_at	ADH1C	2	1	1	1	1	1	0	0	0	0	0	0	1	1	1
206641_at	TNFRSF17	4	0	0	0	1	1	1	0	0	0	0	0	0	0	0
208960_s_at	KLF6	1,2,3,4	0	0	0	1	1	1	1	3	2	4	0	0	0	0
209374_s_at	IGHM	1,2,3,4	0	0	0	0	1	2	1	3	4	4	0	0	0	0
209875_s_at	SPP1	1,2,3,4	2	3	3	4	4	4	4	4	4	4	0	1	1	1
210119_at	KCNJ15	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
210387_at		4	0	1	1	1	1	1	0	0	0	0	0	0	0	0
210397_at	DEFB1	1,2,3,4	2	2	2	2	2	3	4	4	4	4	0	1	1	1
212327_at	LIMCH1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
212531_at	LCN2	2	1	1	1	1	1	1	1	1	1	0	0	0	1	0
215780_s_at		1,2,3,4	1	1	2	3	4	4	4	4	4	4	0	0	0	0
217388_s_at	KYNU	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0
217678_at	SLC7A11	3	1	1	1	1	1	1	0	0	0	0	0	0	1	0
219148_at	PBK	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
222073_at	COL4A3	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1
226145_s_at	FRAS1	1,2,3,4	1	2	2	2	4	4	4	4	4	4	0	0	0	0
227081_at	DNALI1	4	0	0	1	1	0	0	0	0	0	0	0	0	0	0

227662_at	SYNPO2	1	1	1	0	0	0	0	0	0	0	0	0	0	0
227702_at	CYP4X1	4	0	1	1	1	1	1	1	0	0	0	0	0	0
228782_at	SCGB3A2	4	0	0	0	0	0	0	0	0	0	0	0	0	1
228824_s_at	PTGR1	3	1	1	1	1	1	1	1	0	0	0	0	0	0
230193_at	WDR66	4	1	1	1	0	0	0	0	0	0	0	0	1	0	1
230746_s_at		1	0	1	1	1	1	0	0	0	0	0	0	0	0	0
231484_at		2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
232481_s_at	SLITRK6	4	0	0	0	0	1	1	0	1	0	0	0	0	0	0
235811_at		1,2,3,4	1	2	2	4	4	4	4	4	4	4	0	0	1	0
238751_at		1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
242299_at		4	0	0	0	0	0	0	0	0	0	0	1	0	0	0
242915_at	ZNF682	2	0	1	1	1	1	1	1	1	0	0	0	0	0	0
40665_at	FMO3	1,2,3,4	0	0	1	1	1	1	1	1	2	4	0	0	0	0
AFFX-r2-Bs-thr-5_s_at		2	0	0	1	1	1	0	0	0	0	0	0	0	0	0

TABLE C.11: *Cox models including the top-1000-variance genes. For each subgroup s genes with a mean inclusion frequency larger than 0.4 in any model are selected. Values correspond to the number of selections of each gene in each Cox model across all four subgroups. Subgroups: $s = 1$: GSE29013; $s = 2$: GSE31210; $s = 3$: GSE37745; $s = 4$: GSE50081.*

a	b	Subgroup	TP	FP	TN	FN	Model size
-4.00	0.25	1	6	0	94	0	6
-4.00	0.25	2	3	0	94	3	3
-3.50	0.25	1	6	0	94	0	6
-3.50	0.25	2	4	0	94	2	4
-3.00	0.25	1	6	1	93	0	7
-3.00	0.25	2	6	2	92	0	8
-2.50	0.25	1	6	3	91	0	9
-2.50	0.25	2	6	4	90	0	10
-2.00	0.25	1	6	4	90	0	10
-2.00	0.25	2	6	6	88	0	12
-4.00	0.50	1	6	0	94	0	6
-4.00	0.50	2	3	0	94	3	3
-3.50	0.50	1	6	1	93	0	7
-3.50	0.50	2	6	2	92	0	8
-3.00	0.50	1	6	2	92	0	8
-3.00	0.50	2	6	3	91	0	9
-2.50	0.50	1	6	3	91	0	9
-2.50	0.50	2	6	4	90	0	10
-2.00	0.50	1	6	5	89	0	11
-2.00	0.50	2	6	6	88	0	12
-4.00	0.75	1	6	0	94	0	6
-4.00	0.75	2	6	1	93	0	7
-3.50	0.75	1	6	1	93	0	7
-3.50	0.75	2	6	1	93	0	7
-3.00	0.75	1	6	2	92	0	8
-3.00	0.75	2	6	3	91	0	9
-2.50	0.75	1	6	3	91	0	9
-2.50	0.75	2	6	4	90	0	10
-2.00	0.75	1	6	5	89	0	11
-2.00	0.75	2	6	7	87	0	13
-4.00	1.00	1	6	1	93	0	7
-4.00	1.00	2	6	1	93	0	7
-3.50	1.00	1	6	1	93	0	7
-3.50	1.00	2	6	2	92	0	8
-3.00	1.00	1	6	2	92	0	8
-3.00	1.00	2	6	3	91	0	9
-2.50	1.00	1	6	4	90	0	10
-2.50	1.00	2	6	4	90	0	10
-2.00	1.00	1	6	6	88	0	12
-2.00	1.00	2	6	7	87	0	13

TABLE C.12: *Number of selected variables for varying hyperparameters a and b. Number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The best combinations with the smallest number of incorrectly selected variables (FP+FN) averaged over both subgroups are highlighted in gray.*

