2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Bi-level selection in Longitudinal Analysis with Time Dependent Covariates

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. ADRIAAN BLOMMAERT

Madona Wijaya

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Bi-level selection in Longitudinal Analysis with Time Dependent Covariates

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. ADRIAAN BLOMMAERT

Madona Wijaya

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Acknowledgements

I am highly delighted and give thanks to The Almighty God who has given me the opportunity of being among the luckiest student from the developing country to study in University of Hasselt. I have to show my sincere appreciation to the Flemsih Interuniversity Council (VLIR) for granting me a scholarship in order to fulfill my ambition of becoming a professional Statistician.

The completion of this thesis could not have been possible without the help and support of a number of people. First and foremost, I would like to thank my internal supervisor who shared with me his long-term experience and for his thorough guidance. In addition, I wish to say a big thanks to my external supervisor for sharing their invaluable knowledge with me, provision of useful materials to ease my task and always attending to my request and questions.

I have to commend the persistent patience, perseverance, and support of my parents, family, and friends during the period I stayed away from home in the course of acquiring knowledge. Finally, I acknowledge those that I used their materials in the cause of this research work.

Abstract

Penalization methods such as the Lasso (Least absolute shrinkage and selection operator) (Tibshirani 1996) have been used in a variety of contexts to automatically select relevant variables and enhance predictive performance in regression models. Examples are analysis of genetic data and feature selection in image processing. Recently the group Lasso, group bridge, and group MCP have been proposed to deal with group structure in the data. The aim of group Lasso is to select a priori defined groups of variables as a whole. The group bridge and group MCP, in contrast, can perform bi-level selection by encouraging sparse solutions at the group and individual variable levels. In this thesis, we consider the problem of time dependent covariates in longitudinal data analysis to select relevant variable as well as to select the correct lag for each variable. Fu (2003) developed the so-called penalized GEE in longitudinal studies. Since working independent correlation is assumed, this comes down to ordinary lasso. The aforementioned group penalization methods are also applied to time dependent covariates. The performance of these methods in terms of model error and model complexity is compared via simulation studies. It is found that the performance of group bridge is superior to the other methods. In general, penalization method improves the fitted model but cannot consistently select the true coefficients in the model. Due to high multicolinearity within the covariate variable, an alternative method is proposed by approximating the lagged coefficients with the B-spline basis functions. The proposed method does not perform variable selection, however, it is quite useful as it can approximate the true parameters.

Keywords: B-splines, Basis, Bridge, Collinearity, GEE, Lag Selection, Lasso, Longitudinal Data, MCP, Model Error, Penalized Estimating Equation, Time Dependent Covariates

Table of Contents

Chapter 1

Introduction

1.1 Background

Longitudinal study involves following individuals over time, thereby measuring a random outcome variable (response) and risk factors (covariates) at least at two different points in time, and often more. It is in constrast to cross-sectional study, in which a single outcome is measured for each individual. Conceptually and practically, it is useful to distinguish two types of risk variables in longitudinal setting: *time-fixed covariates* and *time-dependent covariates*. A time fixed covariate, like sex, remains the same over all longitudinal observations of the same subject. On the other hand, a time dependent covariate, such as blood pressure, may vary from observation to observation on the same subject.

Time dependent covariates will be the main objective in this study as it is important in epidemiological modeling to explore the structure of the underlying system, that is, to correctly characterize the lag relationship between exposure and the disease outcome (Diggle *et al*., 2002). This work focuses on penalization methods to select relevant variables as well as to select the correct lag for each variable.

Variable selection is fundamental in many kinds of statistical modeling. To select the relevant variables among a large number of potential predictors is not an easy task. A poor choice of covariates will make the resulting prediction model uninterpretable, unrealistic, unreliable, or useless. It is undesirable to keep irrelevant covariates in the final model since this makes it difficult to interpret the result as the model becomes more complex. If an important covariate is eliminated, then coefficient estimates will be biased and may decrease the model's predictive ability. To reduce variability and to obtain a more interpretable model, we are often interested in selecting a smaller number of important variables.

A common approach to variable selection is to identify the best subset of variables according to some criterion. However, this approach is unstable (Breiman, 1996) and becomes

computationally infeasible as the number of variables grows to even moderate size (Breheny and Huang, 2009). For these reasons, penalization methods for variable selection have received much attention in the recent literature.

Penalization method such as bridge regression is proposed by Frank and Friedman (1993) for variable selection in linear model. The *Lasso* (Least square absolute shrinkage and selection operator) is introduced by Tibshirani (1996) and has been used in a variety of contexts to automatically select relevant variables and enhance predictive performance in regression models. Examples are analysis of genetic data and feature selection in image processing. Fan and Li (2001) provided insights into how to construct a penalty function that gives the best performance in selecting significant variables without creating excessive biases via *Smoothly Clipped Absolute Deviation* (SCAD) penalty. Zhang (2007) further proposed the *Minimax Concave Penalty* (MCP), as an improvement of the SCAD penalty. These methods focuses on the selection of individual variables. To accommodate selection at the group level, the *group Lasso* has been proposed by Yuan and Lin (2006). The aim of this method is to select a priori defined groups of variables as a whole. This can be used when for example factors appear in regression. This approach performs at group level but not at an individual level variable selection. The *group bridge* (Huang *et al.*, 2007) and the *group MCP* (Breheny and Huang, 2009), in contrast, perform bi-level selection by encouraging sparse solution at the group and individual variable levels.

To apply variable selection for marginal regression with longitudinal data, Fu (2003) introduced penalized estimating equation via bridge penalty to the GEE. The Lasso is a special case of this and will be used as a penalty model in time dependent covariates. When there are time dependent covariates, Pepe and Anderson (1994) and Diggle (2002) have suggested that the marginal models be estimated by GEE with the independent working correlation to produce unbiased estimates. As independent working correlation is assumed, the aforementioned penalization methods can be directly implemented in the simulation study to perform bi-level selection although we may lose efficiency. However, this assumption is sufficient as we conduct the simulation with large sample size. Besides, applying the correct working correlation is beyond the scope of this thesis. In addition, we propose variable selection for time dependent covariates by combining the B-splines method with the group Lasso, where the lagged covariates can be represented as a linear model with appropriate basis elements.

1.2 Data Structure and Notation

In this thesis, the following notation and definitions are adopted. Let Y_{it} be the t^{th} measurement available for the i^{th} subject, $i = 1, 2, ..., N$; $t = 1, 2, ..., T$. We assume a common set of discrete follow-up times (t), with a well-defined final study measurement time T. Let X_{it} be a time varying covariate on subject i at time t. It is assumed that Y_{it} and X_{it} are simultaneously measured and that for cross-sectional analyses Y_{it} is directly correlated with X_{it}. Furthermore, it is assumed that X_{it} can be divided into L groups. In other words, the underlying model structure is allowed to include more than one time varying covariate leading to the following mechanism:

$$\begin{cases} Y_{it} = \sum_{\ell=1}^{L} \sum_{m=0}^{p-1} \beta_{\ell(m+1)} X_{\ell i(t-m)} + b_i + \varepsilon_{it} \\ X_{\ell it} = \rho X_{\ell i(t-1)} + e_{it} \end{cases} \tag{1}$$

where $\boldsymbol{\beta}_\ell = \left(\beta_{\ell 1}, \beta_{\ell 2}, ..., \beta_{\ell p}\right)'$ is a p-vector of unknown regression coefficients belong to the lth group; $\mathbf{x}_\ell = \left(X_{\ell i(t)}, X_{\ell i(t-1)}, ..., X_{\ell i(t-p+1)}\right)$ is a p-vector of lagged covariates for the i^{th} subject belong to the lth group; b_i is a random effect; ε_{it} and e_{it} are residuals; b_{i}, ε_{it}, and e_{it} are mutually independent. Model (1) shows that the lagged covariates follow an AR(1) structure. Througout this thesis, lagged covariates are standardized prior to fitting to ensure that the penalty is applied equally.

Furthermore, we will assume that the full model is correctly specified and that the functional forms of the covariates and response are correctly specified, that is, linear. We will also assume that some of the true coefficients of the full model are zero while the others are not zero. In this simpler version of the model selection problem, the goal is to find the true subset, that is, to identify which coefficients are zero and which are not.

1.3 Outline of Thesis

This thesis is organized as follows. In Chapter 2, we provide a brief description of marginal model, i.e. GEE, to time dependent covariates. Penalization methods are also introduced. We start by motivating the use of classical penalty models, following with a discussion on the penalty model to the GEE in longitudinal studies proposed by Fu (2003). Group penalization is further elaborated to deal in such situation where it is practically more meaningful to identify not just the correct time lag within each variable but also to select relevant variables. Finally, combining the B-splines method with the group Lasso procedure is discussed. Chapter 3 explores the performance of the different penalization methods in terms of estimation model and variable selection via simulation studies. The discussion and conclusion is in Chapter 4. Most of the outputs resulting from the simulation studies carried out in Chapter 3 are presented in Appendix A and B. Example of R 2.12 codes to generate the simulated data and to carry out the analysis presented in Chapter 3 are given in Appendix C.

Chapter 2

Methodology

In longitudinal datasets, there typically is correlation among a subject's repeated measurements. In a marginal model, this correlation is not of primary interest but it must be taken into account to make proper inferences. Marginal models are appropriate when inferences about the population's average are the primary focus. Liang and Zeger (1986) proposed so-called Generalized Estimating Equations (GEE) to fit a marginal model, which only require the correct specification of the univariate marginal distributions provided one is willing to adopt a working assumption about the association structure (Molenberghs and Verbeke, 2005). It produces efficient estimates if the working correlation structure is correctly specified but remains consistent and provides correct standard errors if the working correlation structure is incorrectly specified.

When there are time-dependent covariates, Pepe and Anderson (1994) have pointed out that the consistency of GEE is not assured with arbitrary working correlation structures unless the *full covariate conditional mean* assumption (FCCM) is satisfied, that is,

$$\mu_{it} = E\left[Y_{it} \mid X_{it}\right] = E\left[Y_{it} \mid X_{i1}, X_{i2}, \ldots, X_{iT}\right] \tag{2}$$

This conclusion is supported and reported in Diggle *et al.* (2002) via simulation study with continuous response data that depends on both current and lagged values of the covariate. The left hand side of equation (2) states that in the fitted model must contain all the information with respect to the response. Since the underlying model considered in this thesis is adopted from Diggle *et al.* (2002), all the lagged covariates taken up in the fitted model must represent all information regarding the response if a general working correlation is to be used. If this assumption does not hold then the independent working correlation should be used otherwise biased regression estimates may be obtained. Consequently, the independent working correlation is suggested to be adopted as a "safe" analysis choice when using GEE with time dependent covariates (Lai and Small, 2006).

2.1. The Classical Penalization Method

A major challenge in regression analysis is to decide which covariates, among many potential ones, are to be included in the model. It is customary to use stepwise selection and subset selection. These procedures, however, are unstable and ignore the stochastic errors introduced by the selection process. Several methods, including bridge regression and Lasso have been proposed to select variables and estimate their regression coefficients simultaneously.