ν_0	h	Subgroup	TP	FP	TN	FN	Model size
0.01	10	1	6	2	92	0	8
0.01	10	2	6	4	90	0	10
0.02	10	1	6	2	92	0	8
0.02	10	2	6	4	90	0	10
0.05	10	1	6	1	93	0	7
0.05	10	2	6	3	91	0	9
0.10	10	1	6	1	93	0	7
0.10	10	2	6	2	92	0	8
0.01	50	1	6	2	92	0	8
0.01	50	2	6	3	91	0	9
0.02	50	1	6	1	93	0	7
0.02	50	2	6	2	92	0	8
0.05	50	1	6	1	93	0	7
0.05	50	2	6	2	92	0	8
0.10	50	1	6	1	93	0	7
0.10	50	2	6	1	93	0	7
0.01	100	1	6	1	93	0	7
0.01	100	2	6	2	92	0	8
0.02	100	1	6	1	93	0	7
0.02	100	2	6	2	92	0	8
0.05	100	1	6	1	93	0	7
0.05	100	2	6	2	92	0	8
0.10	100	1	6	1	93	0	7
0.10	100	2	6	1	93	0	7

TABLE C.13: *Number of selected variables for varying hyperparameters ν_0 and $\nu_1 = h \cdot \nu_0$. Number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The best combinations with the smallest number of incorrectly selected variables (FP+FN) averaged over both subgroups are highlighted in gray.*

n	p	Model	TP	FP	TN	FN	Model size
50	20	Subgroup s=1	3.4 (1.26)	0.4 (0.52)	13.6 (0.52)	2.6 (1.26)	3.8 (1.23)
50	20	Subgroup s=2	3.3 (1.06)	0.2 (0.42)	13.8 (0.42)	2.7 (1.06)	3.5 (1.08)
50	20	Combined	3.6 (0.84)	0.2 (0.42)	10.8 (0.42)	5.4 (0.84)	3.8 (0.79)
50	20	CoxBVSSL s=1	1.7 (1.25)	0.3 (0.48)	13.7 (0.48)	4.3 (1.25)	1.8 (1.48)
50	20	CoxBVSSL s=2	1.4 (0.52)	0.1 (0.32)	13.9 (0.32)	4.6 (0.52)	1.2 (0.92)
100	20	Subgroup s=1	6.0 (0.00)	0.5 (0.71)	13.5 (0.71)	0.0 (0.00)	6.5 (0.71)
100	20	Subgroup s=2	5.9 (0.32)	0.2 (0.42)	13.8 (0.42)	0.1 (0.32)	6.1 (0.57)
100	20	Combined	5.7 (2.45)	0.1 (0.32)	10.9 (0.32)	3.3 (2.45)	5.8 (2.35)
100	20	CoxBVSSL s=1	6.0 (0.00)	0.3 (0.48)	13.7 (0.48)	0.0 (0.00)	6.3 (0.48)
100	20	CoxBVSSL s=2	6.0 (0.00)	0.1 (0.32)	13.9 (0.32)	0.0 (0.00)	6.1 (0.32)
25	100	Subgroup s=1	0.2 (0.42)	2.0 (0.47)	92.0 (0.47)	5.8 (0.42)	2.2 (0.42)
25	100	Subgroup s=2	0.6 (0.52)	1.4 (0.52)	92.6 (0.52)	5.4 (0.52)	2.0 (0.00)
25	100	Combined	0.9 (0.57)	1.0 (0.67)	90.0 (0.67)	8.1 (0.57)	1.9 (0.74)
25	100	CoxBVSSL s=1	0.0 (0.00)	2.2 (0.42)	91.8 (0.42)	6.0 (0.00)	2.2 (0.42)
25	100	CoxBVSSL s=2	0.5 (0.53)	1.6 (0.52)	92.4 (0.52)	5.5 (0.53)	2.1 (0.32)
50	100	Subgroup s=1	1.6 (1.17)	0.8 (0.63)	93.2 (0.63)	4.4 (1.17)	2.4 (1.07)
50	100	Subgroup s=2	1.5 (1.27)	0.8 (0.79)	93.2 (0.79)	4.5 (1.27)	2.3 (0.67)
50	100	Combined	2.9 (0.99)	0.5 (0.53)	90.5 (0.53)	6.1 (0.99)	3.4 (0.70)
50	100	CoxBVSSL s=1	2.0 (1.25)	0.6 (0.52)	93.4 (0.52)	4.0 (1.25)	2.6 (1.17)
50	100	CoxBVSSL s=2	1.8 (1.40)	0.6 (0.70)	93.4 (0.70)	4.2 (1.40)	2.4 (1.07)
75	100	Subgroup s=1	4.0 (1.49)	0.2 (0.42)	93.8 (0.42)	2.0 (1.49)	4.2 (1.48)
75	100	Subgroup s=2	3.3 (1.42)	0.6 (0.52)	93.4 (0.52)	2.7 (1.42)	3.9 (1.10)
75	100	Combined	3.3 (1.42)	0.4 (0.52)	90.6 (0.52)	5.7 (1.42)	3.7 (1.25)
75	100	CoxBVSSL s=1	5.4 (1.58)	0.7 (0.48)	93.3 (0.48)	0.6 (1.58)	6.1 (1.60)
75	100	CoxBVSSL s=2	3.9 (1.20)	0.8 (0.79)	93.2 (0.79)	2.1 (1.20)	4.7 (1.16)
100	100	Subgroup s=1	5.7 (0.67)	0.6 (0.52)	93.4 (0.52)	0.3 (0.67)	6.3 (0.82)
100	100	Subgroup s=2	4.6 (1.17)	0.5 (0.53)	93.5 (0.53)	1.4 (1.17)	5.1 (0.99)
100	100	Combined	3.7 (0.48)	0.3 (0.48)	90.7 (0.48)	5.3 (0.48)	4.0 (0.00)
100	100	CoxBVSSL s=1	6.0 (0.00)	0.9 (0.32)	93.1 (0.32)	0.0 (0.00)	6.9 (0.32)
100	100	CoxBVSSL s=2	5.8 (0.63)	0.9 (0.32)	93.1 (0.32)	0.2 (0.63)	6.7 (0.67)
150	100	Subgroup s=1	6.0 (0.00)	0.8 (0.42)	93.2 (0.42)	0.0 (0.00)	6.8 (0.42)
150	100	Subgroup s=2	6.0 (0.00)	0.9 (0.32)	93.1 (0.32)	0.0 (0.00)	6.9 (0.32)
150	100	Combined	4.3 (1.16)	0.1 (0.32)	90.9 (0.32)	4.7 (1.16)	4.4 (1.07)
150	100	CoxBVSSL s=1	6.0 (0.00)	1.0 (0.00)	93.0 (0.00)	0.0 (0.00)	7.0 (0.00)
150	100	CoxBVSSL s=2	6.0 (0.00)	1.0 (0.00)	93.0 (0.00)	0.0 (0.00)	7.0 (0.00)