The Lasso is a shrinkage method proposed by Tibshirani (1996) that minimizes the residual sum of squares subject to a constraint on the sum of absolute values of the regression coefficients. In other words, the solution to the Lasso is defined to be the value β that minimizes the objective function:

$$Q(\beta) = \frac{1}{2n} || \mathbf{Y} - \mathbf{X}\beta ||^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \tag{3}$$

The Lasso shrinks some coefficients and sets others to zero, and hence tries to retain the good features of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression (Tibshirani, 1996). A limitation of the Lasso method is its inability to do grouped selection. When a group of highly correlated feature exists, the method will only select one of the features and discard the others.

Bridge regression is a broad class of the penalized regression method proposed by Frank and Friedman (1993). The bridge estimate can be obtained by minimizing the objective function (3) with a penalty function $\lambda \sum_{j=1}^{p} \left| \beta_j \right|^{\gamma}$ where $\gamma \geq 0$ and $\lambda \geq 0$. It includes ridge regression with $\gamma = 2$, the Lasso with $\gamma = 1$, and subset selection with $\gamma = 0$ as special cases. If $0 < \gamma \leq 1$, bridge estimators produce sparse model. Due to the general penalty form, bridge regression naturally fits any situation where it needs variable selection or where there exists multicollinearity (Park and Yoon, 2011).

2.2. Penalized GEE

The unavailability of a joint likelihood in the GEE method is a great challenge to implement penalty model since the classical approach requires specification of the full joint likelihood. Fu (2003) worked out this problem by introducing penalized estimating equations with bridge penalty, and the Lasso as a special case, to the GEE. The usual bridge regression minimized the penalized deviance criterion. To obtain the estimator of the penalty model, one needs to solve the penalized equations:

$$\begin{cases} F_1(\beta, X, y) + \lambda d(\beta_1, \gamma) = 0 \\ \vdots \\ F_p(\beta, X, y) + \lambda d(\beta_p, \gamma) = 0 \end{cases} \tag{4}$$

where $F_j(\beta, X, y)$ is minus the j-th score of the likelihood and $d(\beta_j, \gamma)$ is the partial derivative of the penalty function with respect to β_j. Fu has studied that the function F_j can take the form of estimating functions. It can be generalized to let F_j be the minus quasi-score functions of GEE. Since the minus estimating function of the GEE satisfies the Jacobian condition, a unique estimator is determined by the penalized GEE (Fu, 2003). It potentially improves the performance of the GEE estimator and enjoys the same properties as linear penalty models. Further, it yields an asymptotically consistent and normally distributed estimator.

The GEE model with Lasso penalty is considered in this thesis. This comes down to the ordinary Lasso fit as independent working correlation is assumed for the time dependent covariate model.

2.3. Group Penalization

All (past) time measurements of a time dependent variable form a group. When the covariates can be naturally grouped, it is also of interest to accommodate selection at the group level. Group penalties can be considered to have a form in which an outer penalty f_O is applied to a sum of inner penalties f_I. The penalty applied to a group of covariates is

$$f_O\left(\sum_{j=1}^{p_\ell} f_I\left(|\beta_{\ell j}|\right)\right). \tag{5}$$

Note that group Lasso and group bridge fit into this framework with an outer bridge penalty; the former possesses an inner ridge penalty while the latter has an inner Lasso penalty (Breheny and Huang, 2009).

2.3.1. Group Lasso

In some problems, the predictors belong to pre-defined groups, for example in time dependent covariates or collection of indicator (dummy) variables for representing the levels of a categorical predictor. In this situation, it may be desirable to shrink and select the members of a group together. Group Lasso is one way to achieve this and is a natural extension of Lasso. This was proposed by Yuan and Lin (2007) which solves the convex optimization problem:

$$Q(\beta) = \frac{1}{2n} || \mathbf{Y} - \mathbf{X}\beta ||^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \, || \beta_l ||_2 \tag{6}$$

where the $\sqrt{p_l}$ terms accounts for the varying group sizes and $||\ .\ ||_2$ is the Euclidean norm (not squared). This procedure acts like the Lasso at the group level depending on λ, an entire group of predictors may drop from the model. In other words, within a group, the covariates are either all equal to zero or else all nonzero. In fact, if the group sizes are all one, it reduces to the Lasso (Friedman *et al.*, 2010; Hastie *et al.*, 2009). It produces a strong bias towards zero, it tends to overselect the true number of groups, and it is incapable of selecting important elements within a group.

2.3.2. Group Bridge

Unlike the group Lasso, the group bridge produces sparse solution both at group level and at the level of the individual covariates within a group. It was proposed by Huang *et al*. (2007), whose estimate minimizes:

$$Q(\beta) = \frac{1}{2n} || \mathbf{Y} - \mathbf{X}\beta ||^2 + \lambda \sum_{\ell=1}^{L} p_\ell^\gamma || \beta_l ||_1^\gamma \tag{7}$$

When $0 < \gamma < 1$, the group bridge criterion can be used for variable selection at the group and individual levels simultaneously. Group bridge suffers from a number of practical difficulties due to the fact that the bridge penalty is not everywhere differentiable for $\gamma < 1$. For any positive value of λ, zero is a local minimum of the group bridge penalty. This complicates optimization and incurs the potential drawback of dropping groups that would prove to be nonzero when the solution converges (Breheny and Huang, 2009). In the simulation studies, we take $\gamma = ½$ for group bridge.

2.3.3. Group MCP

Zhang (2007) introduced a nonconvex penalty called the minimax concave penalty (MCP) that behaves similarly as the smoothly clipped absolute deviation (SCAD) proposed by Fan and Li (2001). MCP begins by applying the same rate of penalization as the Lasso, but continuously relaxes that penalization until, when $|\beta| > \gamma\lambda$, the rate of penalization drops to 0. The goal of this penalty is to eliminate the unimportant variables from the model while leaving the important variables unpenalized. This would be equivalent to fitting an unpenalized model in which the truly nonzero variables are known in advance (the so-called oracle model). Both MCP and SCAD accomplish this asymptotically and are said to have the oracle property (Fan and Li, 2001, Breheny and Huang, 2009; Zhang, 2007). Breheny and Huang (2009) proposed group MCP whose estimate minimizes:

$$Q(\beta) = \frac{1}{2n} || \mathbf{Y} - \mathbf{X}\beta ||^2 + \sum_{\ell=1}^{L} f_{\lambda,b}\left(\sum_{j=1}^{p_\ell} f_{\lambda,\gamma}\left(|\beta_{\ell m}| \right) \right) \tag{8}$$

where λ is a regularization parameter that determines the magnitude of penalization, γ is a tuning parameter that affects the range over which the penalty is applied, and *b* is the tuning parameter of the outer family chosen to be $\gamma\lambda p_l/2$ in order to ensure that the group level penalty attains is maximum if and only if each of its components are at their maximum. When γ is small, MCP penalty has a broader influence and is best at retaining the unbiasedness of the MCP penalty for large coefficients. It also has the risk of creating objective functions with nonconvexity problem that is difficult to optimize and thus gives solution that is discontinuous with respect to λ. Note that γ is not scale invariant with respect to the response variable. Breheny and Huang (2009) recommend to standardize the variables and use $\gamma = 3$ as it works well in their simulations.

2.4. Regularization Parameter Selection

Penalization methods require the selection of the regularization parameter. Prediction error (PE) is customary to select optimal λ and is defined as the average error in the prediction of the response variable given the covariate for future cases not used in the construction of a prediction equation. Let $\hat{\mu}(\mathbf{X})$ be a prediction procedure constructed using the present data, the prediction error can be expressed as follows:

$$\mathrm{PE}(\hat{\mu}) = E\left[Y - \hat{\mu}(\mathbf{X})\right]^2 \tag{9}$$

where the expectation is taken only with respect to the new observation. The prediction error can be decomposed as:

$$\mathrm{PE}(\hat{\mu}) = E\left[Y - E(Y \mid \mathbf{X})\right]^2 + E\left[E(Y \mid \mathbf{X}) - \hat{\mu}(\mathbf{X})\right]^2 \tag{10}$$

The first component is the inherent prediction error due to the noise; the second is due to lack of fit with an underlying model. This component is called *model error* and is denoted by $\mathrm{ME}(\hat{\mu})$. The size of the model error reflects performance of different model selection procedure and has the form (Fan and Li, 2001):

$$\mathrm{ME}(\hat{\mu}) = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T E\left[\mathbf{X}\mathbf{X}^T\right]\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \tag{11}$$

where $E\left[\mathbf{X}\mathbf{X}^T\right] = \Sigma$ is the covariance matrix of $\mathbf{X}$. The simulation results are reported in terms of model error rather than prediction error.

2.5. Estimation using B-splines

Splines are polynomial curves that are joined at points called knots. These offer more flexibility than traditional polynomial regression to fit non-linear and non-polynomial relationships. The B-spline (De Boor, 1978) is a class of splines that can be seen as a generalization of the truncated power. It is more complicated than the truncated power but numerically more stable and efficient (Costa, 2008; Hastie *et al.*, 2009; Ruppert *et al.*, 2003). The B-spline basis function is, in essence, a rescaling of each of the piecewise functions. The idea is similar to rescaling a set of covariate variables by mean subtraction to reduce collinearity. The rescaling in the B-spline basis reduces the collinearity in the basis function of the model matrix (Keele, 2008).

Combining the smoothing spline method with the group Lasso procedure is proposed to perform variable selection to time dependent covariates. We consider B-splines to represent the lagged covariates in terms of B-spline basis functions as a remedy to handle collinearity within the variables. The penalized estimation procedure is then used to select the sets of basis functions. Time dependent covariates model (1) may be over specified, however, and lead to estimate of lagged covariates that are difficult to interpret (Diggle *et al.*, 2002). Thus, we can further assume that the lagged coefficients follow a lower order smooth function $p^* < p$:

$$\beta_{\ell(m+1)} = \gamma_0 + \sum_{j=1}^{p^*} \gamma_j B_j(m), \text{ where } \ell = 1,2,...,L; m = 0,1,..., p-1 \tag{12}$$

where $B_i(m)$ is a B-spline basis vector. Replacing $\beta_{\ell m}$ by its B-spline approximation in equation (12), model (1) can be approximated as:

$$\begin{aligned} Y_{it} &= \gamma_0 + \sum_{\ell=1}^{L}\sum_{m=0}^{p-1}\sum_{j=1}^{p^*} \gamma_j B_j(m) X_{\ell i(t-m)} + b_i + \varepsilon_{it} \\ &= \gamma_0 + \sum_{\ell=1}^{L}\sum_{j=1}^{p^*} \gamma_j \left[\sum_{m=0}^{p-1} B_j(m) X_{\ell i(t-m)}\right] + b_i + \varepsilon_{it} \\ &= \gamma_0 + \sum_{\ell=1}^{L}\sum_{j=1}^{p^*} \gamma_j X^*_{\ell i} + b_i + \varepsilon_{it} \end{aligned} \tag{13}$$

where $X^*_{\ell i} = \sum_{m=0}^{p-1} B_j(m) X_{\ell i(t-m)}$.