TABLE C.14: *Results of variable selection. Mean number (standard deviation) of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), and total number of selected variables (model size), computed over all ten training sets.*

n	p	s	Combined	KM-Combined	Subgroup	KM-Subgroup	CoxBVSSL
50	20	1	0.18 (0.03)	0.24 (0.01)	0.17 (0.03)	0.24 (0.01)	0.19 (0.04)
50	20	2	0.14 (0.02)	0.20 (0.01)	0.15 (0.02)	0.20 (0.01)	0.17 (0.02)
100	20	1	0.17 (0.02)	0.24 (0.01)	0.12 (0.02)	0.24 (0.01)	0.11 (0.02)
100	20	2	0.15 (0.04)	0.21 (0.01)	0.12 (0.03)	0.22 (0.01)	0.12 (0.03)
25	100	1	0.23 (0.03)	0.24 (0.02)	0.25 (0.04)	0.23 (0.01)	0.27 (0.04)
25	100	2	0.24 (0.07)	0.21 (0.02)	0.21 (0.03)	0.22 (0.03)	0.22 (0.03)
50	100	1	0.20 (0.03)	0.23 (0.01)	0.20 (0.02)	0.23 (0.01)	0.19 (0.03)
50	100	2	0.15 (0.03)	0.21 (0.02)	0.19 (0.05)	0.22 (0.03)	0.19 (0.04)
75	100	1	0.19 (0.02)	0.23 (0.01)	0.17 (0.03)	0.23 (0.01)	0.14 (0.03)
75	100	2	0.15 (0.02)	0.20 (0.01)	0.15 (0.03)	0.21 (0.01)	0.14 (0.03)
100	100	1	0.17 (0.02)	0.24 (0.01)	0.13 (0.01)	0.23 (0.01)	0.12 (0.01)
100	100	2	0.15 (0.02)	0.21 (0.01)	0.13 (0.03)	0.21 (0.01)	0.11 (0.02)
150	100	1	0.17 (0.03)	0.23 (0.01)	0.12 (0.01)	0.23 (0.01)	0.12 (0.01)
150	100	2	0.15 (0.01)	0.22 (0.01)	0.10 (0.01)	0.22 (0.01)	0.10 (0.01)

TABLE C.15: *Mean (standard deviation) of the integrated Brier score (computed over all test sets) for the prediction of subgroup $s = 1, 2$. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (KM-Subgroup) or combined (KM-Combined) training data.*

n	p	Model	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$	$X8$	$X9$	$X10$	$X11$	$X12$
50	20	CoxBVSSL	0.297	0.260	0.050	0.102	0.174	0.038	0.298	0.194	0.267	0.061	0.018	0.009
50	20	Subgroup	0.534	0.495	0.094	0.131	0.369	0.087	0.417	0.207	0.475	0.176	0.120	0.045
50	20	Combined	0.128	0.047	0.062	0.199	0.160	0.131	0.454	0.325	0.761	0.549	0.106	0.041
100	20	CoxBVSSL	1.000	1.000	0.357	0.387	0.967	0.940	0.993	0.669	1.000	1.000	0.983	0.292
100	20	Subgroup	0.998	1.000	0.043	0.144	0.855	0.707	0.893	0.089	0.994	0.994	0.862	0.130
100	20	Combined	0.295	0.259	0.210	0.655	0.035	0.040	0.552	0.314	0.972	0.960	0.840	0.126
50	100	CoxBVSSL	0.093	0.071	0.034	0.092	0.238	0.027	0.216	0.068	0.190	0.014	0.012	0.007
50	100	Subgroup	0.107	0.110	0.021	0.069	0.293	0.036	0.197	0.063	0.178	0.017	0.009	0.009
50	100	Combined	0.094	0.009	0.014	0.163	0.022	0.015	0.223	0.303	0.549	0.308	0.191	0.005
75	100	CoxBVSSL	0.807	0.801	0.028	0.038	0.745	0.143	0.299	0.106	0.785	0.374	0.210	0.018
75	100	Subgroup	0.649	0.653	0.008	0.012	0.717	0.079	0.158	0.032	0.705	0.239	0.086	0.014
75	100	Combined	0.050	0.009	0.007	0.273	0.181	0.092	0.090	0.180	0.851	0.599	0.399	0.029
100	100	CoxBVSSL	0.999	1.000	0.056	0.057	0.672	0.658	0.744	0.333	0.982	0.917	0.731	0.107
100	100	Subgroup	0.828	0.832	0.012	0.021	0.632	0.213	0.474	0.013	0.819	0.581	0.262	0.008
100	100	Combined	0.177	0.111	0.010	0.161	0.005	0.009	0.340	0.144	0.959	0.869	0.515	0.011
200	100	CoxBVSSL	1.000	1.000	0.297	0.386	0.999	0.996	0.997	0.886	1.000	1.000	1.000	0.337
200	100	Subgroup	1.000	1.000	0.005	0.007	0.993	0.774	0.892	0.244	1.000	0.994	0.979	0.094
200	100	Combined	0.258	0.079	0.071	0.243	0.016	0.005	0.695	0.620	1.000	1.000	0.997	0.115

TABLE C.16: *Mean posterior inclusion frequencies (averaged over all training sets) of the prognostic variables for subgroup 1. Variables included on average are highlighted in red.*

n	p	Model	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	$X7$	$X8$	$X9$	$X10$	$X11$	$X12$
50	20	CoxBVSSL	0.032	0.023	0.204	0.184	0.047	0.613	0.100	0.480	0.306	0.220	0.013	0.008
50	20	Subgroup	0.102	0.070	0.373	0.346	0.114	0.752	0.085	0.530	0.395	0.343	0.056	0.040
50	20	Combined	0.128	0.047	0.062	0.199	0.160	0.131	0.454	0.325	0.761	0.549	0.106	0.041
100	20	CoxBVSSL	0.071	0.150	0.906	0.881	0.637	0.852	0.579	0.905	0.998	0.945	0.822	0.117
100	20	Subgroup	0.056	0.059	0.849	0.848	0.413	0.692	0.180	0.761	0.901	0.787	0.603	0.059
100	20	Combined	0.295	0.259	0.210	0.655	0.035	0.040	0.552	0.314	0.972	0.960	0.840	0.126
50	100	CoxBVSSL	0.011	0.020	0.044	0.074	0.025	0.164	0.033	0.479	0.397	0.150	0.021	0.011
50	100	Subgroup	0.011	0.018	0.049	0.080	0.016	0.181	0.020	0.473	0.352	0.133	0.010	0.008
50	100	Combined	0.094	0.009	0.014	0.163	0.022	0.015	0.223	0.303	0.549	0.308	0.191	0.005
75	100	CoxBVSSL	0.007	0.020	0.316	0.349	0.016	0.461	0.039	0.610	0.594	0.223	0.071	0.010
75	100	Subgroup	0.006	0.014	0.232	0.287	0.014	0.480	0.019	0.642	0.483	0.203	0.058	0.014
75	100	Combined	0.050	0.009	0.007	0.273	0.181	0.092	0.090	0.180	0.851	0.599	0.399	0.029
100	100	CoxBVSSL	0.009	0.037	0.851	0.909	0.586	0.868	0.382	0.900	0.669	0.557	0.362	0.021
100	100	Subgroup	0.007	0.015	0.682	0.779	0.075	0.827	0.085	0.845	0.499	0.377	0.189	0.010
100	100	Combined	0.177	0.111	0.010	0.161	0.005	0.009	0.340	0.144	0.959	0.869	0.515	0.011
200	100	CoxBVSSL	0.010	0.045	1.000	1.000	0.997	0.998	0.885	1.000	1.000	1.000	1.000	0.427
200	100	Subgroup	0.007	0.006	1.000	1.000	0.698	0.884	0.238	1.000	0.943	0.939	0.910	0.197
200	100	Combined	0.258	0.079	0.071	0.243	0.016	0.005	0.695	0.620	1.000	1.000	0.997	0.115

TABLE C.17: *Mean posterior inclusion frequencies (averaged over all training sets) of the prognostic variables for subgroup 2. Variables included on average are highlighted in red.*

n	p	Model	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$
50	20	Cox.	0.28 (0.18)	0.24 (0.17)	-0.03 (0.09)	0.06 (0.11)	-0.13 (0.17)	0.01 (0.08)
50	20	Sub.	0.48 (0.25)	0.46 (0.26)	-0.04 (0.12)	0.07 (0.13)	-0.26 (0.21)	0.03 (0.12)
50	20	Comb.	0.05 (0.11)	0.01 (0.07)	0.02 (0.07)	0.13 (0.11)	-0.08 (0.09)	-0.06 (0.10)
100	20	Cox.	0.98 (0.18)	1.08 (0.19)	-0.00 (0.12)	0.05 (0.12)	-0.62 (0.19)	0.58 (0.20)
100	20	Sub.	0.96 (0.18)	1.05 (0.18)	0.00 (0.06)	0.05 (0.08)	-0.59 (0.22)	0.45 (0.24)
100	20	Comb.	0.14 (0.13)	0.14 (0.13)	0.11 (0.10)	0.32 (0.12)	-0.01 (0.05)	0.01 (0.05)
50	100	Cox.	0.08 (0.12)	0.06 (0.13)	-0.02 (0.09)	0.05 (0.12)	-0.18 (0.20)	0.01 (0.08)
50	100	Sub.	0.09 (0.16)	0.09 (0.18)	-0.01 (0.07)	0.05 (0.11)	-0.22 (0.22)	0.02 (0.09)
50	100	Comb.	0.05 (0.11)	0.01 (0.05)	0.00 (0.05)	0.11 (0.10)	-0.02 (0.06)	-0.01 (0.05)
75	100	Cox.	0.74 (0.23)	0.79 (0.25)	-0.00 (0.05)	-0.00 (0.06)	-0.61 (0.25)	0.09 (0.12)
75	100	Sub.	0.59 (0.27)	0.63 (0.29)	-0.00 (0.04)	0.00 (0.05)	-0.57 (0.26)	0.05 (0.11)
75	100	Comb.	0.04 (0.07)	0.01 (0.04)	0.00 (0.04)	0.15 (0.11)	-0.10 (0.10)	-0.05 (0.07)
100	100	Cox.	0.99 (0.19)	1.11 (0.20)	0.00 (0.06)	-0.01 (0.07)	-0.46 (0.22)	0.46 (0.24)
100	100	Sub.	0.80 (0.24)	0.89 (0.26)	-0.00 (0.05)	-0.00 (0.06)	-0.44 (0.23)	0.16 (0.21)
100	100	Comb.	0.10 (0.10)	0.07 (0.08)	0.01 (0.04)	0.08 (0.08)	-0.01 (0.04)	-0.01 (0.04)
200	100	Cox.	1.05 (0.13)	1.25 (0.15)	-0.01 (0.09)	0.02 (0.10)	-0.70 (0.14)	0.66 (0.14)
200	100	Sub.	1.00 (0.13)	1.19 (0.14)	0.00 (0.04)	0.00 (0.04)	-0.70 (0.14)	0.49 (0.18)
200	100	Comb.	0.11 (0.09)	0.05 (0.07)	0.05 (0.06)	0.11 (0.10)	-0.03 (0.04)	0.00 (0.04)