Selecting the optimal regularization parameter to model (13) cannot be done using model error as the true structure of the underlying model is unknown. We propose to use the prediction error of the form:

$$R = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^T V^{-1} \left(Y_i - \hat{Y}_i\right) \tag{14}$$

where V is the covariance matrix of Y. In matrix notation, model (13) can be expressed as $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$ where $\mathbf{b}_i \sim N(0, D)$ and $\boldsymbol{\varepsilon}_i \sim N(0, \Sigma_i)$. It follows that $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$ where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \Sigma_i$ (Verbeke and Molenberghs, 2002). Prediction error (14) is used to estimate the test error based on the training set in order to choose the model complexity. Unfortunately training error is not a good estimate of test error, as it does not properly account for model complexity (Hastie, 2009). Thus, test error based on the test set is also considered to see the performance of the selected model.

2.6. Software

R version 2.12 is the software used for simulation studies, graphical, and data analysis. Package 'grpreg' is available to fit group Lasso, group bridge, and group MCP and is developed by Breheny (2009).

Chapter 3

Simulation Studies

3.1. Data Generating Models

A time dependent covariate model under mechanism (1) is considered for the simulation study to demonstrate the performance of the different penalization methods with $b_i \sim N(0,1), \varepsilon_{it} \sim N(0,1)$, and, $e_{it} \sim N\left(0, 1-\rho^2\right)$. 800 datasets each of which contained data on N subjects with 10 observations per subject were generated, where the number of subject is taken to be $N = 20$, $N = 50$, or $N = 100$. A range of correlation ($\rho = 0 - 0.9$) is considered in the simulation study to investigate the influence of time dependence in the covariates. Six scenarios are generated in the simulation by varying the number of groups and the number of covariates within a group. Additional scenario is considered to estimation using B-splines.

1. **Scenario I**: 1 group of time dependent covariate is considered with 3 lagged covariates. The true parameter takes values $\boldsymbol{\beta} = (0, 0.5, 0)^T$.
2. **Scenario II**: 3 more lagged covariates are added to scenario I with the true parameter is $\boldsymbol{\beta} = (0, 0.75, 0, 0.5, -0.5, 0)^T$.
3. **Scenario III**: 2 groups of time dependent covariates are considered with 3 lagged covariates for each group. The true parameters take values $\boldsymbol{\beta}_1 = (0.5, 0, 0.75)^T$ and $\boldsymbol{\beta}_2 = (0, 0, 0)^T$.
4. **Scenario VI**: The same number of groups as in scenario III with 6 lagged covariates for each group. The coefficients in the first group are all nonzero and in the second group are all zero. The true parameters are $\boldsymbol{\beta}_1 = (1, 1, 0.75, 0.75, 0.5, 0.5)^T$ and $\boldsymbol{\beta}_2 = (0, 0, 0, 0, 0, 0)^T$.
5. **Scenario V**: 8 groups of time dependent covariates are considered with 3 lagged covariates for each group. The true parameters for the first 4 groups are zero while the

remaining 4 nonzero groups are $\boldsymbol{\beta}_5 = (1,0,0)^T$, $\boldsymbol{\beta}_6 = (1,0.75,0)^T$, $\boldsymbol{\beta}_7 = (1,0.75,0.5)^T$, and $\boldsymbol{\beta}_8 = (-1,0.75,0.5)^T$.

6. **Scenario VI**: The same number of groups as in scenario V with 10 lagged covariates for each group. The 4 nonzero groups are $\boldsymbol{\beta}_5 = (1,0,0,..,0)^T$, $\boldsymbol{\beta}_6 = (1,0.75,0,...,0)^T$, $\boldsymbol{\beta}_7 = (1,0.75,0.5,0,...,0)^T$, and $\boldsymbol{\beta}_8 = (-0.5,0.5,-0.5,1,1,0.75,0.5,-0.75,0.5,-0.5)^T$.
7. **Scenario VII**: 5 groups of time dependent covariates are considered with 12 lagged covariates for each group. All the coefficients in first 2 groups are set to 0, thus the true underlying model only depends on a single lagged covariate from each nonzero group. The number of subject is 100 and the number of observations for each subject is 15.

The performance of each penalization method is measured by means of model error. The model complexity of the selected model is summarized in terms of correct deletions (**C**), erroneous deletions (**E**), and proportion correct models (**P**) at group as well as individual level. Correct deletions are defined as the average number (per simulation) of truly zero coefficients correctly estimated as zero, and erroneous deletions as the average number of truly nonzero coefficients erroneously set to zero. Because of the true parameter in scenario II is $\boldsymbol{\beta} = (0,0.75,0,0.5,-0.5,0)^T$, up to 3 correct deletions and 3 wrong deletions are possible. Meanwhile proportion correct models are defined as the proportion of trials in which exactly the true subset of nonzero covariates is chosen.

To illustrate how the data was generated, a sample of 3 simulated data under scenario I with weak and strong correlation within covariate are shown in Figure 3.1 and 3.2, respectively. The individual profiles are plotted from 20 randomly selected subjects. Note that the covariates also follow a longitudinal structure and are depicted in bottom panel while the response profiles are depicted in top panel. It can be observed that the presence of correlation influences the longitudinal profiles of the covariate. This can cause a problem as collinearity leads to poor performance in linear models with inaccurate estimation and prediction.

Figure 3.1. *Individual profiles from the generated data under scenario I with ρ = 0.1*

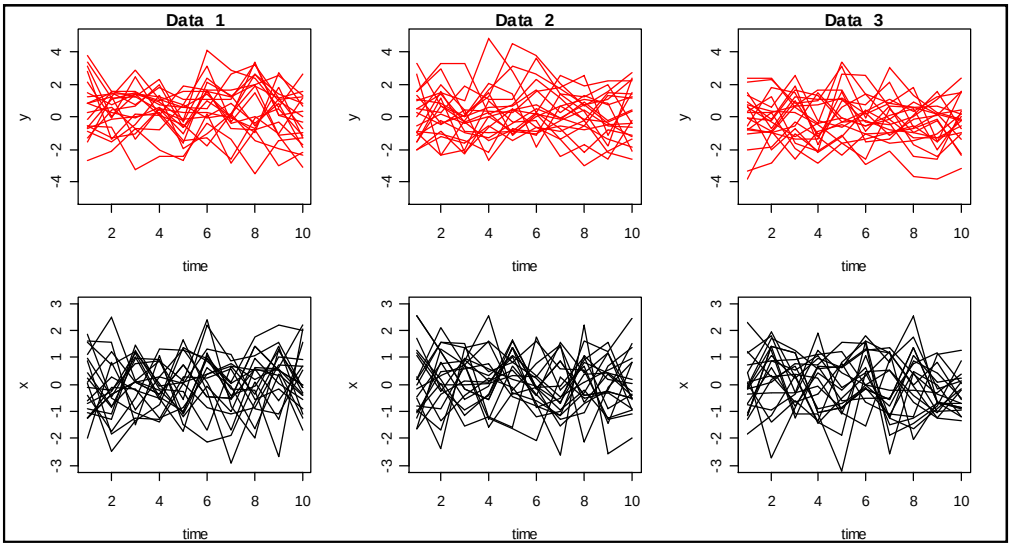
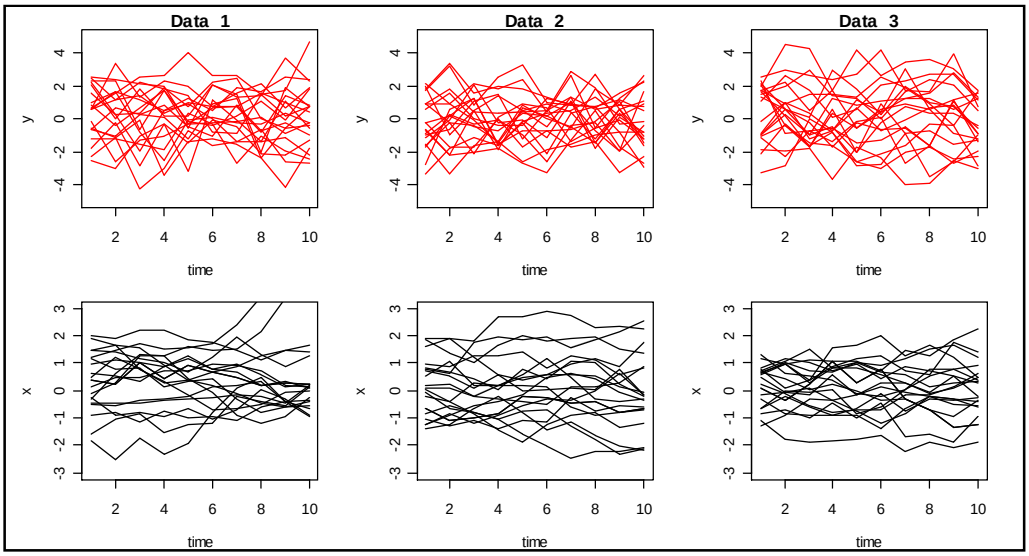


Figure 3.2. *Individual profiles from the generated data under scenario I with ρ = 0.9*



3.2. Model Error Comparison

The optimal regularization parameter and the corresponding estimated model error for the different scenarios are summarized in Table 3.1. The table only presents a small part of the results for scenario I-VI with $N = 50$ and $\rho = 0.3$. The complete results can be found in Table A1 – A10 (Appendix A). Extreme case with time independent covariate ($\rho = 0$) was also generated as a benchmark. Group penalization was not fitted to scenario I and II since no grouping structure is present. Therefore only the Lasso method was fitted to the simulated data and then compared to the GEE method. It can be observed that the penalty method improves the fitted model since the estimated model error from the Lasso method is smaller than the GEE method. However, the performance of the Lasso in terms of variable selection is still poor. For scenario I, it is found that the Lasso method can only make at most one correct deletion when there are 2 true zero coefficients. The proportion of correct models for this scenario is 21%. The performance of the Lasso method becomes worse as the more lagged covariates in the model (scenario II).

Group penalization is applied to scenario III – VI. Again, we can observe that the different penalization methods improve the fit as compared to the GEE. Overall, the performance of group bridge is superior to the other methods, followed by Lasso, group Lasso, and then group MCP in terms of model error, the number of true zero groups deleted, the number of true zero lagged covariates deleted, and the percentage of correct model selected. Although the group bridge tends to outperform the other methods in terms of variable selection, its performance is still below our expectation at the individual level. For example, out of 64 zero coefficients, the group bridge can only make 50 correct deletions under scenario VI. In addition, the proportion of correct model selected becomes very poor as the number of groups increases.

Table 3.1. *Model error (standard error) and model complexity for the generated data under scenario I – VI with the number of subject is 50 and the correlation within covariate is 0.3*

				Group		Variable		
Scenario*	**Method**	**λ**	**ME(s.e.)**	**C**	**E**	**C**	**E**	**P**
I	Lasso	0.041	0.0099(0.0004)			0.88	0.00	0.21
(zc = 2)	GEE	0	0.0139(0.0005)					
II	Lasso	0.020	0.0226(0.0007)			0.77	0.00	0.01
(zc = 3)	GEE	0	0.0262(0.0007)					
	Lasso	0.045	0.0198(0.0006)	0.47	0.00	1.97	0.00	0.06
III	gLasso	0.020	0.0233(0.0006)	0.04	0.00	0.13	0.00	0.00
(zg = 1;	gBridge	0.026	0.0164(0.0006)	0.91	0.00	3.02	0.00	0.25
zc = 4)	gMCP	0.034	0.0269(0.0007)	0.01	0.00	0.04	0.00	0.00
	GEE	0	0.0278(0.0007)					
	Lasso	0.038	0.0414(0.0011)	0.41	0.00	2.68	0.00	0.01
IV	gLasso	0.022	0.0421(0.0010)	0.01	0.00	0.05	0.00	0.01
(zg = 1;	gBridge	0.029	0.0312(0.0009)	0.95	0.00	5.74	0.00	0.94
zc = 6)	gMCP	0.022	0.0506(0.0010)	0.00	0.00	0.01	0.00	0.00
	GEE	0	0.0517(0.0010)					
	Lasso	0.041	0.0801(0.0013)	1.75	0.00	6.99	0.00	0.00
V	gLasso	0.017	0.0941(0.0014)	0.13	0.00	0.38	0.00	0.00
(zg = 4;	gBridge	0.026	0.0663(0.0012)	3.69	0.00	12.01	0.00	0.01
zc = 15)	gMCP	0.017	0.1089(0.0014)	0.01	0.00	0.04	0.00	0.01
	GEE	0	0.1112(0.0015)					
	Lasso	0.041	0.2144(0.0024)	1.87	0.00	32.73	0.00	0.00
VI	gLasso	0.021	0.2928(0.0028)	0.01	0.00	0.05	0.00	0.00
(zg = 4;	gBridge	0.026	0.1559(0.0022)	3.89	0.00	49.79	0.00	0.00
zc = 64)	gMCP	0.041	0.3736(0.0030)	0.04	0.00	0.85	0.00	0.00
	GEE	0	0.3913(0.0032)					

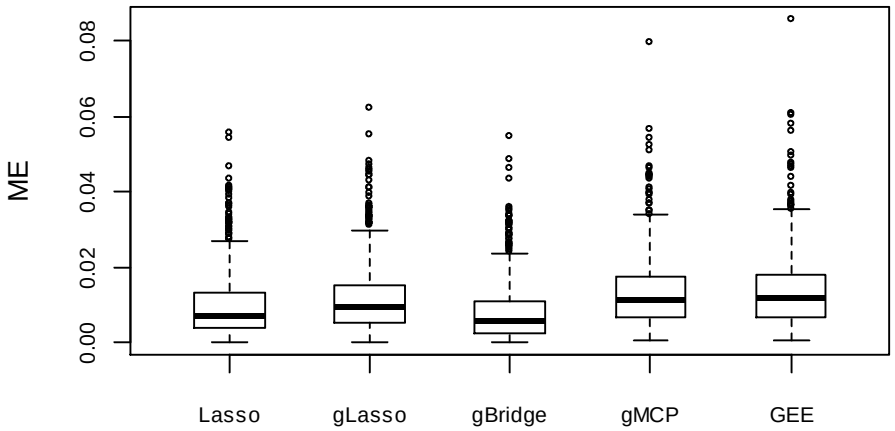
**zg is the number of zero groups in the model; zc is the number of zero coefficients in the model*

The group Lasso tends to select more groups and variables than there actually are in the generating models. On average, the model selected by the group Lasso method retains all the groups in the model. Note that group MCP performs better than group Lasso in terms of variable selection at individual level when the groups are fairly sparse, that is, when half or more members are zero, but it is a poor group selector as its number of correct deletions at the group level is very small. This is in agreement with the simulation results reported in Huang and Breheny (2009). The results can be seen in Table A6 (Appendix A) under scenario VI. However, this is only true for small *N* and the selection is inconsistent as *N* goes larger.

Varying the values of correlation within covariate, the model error increases as the correlation is increased. This is found for all the penalty methods but it is always smaller than the GEE. This suggests that applying penalization to a time dependent covariate model may potentially achieve better estimation and prediction when collinearity is present in covariates. In addition, the consistency of the estimator and the standard error performs rather well when taking more subjects in the simulated data. It is found that the Lasso method does not consistently select the correct groups and lagged covariates. Similar behavior is observed for group Lasso as collinearity is present. For group bridge, the number

of correct deletions decreases as the correlation within covariate is increased. On the other hand, collinearity seems to improve the number of correct deletions for group MCP although it is very small.

Figure 3.3. *Boxplot of the estimated model error from the generated data under scenario III with the number of subject is 50 and the correlation within covariate is 0.3*

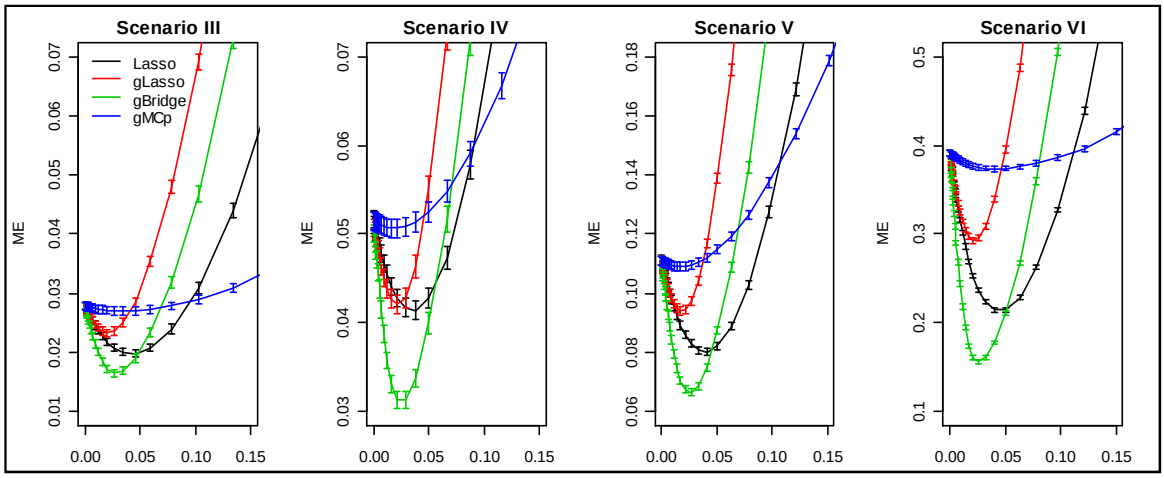


To see the variability of the estimated model error of the different penalization methods, boxplots are plotted in Figure 3.3 (see Figure B1 – B6 in Appendix B for other scenarios). It can be observed that the group bridge has the smallest variability. Similar results are observed across the range of ρ and the different value of N. In general, the distribution of the model error produced by the penalization methods is skewed to the right and bounded at zero.

3.3. The Estimated Model Error Curves

Figure 3.4 demonstrates the shape of the estimated model error for the various selection penalty methods under scenario III – VI (the others curves are given in Table B7 – B12 in Appendix B). The means of these estimates (+/- 1SE) are plotted as a function of the corresponding regularization parameter. The shapes are convex with respect to λ and they reach minimum values at a very small λ. This might be the reason that the penalization methods fail to yield a sufficiently sparse solution as we have already seen in Section 3.2. Again, these figures show that the group bridge attains the smallest model error as compared to the others. The same results can be observed for the different scenarios under the different settings.

Figure 3.4. *Estimated model error curves and their standard errors for the various selection penalty methods from the generating data with the number of subject is 50 and the correlation within covariate is 0.3*



3.4. The Lagged Coefficient Paths

To illustrate the shrinkage behavior of the penalization methods toward zero, profiles of the Lasso, group Lasso, group bridge, and group MCP coefficients are plotted in Figure 3.5 as the regularization parameter is varied. These plots are obtained from fitting the penalization methods to 5 randomly simulated data under scenario III with the number of subject is 50 and the correlation within covariate is 0.3. The black dashed line indicates the first group with 2 nonzero and 1 zero lagged covariates and the red solid line indicates the second group with 3 zero lagged covariates. The coefficients decrease to 0 as the regularization parameter grows larger. The decrease is not always strictly monotonic as can be seen from group Lasso, group bridge, and group MCP. In addition, all the group bridge coefficients hit zero at sufficiently small λ while the group MCP coefficient hit zero at larger λ.

Figure 3.5. *Profiles of the Lasso, group Lasso, group bridge and group MCP coefficients*