n	p	Model	$X7$	$X8$	$X9$	$X10$	$X11$	$X12$
50	20	Cox.	0.25 (0.22)	0.12 (0.17)	-0.17 (0.23)	-0.04 (0.12)	-0.01 (0.06)	-0.00 (0.05)
50	20	Sub.	0.31 (0.26)	0.13 (0.23)	-0.34 (0.34)	-0.11 (0.20)	-0.05 (0.12)	-0.00 (0.07)
50	20	Comb.	0.26 (0.22)	0.20 (0.21)	-0.59 (0.26)	-0.35 (0.23)	-0.05 (0.11)	0.00 (0.06)
100	20	Cox.	0.79 (0.20)	0.17 (0.18)	-1.10 (0.20)	-1.10 (0.20)	-0.67 (0.20)	-0.11 (0.14)
100	20	Sub.	0.65 (0.22)	0.04 (0.09)	-1.08 (0.21)	-1.03 (0.22)	-0.56 (0.23)	-0.05 (0.10)
100	20	Comb.	0.27 (0.12)	0.17 (0.11)	-0.79 (0.15)	-0.71 (0.16)	-0.45 (0.16)	-0.05 (0.08)
50	100	Cox.	0.17 (0.17)	0.05 (0.13)	-0.14 (0.16)	-0.01 (0.06)	-0.01 (0.05)	0.00 (0.04)
50	100	Sub.	0.16 (0.17)	0.04 (0.10)	-0.14 (0.16)	-0.01 (0.06)	-0.01 (0.05)	0.00 (0.05)
50	100	Comb.	0.17 (0.14)	0.21 (0.16)	-0.49 (0.22)	-0.26 (0.20)	-0.14 (0.13)	0.00 (0.04)
75	100	Cox.	0.25 (0.18)	0.05 (0.11)	-0.75 (0.24)	-0.34 (0.21)	-0.15 (0.18)	-0.00 (0.06)
75	100	Sub.	0.13 (0.18)	0.02 (0.08)	-0.63 (0.27)	-0.20 (0.18)	-0.06 (0.14)	-0.00 (0.05)
75	100	Comb.	0.06 (0.08)	0.12 (0.11)	-0.77 (0.20)	-0.46 (0.19)	-0.23 (0.19)	-0.01 (0.06)
100	100	Cox.	0.66 (0.25)	0.12 (0.15)	-1.07 (0.24)	-0.90 (0.27)	-0.49 (0.25)	-0.06 (0.09)
100	100	Sub.	0.36 (0.24)	0.01 (0.05)	-0.74 (0.31)	-0.50 (0.30)	-0.17 (0.21)	-0.00 (0.04)
100	100	Comb.	0.18 (0.13)	0.09 (0.10)	-0.77 (0.19)	-0.54 (0.20)	-0.26 (0.19)	-0.01 (0.04)
200	100	Cox.	0.92 (0.15)	0.32 (0.14)	-1.30 (0.15)	-1.22 (0.15)	-0.85 (0.15)	-0.13 (0.11)
200	100	Sub.	0.74 (0.17)	0.13 (0.13)	-1.32 (0.16)	-1.19 (0.16)	-0.76 (0.15)	-0.06 (0.08)
200	100	Comb.	0.33 (0.12)	0.27 (0.15)	-0.88 (0.12)	-0.80 (0.11)	-0.54 (0.10)	-0.06 (0.05)