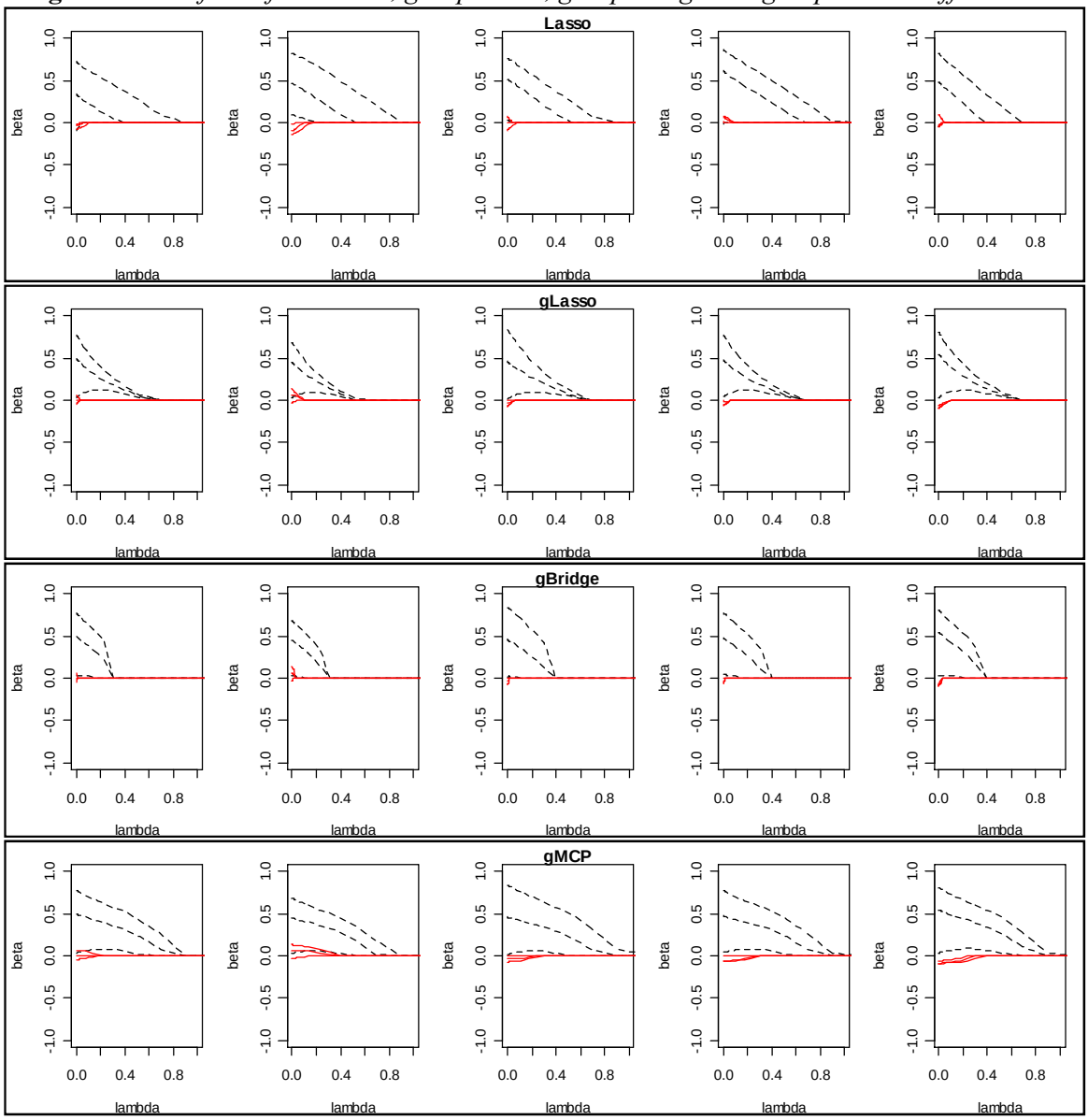


Table 3.2 shows the number of coefficients over the 800 simulated data that was set to zero across the values of regularization parameter. The same data generating model is used where β_{11} and β_{13} are the true nonzero coefficients and the others are zero. The number of zero coefficients selected by the optimal λ is shown in red font color. This indicates that the model error is a poor method for variable selection in time dependent covariates data. The number is even worse for group Lasso and group MCP. In general, we would expect a method that can select the optimal λ where the parameter β_{12}, β_{21}, β_{22}, and β_{23} are simulatenously set to zero for all 800 simulated data. For this setting, a good method would have picked the values somewhere in 0.304, 0.231, 0.134, and 0.524 as the optimal λ for the Lasso, group Lasso, group bridge, and group MCP, respectively, and these are shown in blue font color. These results are also plotted in Figure B13 (Appendix B).

Table 3.2. *The number of zero coefficients across the range of regularization parameter*

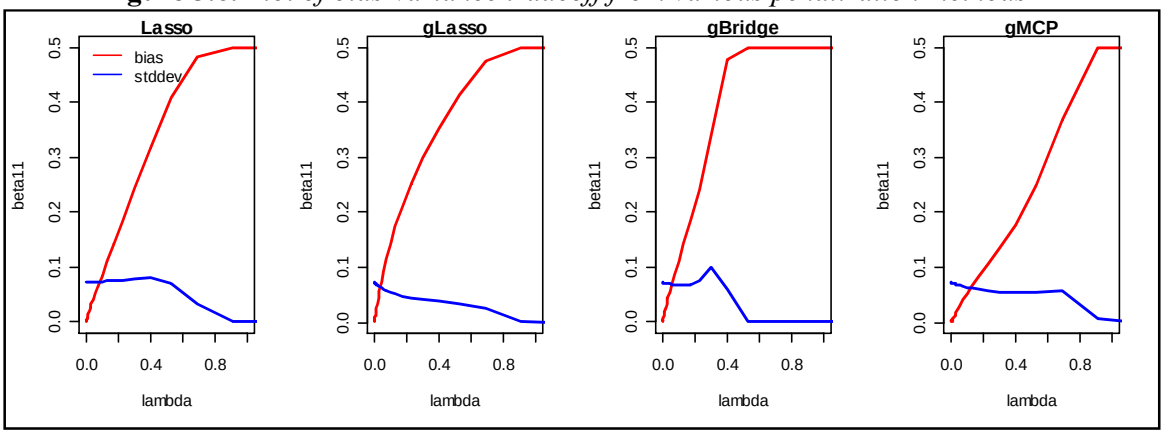
	Regularization Parameter (λ)													
Par.	**0.000**	**0.003**	**0.012**	**0.020**	**0.026**	**0.034**	**0.045**	**0.078**	**0.134**	**0.231**	**0.304**	**0.524**	**0.902**	**1.185**
Lasso														
β_{11}	0	0	0	0	0	0	**0**	0	0	0	**0**	170	798	800
β_{12}	1	37	127	204	259	333	**413**	548	659	752	**793**	800	800	800
β_{13}	0	0	0	0	0	0	**0**	0	0	0	**0**	0	496	800
β_{21}	1	33	107	175	235	312	**394**	561	725	793	**800**	800	800	800
β_{22}	1	30	116	188	242	312	**387**	575	722	797	**799**	800	800	800
β_{23}	3	30	107	196	248	308	**379**	545	729	795	**800**	800	800	800
gLasso														
β_{11}	0	0	0	**0**	0	0	0	0	0	**0**	0	110	800	800
β_{12}	0	0	0	**0**	0	0	0	0	0	**0**	0	110	800	800
β_{13}	0	0	0	**0**	0	0	0	0	0	**0**	0	110	800	800
β_{21}	0	0	13	**34**	67	151	276	552	757	**800**	800	800	800	800
β_{22}	0	0	13	**34**	67	151	276	552	757	**800**	800	800	800	800
β_{23}	0	0	13	**34**	67	151	276	552	757	**800**	800	800	800	800
gBridge														
β_{11}	0	0	0	0	**0**	0	0	0	**0**	52	422	800	800	800
β_{12}	0	30	114	170	**224**	286	340	410	**504**	689	777	800	800	800
β_{13}	0	0	0	0	**0**	0	0	0	**0**	43	413	800	800	800
β_{21}	0	128	506	670	**732**	772	788	800	**800**	800	800	800	800	800
β_{22}	0	129	517	678	**734**	767	785	800	**800**	800	800	800	800	800
β_{23}	0	122	497	664	**727**	771	788	800	**800**	800	800	800	800	800
gMCP														
β_{11}	0	0	0	0	0	**0**	0	0	0	0	0	**0**	798	800
β_{12}	0	0	0	1	2	**4**	6	19	50	114	172	**511**	800	800
β_{13}	0	0	0	0	0	**0**	0	0	0	0	0	**0**	482	800
β_{21}	0	0	0	2	7	**13**	18	45	164	452	623	**800**	800	800
β_{22}	0	0	1	5	4	**6**	14	54	154	449	648	**800**	800	800
β_{23}	0	0	3	4	3	**6**	18	53	171	439	641	**800**	800	800

3.5. Bias-Variance Tradeoff

Prediction accuracy can sometimes be improved by shrinking or setting some coefficients to zero. By doing so, a little bit of bias is sacrificed to reduce the variance of the predicted values, and hence may improve the overall prediction accuracy (Hastie *et al.*, 2009). In this section, the bias and variance tradeoff obtained from the penalization methods are explored and plotted in Figure 3.6. The simulated data from scenario III with 50 subjects is used for investigation. These figures show the bias-variance tradeoff when the correlation within lagged covariate is 0.5 and other figures can be seen in Figure B14 – B17 (Appendix) for different values of correlation. Standard deviation is plotted instead of variance in order to have a clear picture of its evolution across the values of regularization parameter. Further, the bias is shown in terms of its magnitude. As can be expected, the penalization methods yield biased estimates. The bias increases with increasing penalty and the standard deviation decreases with increasing penalty. The group bridge seems to produce more bias with small λ while the opposite phenomenon is observed for the group MCP. A similar picture is observed for different values of correlation for each penalization method but the evolution of standard deviation is a little bit higher as the correlation increases. In addition, Table A12 (Appendix) provides the resulting bias obtained from the model selected by the optimal λ. In

general, it is found that the bias produced by the group MCP is smaller compare to the others. This confirms the results in Figure 3.4. Furthermore, the bias increases as the lagged covariates are more correlated within a group. The negative sign of the bias shows that the estimated coefficients resulting from the penalization methods are smaller than the true parameters. When the true coefficients are zero (β_{12}, β_{21}, β_{22}, β_{23}), we observe that the bias is not significant.

Figure 3.6. *Plot of bias-variance tradeoff from various penalization methods*



3.6. Sensitivity Analysis

In practice, selection of tuning parameter (γ) for group bridge is not easy (Fu, 2002). The results shown in previous sections are obtained by taking γ fixed for group bridge as well as for group MCP. This section is devoted to conduct a small sensitivity analysis to see whether different values of γ can improve the performance of the fitted model in terms of model error. The same data generating model as in Section 3.4 is used and the path of model error as a function of γ is plotted in Figure 3.7 and 3.8 for group bridge and group MCP, respectively. Different values of ρ are taken into account. It can be seen that the optimal γ for group bridge is 0.3. However, it does not tremendously affect the fitted model. Only a very small decline of the estimated model error is observed from $\gamma = 0.5$ to $\gamma = 0.3$. On the other hand, a monotonically decreasing function is observed for group MCP.

Figure 3.7. *Tuning parameter path of group bridge*

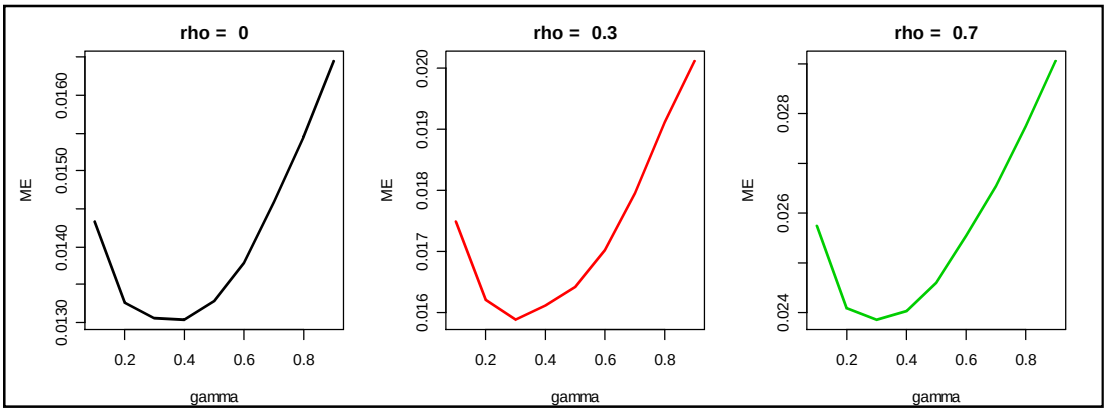
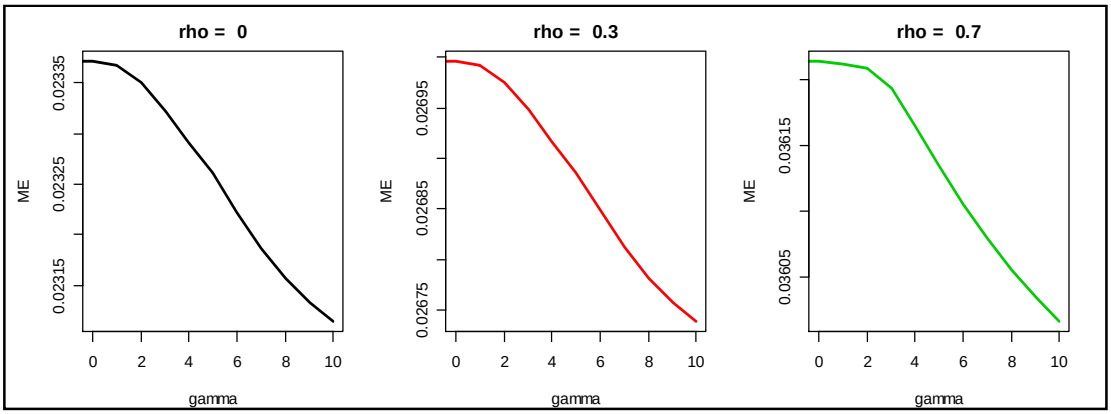


Figure 3.8. *Tuning parameter path of group MCP*



3.7. Results of Estimation Using B-splines

Scenario VII is used to investigate the proposed method of combining the B-splines method with the group Lasso approach. The simulated data was generated with $\rho = 0.7$ as the correlation within covariate. Figure 3.9 shows the estimated prediction error curves plotted as a function of regularization parameter. Different numbers of b-splines basis functions (df) are considered. It can be observed that the path of prediction error is monotonically increasing. It leads to a smaller prediction error as the number of basis functions is increased. The parsimonious model selected is the one without applying any penalty since $\lambda = 0$.

Figure 3.9. *Estimated prediction error curves for the group Lasso method*

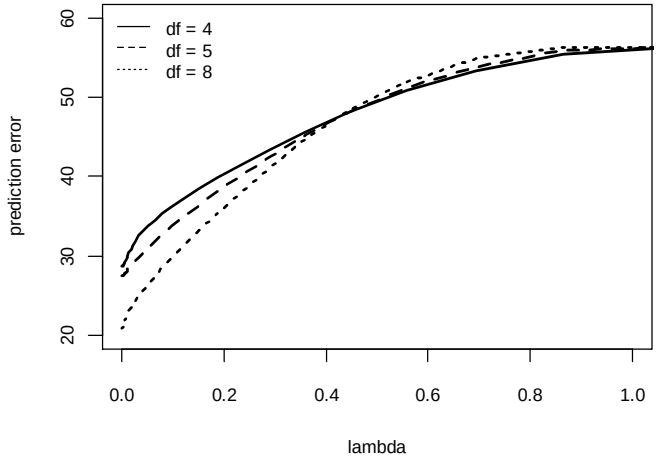
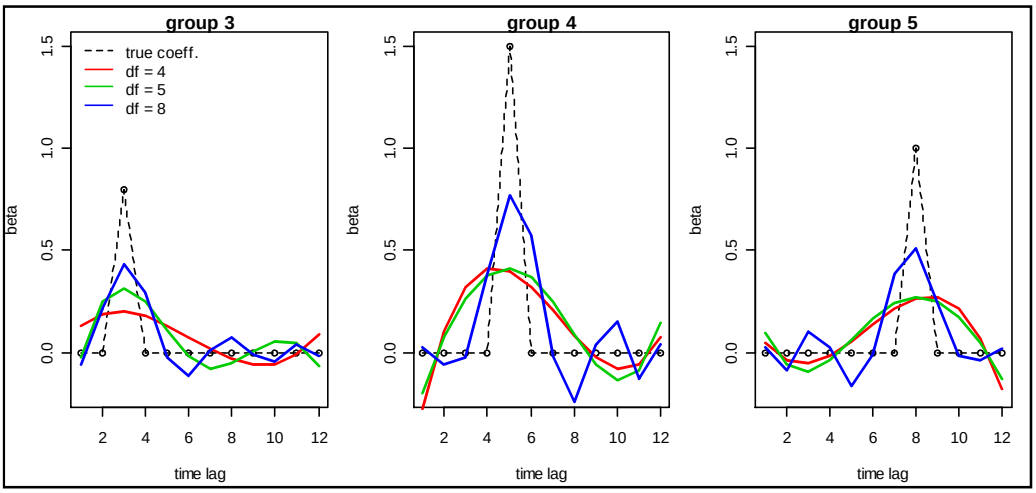


Table 3.3 presents the estimated prediction error selected by the optimal regularization parameter. It was calculated based on training and test set, where the test set contains 30 subjects. It shows that the training error is a good estimate of prediction error as its estimated value is closer to the test error. Figure 3.10 depicts the true and estimated lagged coefficients for the nonzero groups for each lagged time point. These estimates cannot fully capture the trend of the true coefficients especially for the peak of the curve. However, the splines fit looks quite good. Higher order of basis functions might lead to a better approximation of the true structure of the parameters.

Table 3.3. *Prediction error (standard error)*

	df = 4	**df = 5**	**df = 8**
training set	28.82(0.0396)	27.51(0.0373)	21.02(0.0287)
test set	28.78(0.0727)	27.58(0.0694)	21.23(0.0519)

Figure 3.10. *True and estimated lagged coefficients using the group Lasso method*



Chapter 4

Discussion and Conclusion

4.1.Summary of Findings

Penalization method is an attractive statistical tool that has received much attention in recent literature. It has been widely used in linear model for variable selection. In this thesis, penalization method is utilized to longitudinal data with time dependent covariates to select relevant variables as well as to select the correct lag for each variable. Pepe and Anderson (1994) have pointed out that the consistency of GEE is not assured with arbitrary working correlation structure unless the FCCM assumption is satisfied. In genaral, to verify this assumption is not easy, therefore, it is suggested to use independent working correlation when using GEE with time-dependent covariates. This idea is adopted in this thesis for bi-level variable selection in time-dependent covariates so that the common penalization methods still can be applied. Beside, assuming another working correlation aside from independent need extra programming and is outside the scope of this thesis.

Model selection method using Lasso, group Lasso, group bridge, and group MCP were explored for time dependent covariates data. The group Lasso lacks the ability to do variable selection at the individual level and heavily shrinks large coefficients. The group bridge and group MCP are very appealing from the perspective of performing at group and individual level. Most of these methods are not consistent in selecting the correct model except the group bridge. Table 4.1 summarizes the comparison between these methods.

Table 4.1. *Some characteristics of different penalization methods. Key:* √ *= yes;* O *= moderate;* × *= no*

Characteristic	Lasso	Group Lasso	Group Bridge	Group MCP
Encouraging sparsity at the individual level	√	×	√	√
Encouraging sparsity at the group level	×	√	√	√
Bi-level variable selection	×	×	√	√
Heavily shrinks large coefficients	×	√	×	×
Model selection consistency	×	×	√	×
Oracle property	×	×	√	√
Ability to deal with multicollinearity	O	O	O	O

The model error is utilized as a criterion for selecting the optimal regularization parameter. In general, all these methods may potentially improve the estimation and prediction when collinearity is present since the model error is smaller than the no-penalty GEE. It is found that the performance of the group bridge is superior to the other methods in terms of variable selection at both group and individual level. However, its performance is still poor at the individual level. The group Lasso tends to select more groups and variables than there actually are in the generating models. The group MCP performs well at the individual level when there are a larger number of rather sparse groups although the number of correct deletion is still small. In addition, these methods do not consistently select the correct coefficients in the selected model. An alternative method is proposed by approximating the lagged coefficients with B-spline basis functions and then model selection is performed by the group Lasso. This method did not perform variable selection since the optimal λ selected is 0. However, the true coefficients can be well approximated by the proposed method.

4.2.Challenges for Future Research

An important area of future research is to take into account the correct working correlation structure in penalized GEE. It may lead to better performance of the penalization methods. We observed that the optimal regularization parameter selected by model error is very small and fails to yield a sufficiently sparse solution via simulation studies. Thus, different criteria to choose the optimal regularization parameter need to be adopted. Furthermore, it might be interesting to extend our proposed method of approximating the lagged coefficients with other basis functions or smoothing splines that could provide better alternatives.

Bibliography

Breheny, P. (2009). Regularization path for regression models with grouped covariates, package 'grpreg'.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, **2**: 369 – 480.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**(6): 2350 – 2383.

De Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer.

Diggle P. J., Heagerty, P., Liang, K.Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data, Second Edition*. New York: Oxford University Press.

Dziak, J.J. (2006). *Penalized Quadratic Inference Functions for Variable Selection in Longitudinal Research*, Dissertation. The Pennsylvania State University.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456): 1348 – 1360.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, **35**: 109 – 148.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso.

Fu, W. J. (2002). Nonlinear GCV and quasi-GCV for shrinkage models. Unpublished manuscript.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics*, **59**: 126 – 132.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, Second Edition*. New York: Springer.

Huang, J., Ma, S., Xie, H., and Zhang, C.H. (2007). A group bridge approach for variable selection. Technical Report #376, Department of Statistics and Actuarial Science, University of Iowa.

Keele, L. J. (2008). *Semiparametric Regression for the Social Science*. Chichester, UK: Wiley & Son.

Lai, T.L. and Small, D. (2006). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalised method of moment approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **69**:79 – 99.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models, *Biometrics*. **73**: 13 – 22.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Park, C. And Yoon, Y.J. (2011). Bridge regression: adaptivity and group selection. Accepted for publication in *Journal of Statistical Planning and Inference*.

Pepe, M.S. and Anderson, G.L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics, Part B – Simulation and Computation*, **23**: 939 – 951.

Ruppert, D., Wand M. P., and Carroll R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

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

Verbeke, G. and Molenberghs, G. (2002). *Linear Mixced Models for Longitudinal Data*. New York: Springer.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **68**: 49 – 67.

Zhang, C. H. (2007). Penalized linear unbiased selection. Technical Report #2007-003, Department of Statistics and Biostatistics, Rutgers University.