TABLE C.18: *Posterior mean (standard deviation) of regression coefficients (averaged over all training sets) for the prognostic variables in subgroup 1. Variables included on average are highlighted in red. Cox.: CoxBVSSL; Sub.: Subgroup; Comb.: Combined.*

n	p	Model	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$
50	20	Cox.	0.01 (0.07)	-0.01 (0.06)	0.22 (0.19)	0.16 (0.17)	0.01 (0.09)	-0.55 (0.30)
50	20	Sub.	0.02 (0.11)	-0.02 (0.09)	0.33 (0.28)	0.27 (0.27)	0.04 (0.16)	-0.65 (0.30)
50	20	Comb.	0.05 (0.11)	0.01 (0.07)	0.02 (0.07)	0.13 (0.11)	-0.08 (0.09)	-0.06 (0.10)
100	20	Cox.	-0.00 (0.07)	-0.01 (0.10)	0.84 (0.27)	0.88 (0.27)	0.37 (0.23)	-0.54 (0.24)
100	20	Sub.	-0.01 (0.07)	-0.01 (0.08)	0.75 (0.27)	0.79 (0.27)	0.28 (0.22)	-0.53 (0.26)
100	20	Comb.	0.14 (0.13)	0.14 (0.13)	0.11 (0.10)	0.32 (0.12)	-0.01 (0.05)	0.01 (0.05)
50	100	Cox.	-0.00 (0.05)	-0.01 (0.07)	0.04 (0.10)	0.06 (0.15)	0.01 (0.09)	-0.12 (0.22)
50	100	Sub.	-0.00 (0.05)	-0.01 (0.07)	0.04 (0.11)	0.06 (0.16)	0.01 (0.06)	-0.14 (0.22)
50	100	Comb.	0.05 (0.11)	0.01 (0.05)	0.00 (0.05)	0.11 (0.10)	-0.02 (0.06)	-0.01 (0.05)
75	100	Cox.	-0.00 (0.04)	-0.01 (0.06)	0.28 (0.21)	0.32 (0.22)	0.01 (0.06)	-0.43 (0.25)
75	100	Sub.	-0.00 (0.04)	-0.01 (0.05)	0.21 (0.17)	0.25 (0.19)	0.00 (0.06)	-0.45 (0.21)
75	100	Comb.	0.04 (0.07)	0.01 (0.04)	0.00 (0.04)	0.15 (0.11)	-0.10 (0.10)	-0.05 (0.07)
100	100	Cox.	0.00 (0.04)	0.00 (0.06)	0.85 (0.24)	0.97 (0.26)	0.35 (0.25)	-0.72 (0.28)
100	100	Sub.	0.00 (0.04)	-0.01 (0.05)	0.64 (0.28)	0.73 (0.30)	0.05 (0.12)	-0.80 (0.24)
100	100	Comb.	0.10 (0.10)	0.07 (0.08)	0.01 (0.04)	0.08 (0.08)	-0.01 (0.04)	-0.01 (0.04)
200	100	Cox.	0.01 (0.04)	-0.00 (0.05)	1.23 (0.15)	1.29 (0.17)	0.63 (0.16)	-0.61 (0.15)
200	100	Sub.	0.01 (0.04)	-0.00 (0.04)	1.15 (0.16)	1.18 (0.19)	0.46 (0.21)	-0.68 (0.18)
200	100	Comb.	0.11 (0.09)	0.05 (0.07)	0.05 (0.06)	0.11 (0.10)	-0.03 (0.04)	0.00 (0.04)

n	p	Model	$X7$	$X8$	$X9$	$X10$	$X11$	$X12$
50	20	Cox.	0.03 (0.12)	0.45 (0.24)	-0.29 (0.27)	-0.18 (0.21)	-0.01 (0.06)	0.00 (0.04)
50	20	Sub.	0.03 (0.12)	0.50 (0.27)	-0.35 (0.31)	-0.26 (0.28)	-0.01 (0.09)	0.01 (0.06)
50	20	Comb.	0.26 (0.22)	0.20 (0.21)	-0.59 (0.26)	-0.35 (0.23)	-0.05 (0.11)	0.00 (0.06)
100	20	Cox.	0.19 (0.18)	0.75 (0.23)	-1.07 (0.25)	-0.90 (0.24)	-0.55 (0.23)	-0.03 (0.10)
100	20	Sub.	0.10 (0.15)	0.66 (0.27)	-0.98 (0.30)	-0.77 (0.28)	-0.41 (0.23)	-0.02 (0.08)
100	20	Comb.	0.27 (0.12)	0.17 (0.11)	-0.79 (0.15)	-0.71 (0.16)	-0.45 (0.16)	-0.05 (0.08)
50	100	Cox.	0.01 (0.09)	0.42 (0.34)	-0.40 (0.34)	-0.13 (0.24)	-0.01 (0.07)	-0.00 (0.05)
50	100	Sub.	0.00 (0.07)	0.43 (0.34)	-0.37 (0.29)	-0.12 (0.20)	-0.00 (0.05)	-0.00 (0.05)
50	100	Comb.	0.17 (0.14)	0.21 (0.16)	-0.49 (0.22)	-0.26 (0.20)	-0.14 (0.13)	0.00 (0.04)
75	100	Cox.	0.02 (0.08)	0.61 (0.26)	-0.60 (0.29)	-0.22 (0.20)	-0.05 (0.12)	0.00 (0.05)
75	100	Sub.	0.01 (0.06)	0.65 (0.27)	-0.50 (0.25)	-0.20 (0.15)	-0.04 (0.10)	0.00 (0.05)
75	100	Comb.	0.06 (0.08)	0.12 (0.11)	-0.77 (0.20)	-0.46 (0.19)	-0.23 (0.19)	-0.01 (0.06)
100	100	Cox.	0.20 (0.19)	0.79 (0.28)	-0.68 (0.31)	-0.51 (0.28)	-0.26 (0.23)	-0.00 (0.05)
100	100	Sub.	0.06 (0.09)	0.74 (0.25)	-0.51 (0.29)	-0.33 (0.24)	-0.14 (0.19)	0.00 (0.05)
100	100	Comb.	0.18 (0.13)	0.09 (0.10)	-0.77 (0.19)	-0.54 (0.20)	-0.26 (0.19)	-0.01 (0.04)
200	100	Cox.	0.33 (0.16)	1.04 (0.17)	-1.19 (0.17)	-1.11 (0.17)	-0.85 (0.16)	-0.19 (0.13)
200	100	Sub.	0.15 (0.13)	0.96 (0.17)	-1.14 (0.19)	-1.00 (0.19)	-0.74 (0.18)	-0.11 (0.09)
200	100	Comb.	0.33 (0.12)	0.27 (0.15)	-0.88 (0.12)	-0.80 (0.11)	-0.54 (0.10)	-0.06 (0.05)

TABLE C.19: *Posterior mean (standard deviation) of regression coefficients (averaged over all training sets) for the prognostic variables in subgroup 2. Variables included on average are highlighted in red. Cox.: CoxBVSSL; Sub.: Subgroup; Comb.: Combined.*

n	p	Model	TP	FP	TN	FN	Model size
50	20	Subgroup s=1	3.2 (1.32)	0.5 (0.53)	9.5 (0.53)	6.8 (1.32)	3.7 (1.16)
50	20	Subgroup s=2	3.4 (0.97)	0.2 (0.42)	9.8 (0.42)	6.6 (0.97)	3.6 (0.70)
50	20	Combined	3.5 (0.71)	0.0 (0.00)	8.0 (0.00)	8.5 (0.71)	3.5 (0.71)
50	20	CoxBVSSL s=1	1.9 (1.20)	0.1 (0.32)	9.9 (0.32)	8.1 (1.20)	1.8 (1.48)
50	20	CoxBVSSL s=2	2.2 (0.92)	0.0 (0.00)	10.0 (0.00)	7.8 (0.92)	2.2 (0.92)
100	20	Subgroup s=1	7.7 (0.48)	0.1 (0.32)	9.9 (0.32)	2.3 (0.48)	7.8 (0.42)
100	20	Subgroup s=2	6.3 (1.25)	0.2 (0.42)	9.8 (0.42)	3.7 (1.25)	6.5 (1.08)
100	20	Combined	5.4 (1.35)	0.0 (0.00)	8.0 (0.00)	6.6 (1.35)	5.4 (1.35)
100	20	CoxBVSSL s=1	9.1 (0.57)	0.4 (0.70)	9.6 (0.70)	0.9 (0.57)	9.5 (0.85)
100	20	CoxBVSSL s=2	8.0 (1.15)	0.0 (0.00)	10.0 (0.00)	2.0 (1.15)	8.0 (1.15)
50	100	Subgroup s=1	1.6 (1.17)	0.8 (0.79)	89.2 (0.79)	8.4 (1.17)	2.4 (0.52)
50	100	Subgroup s=2	1.9 (0.74)	0.8 (0.63)	89.2 (0.63)	8.1 (0.74)	2.7 (0.67)
50	100	Combined	2.5 (0.85)	0.2 (0.42)	87.8 (0.42)	9.5 (0.85)	2.7 (0.82)
50	100	CoxBVSSL s=1	1.5 (0.85)	0.9 (0.74)	89.1 (0.74)	8.5 (0.85)	2.4 (0.52)
50	100	CoxBVSSL s=2	2.0 (0.94)	0.8 (0.63)	89.2 (0.63)	8.0 (0.94)	2.8 (0.63)
75	100	Subgroup s=1	3.8 (1.14)	0.5 (0.71)	89.5 (0.71)	6.2 (1.14)	4.3 (1.42)
75	100	Subgroup s=2	3.1 (0.99)	0.3 (0.67)	89.7 (0.67)	6.9 (0.99)	3.4 (1.07)
75	100	Combined	3.3 (1.25)	0.2 (0.42)	87.8 (0.42)	8.7 (1.25)	3.5 (0.97)
75	100	CoxBVSSL s=1	4.8 (1.93)	0.4 (0.70)	89.6 (0.70)	5.2 (1.93)	5.2 (1.87)
75	100	CoxBVSSL s=2	3.3 (1.16)	0.3 (0.48)	89.7 (0.48)	6.7 (1.16)	3.6 (1.17)
100	100	Subgroup s=1	5.5 (1.43)	0.1 (0.32)	89.9 (0.32)	4.5 (1.43)	5.6 (1.51)
100	100	Subgroup s=2	4.9 (1.45)	0.5 (0.53)	89.5 (0.53)	5.1 (1.45)	5.4 (1.26)
100	100	Combined	3.8 (0.63)	0.2 (0.42)	87.8 (0.42)	8.2 (0.63)	4.0 (0.67)
100	100	CoxBVSSL s=1	7.8 (1.03)	0.1 (0.32)	89.9 (0.32)	2.2 (1.03)	7.9 (0.99)
100	100	CoxBVSSL s=2	6.9 (1.60)	0.1 (0.32)	89.9 (0.32)	3.1 (1.60)	7.0 (1.41)
200	100	Subgroup s=1	8.5 (0.85)	0.0 (0.00)	90.0 (0.00)	1.5 (0.85)	8.5 (0.85)
200	100	Subgroup s=2	8.2 (0.63)	0.0 (0.00)	90.0 (0.00)	1.8 (0.63)	8.2 (0.63)
200	100	Combined	5.3 (0.67)	0.1 (0.32)	87.9 (0.32)	6.7 (0.67)	5.4 (0.70)
200	100	CoxBVSSL s=1	9.3 (0.48)	0.9 (0.74)	89.1 (0.74)	0.7 (0.48)	10.2 (0.63)
200	100	CoxBVSSL s=2	9.9 (0.32)	0.1 (0.32)	89.9 (0.32)	0.1 (0.32)	10.0 (0.47)