Appendix A

Table A1. *Model error (standard error) for scenario I and II*

			N = 20		*N* = 50		*N* = 100	
Scenario	**ρ**	**method**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**
I	0	Lasso	0.070	0.0203(0.0009)	0.041	0.0083(0.0004)	0.024	0.0041(0.0002)
		GEE	0	0.0295(0.0010)	0	0.0118(0.0004)	0	0.0060(0.0002)
	0.1	Lasso	0.070	0.0224(0.0009)	0.041	0.0082(0.0003)	0.024	0.0042(0.0002)
		GEE	0	0.0319(0.0011)	0	0.0119(0.0004)	0	0.0060(0.0002)
	0.3	Lasso	0.070	0.0243(0.0010)	0.041	0.0099(0.0004)	0.024	0.0045(0.0002)
		GEE	0	0.0342(0.0012)	0	0.0139(0.0005)	0	0.0064(0.0002)
	0.5	Lasso	0.070	0.0286(0.0012)	0.041	0.0119(0.0005)	0.024	0.0059(0.0002)
		GEE	0	0.0376(0.0013)	0	0.0155(0.0006)	0	0.0076(0.0003)
	0.7	Lasso	0.041	0.0373(0.0016)	0.041	0.0138(0.0006)	0.024	0.0074(0.0003)
		GEE	0	0.0476(0.0018)	0	0.0175(0.0006)	0	0.0091(0.0003)
	0.9	Lasso	0.041	0.0525(0.0023)	0.024	0.0210(0.0009)	0.014	0.0100(0.0005)
		GEE	0	0.0637(0.0025)	0	0.0250(0.0009)	0	0.0119(0.0005)
II	0	Lasso	0.038	0.0532(0.0017)	0.027	0.0209(0.0006)	0.020	0.0099(0.0003)
		GEE	0	0.0630(0.0019)	0	0.0251(0.0007)	0	0.0118(0.0003)
	0.1	Lasso	0.038	0.0505(0.0015)	0.027	0.0208(0.0006)	0.020	0.0106(0.0003)
		GEE	0	0.0614(0.0017)	0	0.0246(0.0006)	0	0.0124(0.0004)
	0.3	Lasso	0.038	0.0570(0.0018)	0.020	0.0226(0.0007)	0.014	0.0109(0.0003)
		GEE	0	0.0656(0.0019)	0	0.0262(0.0007)	0	0.0124(0.0003)
	0.5	Lasso	0.027	0.0639(0.0020)	0.020	0.0253(0.0008)	0.014	0.0123(0.0004)
		GEE	0	0.0712(0.0021)	0	0.0286(0.0008)	0	0.0137(0.0004)
	0.7	Lasso	0.020	0.0686(0.0021)	0.014	0.0296(0.0009)	0.010	0.0147(0.0004)
		GEE	0	0.0753(0.0022)	0	0.0322(0.0009)	0	0.0160(0.0005)
	0.9	Lasso	0.014	0.0845(0.0027)	0.007	0.0352(0.0010)	0.005	0.0168(0.0005)
		GEE	0	0.0923(0.0029)	0	0.0374(0.0010)	0	0.0180(0.0006)

Table A2. *Model complexity of the selected model in terms of correct deletion, erroneous deletion, and proportion of correct model for scenario I and II*

		N = 20			N = 50			N = 100		
Scenario	**ρ**	**C**	**E**	**P**	**C**	**E**	**P**	**C**	**E**	**P**
I	0	0.99	0.00	0.25	0.92	0.00	0.23	0.78	0.00	0.18
	0.1	0.98	0.00	0.27	0.92	0.00	0.23	0.79	0.00	0.16
	0.3	0.97	0.00	0.25	0.88	0.00	0.21	0.78	0.00	0.15
	0.5	1.06	0.00	0.30	0.97	0.00	0.25	0.88	0.00	0.18
	0.7	0.79	0.00	0.16	1.03	0.00	0.28	0.91	0.00	0.21
	0.9	1.00	0.05	0.26	0.93	0.00	0.20	0.87	0.00	0.18
II	0	0.85	0.00	0.03	0.92	0.00	0.04	0.98	0.00	0.04
	0.1	0.82	0.00	0.01	0.94	0.00	0.03	0.97	0.00	0.04
	0.3	0.93	0.00	0.03	0.77	0.00	0.01	0.74	0.00	0.02
	0.5	0.80	0.00	0.02	0.82	0.00	0.02	0.86	0.00	0.03
	0.7	0.72	0.00	0.01	0.77	0.00	0.02	0.74	0.00	0.02
	0.9	0.86	0.09	0.02	0.71	0.00	0.01	0.70	0.00	0.02

Table A3. *Model error (standard error) for scenario III*

		N=20		N=50		N=100	
ρ	**method**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**
0	Lasso	0.059	0.0431(0.0014)	0.045	0.0161(0.0005)	0.026	0.0085(0.0003)
	gLasso	0.034	0.0500(0.0015)	0.020	0.0197(0.0005)	0.015	0.0104(0.0003)
	gBridge	0.045	0.0371(0.0014)	0.026	0.0133(0.0005)	0.015	0.0066(0.0002)
	gMCP	0.134	0.0576(0.0015)	0.034	0.0233(0.0005)	0.012	0.0121(0.0003)
	GEE	0	0.0649(0.0017)	0	0.0242(0.0005)	0	0.0122(0.0003)
0.1	Lasso	0.059	0.0472(0.0015)	0.034	0.0176(0.0006)	0.026	0.0085(0.0003)
	gLasso	0.034	0.0530(0.0015)	0.020	0.0209(0.0006)	0.015	0.0104(0.0003)
	gBridge	0.045	0.0406(0.0014)	0.026	0.0146(0.0005)	0.015	0.0069(0.0003)
	gMCP	0.102	0.0605(0.0016)	0.026	0.0243(0.0006)	0.015	0.0121(0.0003)
	GEE	0	0.0648(0.0017)	0	0.0248(0.0006)	0	0.0123(0.0003)
0.3	Lasso	0.059	0.0491(0.0015)	0.045	0.0198(0.0006)	0.026	0.0098(0.0003)
	gLasso	0.034	0.0551(0.0015)	0.020	0.0233(0.0006)	0.015	0.0117(0.0003)
	gBridge	0.045	0.0420(0.0014)	0.026	0.0164(0.0006)	0.020	0.0079(0.0003)
	gMCP	0.134	0.0626(0.0016)	0.034	0.0269(0.0007)	0.015	0.0134(0.0003)
	GEE	0	0.0686(0.0017)	0	0.0278(0.0007)	0	0.0136(0.0003)
0.5	Lasso	0.059	0.0592(0.0020)	0.045	0.0214(0.0008)	0.034	0.0114(0.0004)
	gLasso	0.034	0.0653(0.0020)	0.020	0.0248(0.0008)	0.012	0.0133(0.0004)
	gBridge	0.045	0.0521(0.0019)	0.026	0.0176(0.0007)	0.020	0.0092(0.0003)
	gMCP	0.134	0.0725(0.0020)	0.034	0.0284(0.0008)	0.015	0.0151(0.0004)
	GEE	0	0.0802(0.0022)	0	0.0296(0.0008)	0	0.0154(0.0004)
0.7	Lasso	0.078	0.0669(0.0023)	0.045	0.0285(0.0010)	0.034	0.0133(0.0004)
	gLasso	0.034	0.0728(0.0022)	0.020	0.0326(0.0010)	0.012	0.0159(0.0004)
	gBridge	0.059	0.0589(0.0022)	0.034	0.0246(0.0009)	0.020	0.0112(0.0004)
	gMCP	0.176	0.0792(0.0022)	0.034	0.0362(0.0010)	0.015	0.0177(0.0005)
	GEE	0	0.0899(0.0024)	0	0.0374(0.0010)	0	0.0181(0.0005)
0.9	Lasso	0.102	0.0936(0.0034)	0.059	0.0372(0.0013)	0.045	0.0184(0.0006)
	gLasso	0.059	0.0958(0.0033)	0.020	0.0410(0.0012)	0.012	0.0216(0.0006)
	gBridge	0.078	0.0816(0.0034)	0.045	0.0321(0.0012)	0.026	0.0160(0.0006)
	gMCP	0.231	0.0985(0.0033)	0.045	0.0452(0.0013)	0.020	0.0237(0.0007)
	GEE	0	0.1282(0.0036)	0	0.0487(0.0013)	0	0.0244(0.0007)

Table A4. *Model error (standard error) for scenario IV*

		N=20		N=50		N=100	
ρ	**method**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**
0	Lasso	0.038	0.1069(0.0028)	0.029	0.0417(0.0010)	0.016	0.0199(0.0005)
	gLasso	0.029	0.1068(0.0027)	0.016	0.0428(0.0010)	0.009	0.0206(0.0005)
	gBridge	0.038	0.0847(0.0027)	0.022	0.0315(0.0010)	0.012	0.0145(0.0004)
	gMCP	0.029	0.1239(0.0026)	0.016	0.0498(0.0010)	0.005	0.0235(0.0005)
	GEE	0	0.1287(0.0027)	0	0.0508(0.0010)	0	0.0238(0.0005)
0.1	Lasso	0.050	0.1053(0.0027)	0.029	0.0406(0.0010)	0.022	0.0207(0.0005)
	gLasso	0.029	0.1040(0.0025)	0.016	0.0416(0.0010)	0.012	0.0215(0.0005)
	gBridge	0.038	0.0813(0.0024)	0.022	0.0305(0.0009)	0.012	0.0150(0.0004)
	gMCP	0.038	0.1229(0.0025)	0.016	0.0486(0.0009)	0.004	0.0245(0.0005)
	GEE	0	0.1288(0.0028)	0	0.0492(0.0009)	0	0.0246(0.0005)
0.3	Lasso	0.050	0.1086(0.0026)	0.038	0.0414(0.0011)	0.022	0.0197(0.0005)
	gLasso	0.038	0.1070(0.0028)	0.022	0.0421(0.0010)	0.016	0.0204(0.0005)
	gBridge	0.038	0.0839(0.0025)	0.029	0.0312(0.0009)	0.016	0.0142(0.0004)
	gMCP	0.050	0.1298(0.0028)	0.022	0.0506(0.0010)	0.009	0.0243(0.0005)
	GEE	0	0.1366(0.0028)	0	0.0517(0.0010)	0	0.0245(0.0005)
0.5	Lasso	0.066	0.1163(0.0033)	0.050	0.0436(0.0012)	0.029	0.0215(0.0005)
	gLasso	0.038	0.1129(0.0032)	0.029	0.0436(0.0011)	0.022	0.0220(0.0006)
	gBridge	0.050	0.0917(0.0031)	0.029	0.0327(0.0010)	0.022	0.0159(0.0005)
	gMCP	0.066	0.1378(0.0032)	0.038	0.0547(0.0012)	0.022	0.0274(0.0005)
	GEE	0	0.146(0.0030)	0	0.0570(0.0012)	0	0.0280(0.0005)
0.7	Lasso	0.066	0.1248(0.0034)	0.050	0.0464(0.0012)	0.038	0.0236(0.0006)
	gLasso	0.050	0.1161(0.0034)	0.029	0.0450(0.0012)	0.022	0.0235(0.0006)
	gBridge	0.066	0.0947(0.0032)	0.038	0.0344(0.0011)	0.022	0.0176(0.0005)
	gMCP	0.116	0.1431(0.0037)	0.050	0.0573(0.0013)	0.029	0.0298(0.0006)
	GEE	0	0.1617(0.0035)	0	0.0615(0.0013)	0	0.0310(0.0006)
0.9	Lasso	0.088	0.1502(0.0045)	0.066	0.0563(0.0016)	0.050	0.0269(0.0008)
	gLasso	0.066	0.1332(0.0046)	0.038	0.0513(0.0016)	0.022	0.0253(0.0007)
	gBridge	0.088	0.1090(0.0044)	0.050	0.0405(0.0014)	0.029	0.0195(0.0007)
	gMCP	0.153	0.1587(0.0049)	0.088	0.0641(0.0017)	0.050	0.0321(0.0008)
	GEE	0	0.2049(0.0047)	0	0.0761(0.0017)	0	0.0367(0.0008)