TABLE C.20: *Results of variable selection. Mean number (standard deviation) of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), and total number of selected variables (model size), computed over all ten training sets.*

n	p	s	Combined	KM-Combined	Subgroup	KM-Subgroup	CoxBVSSL
50	20	1	0.22 (0.03)	0.25 (0.01)	0.21 (0.04)	0.25 (0.01)	0.22 (0.03)
50	20	2	0.16 (0.05)	0.21 (0.01)	0.15 (0.03)	0.22 (0.02)	0.16 (0.04)
100	20	1	0.18 (0.01)	0.24 (0.01)	0.12 (0.02)	0.24 (0.01)	0.12 (0.02)
100	20	2	0.15 (0.02)	0.21 (0.01)	0.11 (0.02)	0.22 (0.01)	0.10 (0.02)
50	100	1	0.21 (0.02)	0.23 (0.01)	0.22 (0.02)	0.23 (0.01)	0.23 (0.02)
50	100	2	0.17 (0.02)	0.21 (0.01)	0.18 (0.05)	0.22 (0.01)	0.17 (0.03)
75	100	1	0.21 (0.03)	0.24 (0.01)	0.18 (0.03)	0.24 (0.02)	0.17 (0.03)
75	100	2	0.16 (0.02)	0.21 (0.01)	0.16 (0.04)	0.22 (0.01)	0.15 (0.04)
100	100	1	0.18 (0.02)	0.23 (0.01)	0.15 (0.02)	0.23 (0.00)	0.12 (0.02)
100	100	2	0.17 (0.03)	0.22 (0.01)	0.13 (0.03)	0.22 (0.01)	0.12 (0.02)
200	100	1	0.18 (0.01)	0.23 (0.00)	0.11 (0.01)	0.23 (0.00)	0.11 (0.01)
200	100	2	0.16 (0.02)	0.22 (0.01)	0.10 (0.02)	0.22 (0.01)	0.10 (0.02)

TABLE C.21: *Mean (standard deviation) of the integrated Brier score (computed over all test sets) for the prediction of subgroup $s = 1, 2$. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (KM-Subgroup) or combined (KM-Combined) training data.*

Probe set	Gene	Subgroup	Comb.	Sub.	CoxBVSSL	Biology
204580_at	MMP12	$s = 1$	-	+	+	cancer-related, lung-related
216491_x_at	IGHM	$s = 1$	-	+	+	immune response
206291_at	NTS	$s = 2$	-	+	-	cancer-related
215125_s_at	UGT1A	$s = 2$	-	+	-	cancer-related
224590_at	XIST	$s = 2$	-	-	+	cancer-related
224588_at	XIST	$s = 2$	-	-	+	cancer-related
214218_s_at	XIST	$s = 2$	-	-	+	cancer-related
221728_x_at	XIST	$s = 2$	-	-	+	cancer-related
228782_at	SCGB3A2	$s = 4$	-	+	+	cancer-related, lung-related
230378_at	SCGB3A1	$s = 4$	-	-	+	cancer-related, lung-related
200869_at	RPL18A	$s = 1, 2, 3, 4$	+	-	-	viral diseases, not cancer-related
214777_at	IGKC	$s = 1, 2, 3, 4$	+	+	+	immune response

TABLE C.22: *Top-100-variance genes selected on average in any of the three models (Comb.: Combined; Sub.: Subgroup; CoxBVSSL) for each subgroup ($s = 1$: GSE29013; $s = 2$: GSE31210; $s = 3$: GSE37745; $s = 4$: GSE50081). Selected: +; Non-selected: -.*

Gene filter	s	Combined	KM-Comb.	Subgroup	KM-Sub.	CoxBVSSL
30 Kratz	GSE29013	0.10 (0.02)	0.12 (0.01)	0.13 (0.03)	0.16 (0.02)	0.14 (0.03)
30 Kratz	GSE31210	0.11 (0.02)	0.11 (0.01)	0.14 (0.05)	0.13 (0.03)	0.13 (0.04)
30 Kratz	GSE37745	0.21 (0.02)	0.22 (0.01)	0.20 (0.01)	0.19 (0.01)	0.21 (0.02)
30 Kratz	GSE50081	0.16 (0.02)	0.15 (0.01)	0.16 (0.02)	0.17 (0.02)	0.16 (0.02)
Top-100-v.	GSE29013	0.15 (0.03)	0.12 (0.01)	0.17 (0.04)	0.16 (0.02)	0.17 (0.05)
Top-100-v.	GSE31210	0.11 (0.02)	0.11 (0.01)	0.13 (0.03)	0.13 (0.03)	0.13 (0.03)
Top-100-v.	GSE37745	0.20 (0.01)	0.22 (0.01)	0.20 (0.01)	0.19 (0.01)	0.20 (0.01)
Top-100-v.	GSE50081	0.15 (0.02)	0.15 (0.01)	0.16 (0.02)	0.17 (0.02)	0.16 (0.02)

TABLE C.23: *Mean (standard deviation) of the integrated Brier score (computed over all test sets) for the prediction of subgroup s. Kaplan-Meier estimator (KM) for a reference model without covariates based on subgroup (KM-Sub.) or combined (KM-Comb.) training data.*

Eidesstattliche Erklärung

Hiermit erkläre ich, Katrin Madjar, dass ich die vorliegende Dissertation mit dem Titel "Survival models with selection of genomic covariates in heterogeneous cancer studies" selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Dissertation ist bisher keiner anderen Fakultät vorgelegt worden. Ich erkläre, dass ich bisher kein Promotionsverfahren erfolglos beendet habe und dass keine Aberkennung eines bereits erworbenen Doktorgrades vorliegt.

Ort, Datum

Unterschrift