Table A5. *Model error (standard error) for scenario V*

		N=20		N=50		N=100	
ρ	**method**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**
0	Lasso	0.051	0.1899(0.0030)	0.033	0.0710(0.0011)	0.026	0.0359(0.0005)
	gLasso	0.026	0.2149(0.0032)	0.017	0.0836(0.0011)	0.011	0.0430(0.0006)
	gBridge	0.041	0.1576(0.0028)	0.021	0.0576(0.0010)	0.017	0.0286(0.0005)
	gMCP	0.041	0.2539(0.0034)	0.017	0.0975(0.0012)	0.007	0.0494(0.0006)
	GEE	0	0.2680(0.0037)	0	0.0996(0.0012)	0	0.0499(0.0006)
0.1	Lasso	0.051	0.1961(0.0032)	0.041	0.0767(0.0012)	0.026	0.0365(0.0006)
	gLasso	0.026	0.2233(0.0033)	0.017	0.0900(0.0012)	0.011	0.0435(0.0006)
	gBridge	0.041	0.1663(0.0031)	0.026	0.0628(0.0011)	0.017	0.0293(0.0005)
	gMCP	0.041	0.2618(0.0035)	0.017	0.1042(0.0013)	0.007	0.0498(0.0006)
	GEE	0	0.2744(0.0036)	0	0.1060(0.0013)	0	0.0502(0.0006)
0.3	Lasso	0.063	0.2178(0.0037)	0.041	0.0801(0.0013)	0.026	0.0412(0.0006)
	gLasso	0.033	0.2451(0.0038)	0.017	0.0941(0.0014)	0.011	0.0485(0.0007)
	gBridge	0.041	0.1823(0.0034)	0.026	0.0663(0.0012)	0.017	0.0330(0.0006)
	gMCP	0.041	0.2892(0.0041)	0.017	0.1089(0.0014)	0.007	0.0549(0.0007)
	GEE	0	0.3063(0.0044)	0	0.1112(0.0015)	0	0.0554(0.0007)
0.5	Lasso	0.063	0.2500(0.0045)	0.041	0.0963(0.0016)	0.026	0.0477(0.0007)
	gLasso	0.033	0.2783(0.0046)	0.017	0.1102(0.0016)	0.011	0.0555(0.0008)
	gBridge	0.041	0.2161(0.0043)	0.026	0.0801(0.0015)	0.017	0.0386(0.0007)
	gMCP	0.051	0.3224(0.0048)	0.017	0.1250(0.0017)	0.009	0.0625(0.0008)
	GEE	0	0.3424(0.0051)	0	0.1273(0.0017)	0	0.0631(0.0008)
0.7	Lasso	0.051	0.3131(0.0056)	0.033	0.1189(0.0019)	0.026	0.0604(0.0009)
	gLasso	0.026	0.3365(0.0056)	0.017	0.1327(0.0020)	0.011	0.0679(0.0010)
	gBridge	0.041	0.2705(0.0052)	0.026	0.1014(0.0018)	0.021	0.0498(0.0009)
	gMCP	0.051	0.3825(0.0059)	0.017	0.1486(0.0020)	0.011	0.0756(0.0010)
	GEE	0	0.4100(0.0063)	0	0.1524(0.0021)	0	0.0768(0.0010)
0.9	Lasso	0.041	0.4802(0.0081)	0.021	0.1757(0.0026)	0.017	0.0855(0.0014)
	gLasso	0.026	0.4917(0.0084)	0.011	0.1846(0.0027)	0.007	0.0915(0.0014)
	gBridge	0.033	0.4368(0.0082)	0.021	0.1550(0.0026)	0.017	0.0744(0.0013)
	gMCP	0.051	0.5362(0.0089)	0.021	0.2005(0.0028)	0.011	0.0991(0.0014)
	GEE	0	0.5846(0.0091)	0	0.2091(0.0028)	0	0.1015(0.0014)

Table A6. *Model error (standard error) for scenario VI*

		N=20		N=50		N=100	
ρ	**method**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**	**λ**	**ME(s.e.)**
0	Lasso	0.078	0.5207(0.0064)	0.051	0.1777(0.0021)	0.041	0.0868(0.0010)
	gLasso	0.041	0.7749(0.0078)	0.021	0.2780(0.0026)	0.017	0.1367(0.0013)
	gBridge	0.041	0.4002(0.0057)	0.026	0.1391(0.0018)	0.017	0.0677(0.0009)
	gMCP	0.234	0.9247(0.0082)	0.041	0.3571(0.0029)	0.017	0.1708(0.0014)
	GEE	0	1.3523(0.0134)	0	0.3754(0.0031)	0	0.1744(0.0015)
0.1	Lasso	0.078	0.5291(0.0063)	0.051	0.1851(0.0021)	0.041	0.0908(0.0011)
	gLasso	0.041	0.7743(0.0075)	0.021	0.2798(0.0026)	0.017	0.1377(0.0013)
	gBridge	0.041	0.4015(0.0053)	0.026	0.1422(0.0019)	0.017	0.0694(0.0009)
	gMCP	0.234	0.9213(0.0077)	0.041	0.3578(0.0029)	0.017	0.1712(0.0014)
	GEE	0	1.3252(0.0133)	0	0.3748(0.0031)	0	0.1745(0.0015)
0.3	Lasso	0.063	0.6164(0.0075)	0.041	0.2144(0.0024)	0.033	0.1048(0.0012)
	gLasso	0.041	0.8218(0.0088)	0.021	0.2928(0.0028)	0.014	0.1451(0.0014)
	gBridge	0.041	0.4498(0.0065)	0.026	0.1559(0.0022)	0.017	0.0762(0.0011)
	gMCP	0.234	0.9696(0.0089)	0.041	0.3736(0.003)	0.017	0.1793(0.0015)
	GEE	0	1.4033(0.0139)	0	0.3913(0.0032)	0	0.1827(0.0015)
0.5	Lasso	0.063	0.7511(0.0087)	0.033	0.2601(0.0029)	0.026	0.1280(0.0014)
	gLasso	0.041	0.8888(0.0100)	0.021	0.3205(0.0031)	0.014	0.1598(0.0016)
	gBridge	0.041	0.5324(0.0077)	0.026	0.1844(0.0025)	0.017	0.0894(0.0012)
	gMCP	0.188	1.0554(0.0102)	0.041	0.4110(0.0034)	0.017	0.1986(0.0017)
	GEE	0	1.5674(0.0153)	0	0.4355(0.0035)	0	0.2032(0.0017)
0.7	Lasso	0.041	0.9438(0.0108)	0.021	0.3278(0.0036)	0.017	0.1599(0.0017)
	gLasso	0.033	0.9812(0.0108)	0.017	0.3676(0.0036)	0.011	0.1856(0.0018)
	gBridge	0.033	0.6856(0.0094)	0.021	0.2394(0.0032)	0.014	0.1149(0.0015)
	gMCP	0.188	1.1781(0.0113)	0.041	0.4765(0.0039)	0.017	0.2339(0.0019)
	GEE	0	1.8521(0.0166)	0	0.5249(0.0042)	0	0.2436(0.0019)
0.9	Lasso	0.098	1.0937(0.0106)	0.011	0.4321(0.0048)	0.007	0.2081(0.0021)
	gLasso	0.026	1.0898(0.0147)	0.011	0.4260(0.0045)	0.007	0.2210(0.0020)
	gBridge	0.026	0.9207(0.0134)	0.011	0.3422(0.0044)	0.007	0.1650(0.0020)
	gMCP	0.188	1.2427(0.0144)	0.041	0.5411(0.0047)	0.017	0.2841(0.0022)
	GEE	0	2.5098(0.0244)	0	0.7077(0.0057)	0	0.3250(0.0024)

Table A7. *Model complexity of the selected model in terms of correct deletion, erroneous deletion, and proportion of correct model for scenario III*

		N =20					N = 50					N = 100				
		group		variable			group		variable			group		variable		
ρ	method	C	E	C	E	P	C	E	C	E	P	C	E	C	E	P
0	Lasso	0.40	0.00	1.61	0.00	0.04	0.46	0.00	1.93	0.00	0.06	0.37	0.00	1.60	0.00	0.03
	gLasso	0.05	0.00	0.15	0.00	0.00	0.04	0.00	0.12	0.00	0.00	0.06	0.00	0.18	0.00	0.00
	gBridge	0.93	0.00	3.04	0.00	0.24	0.95	0.00	3.12	0.00	0.26	0.94	0.00	3.03	0.00	0.19
	gMCP	0.11	0.00	0.41	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	Lasso	0.40	0.00	1.69	0.00	0.03	0.37	0.00	1.53	0.00	0.03	0.41	0.00	1.69	0.00	0.04
	gLasso	0.06	0.00	0.17	0.00	0.00	0.05	0.00	0.15	0.00	0.00	0.06	0.00	0.18	0.00	0.00
	gBridge	0.91	0.00	3.05	0.00	0.27	0.94	0.00	3.09	0.00	0.25	0.94	0.00	3.02	0.00	0.19
	gMCP	0.06	0.00	0.23	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.3	Lasso	0.41	0.00	1.73	0.00	0.03	0.47	0.00	1.97	0.00	0.06	0.38	0.00	1.70	0.00	0.04
	gLasso	0.07	0.00	0.21	0.00	0.00	0.04	0.00	0.13	0.00	0.00	0.06	0.00	0.19	0.00	0.00
	gBridge	0.89	0.00	2.98	0.00	0.24	0.91	0.00	3.02	0.00	0.25	0.96	0.00	3.19	0.00	0.28
	gMCP	0.11	0.00	0.44	0.00	0.00	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.5	Lasso	0.44	0.00	1.84	0.00	0.04	0.46	0.00	2.03	0.00	0.07	0.49	0.00	2.06	0.00	0.09
	gLasso	0.07	0.00	0.22	0.00	0.00	0.05	0.00	0.14	0.00	0.00	0.02	0.00	0.07	0.00	0.00
	gBridge	0.85	0.00	2.89	0.00	0.25	0.90	0.00	3.03	0.00	0.27	0.93	0.00	3.14	0.00	0.32
	gMCP	0.14	0.00	0.50	0.00	0.00	0.01	0.00	0.04	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.7	Lasso	0.51	0.00	2.16	0.00	0.09	0.46	0.00	2.02	0.00	0.07	0.50	0.00	2.07	0.00	0.09
	gLasso	0.07	0.00	0.20	0.00	0.00	0.05	0.00	0.16	0.00	0.00	0.04	0.00	0.12	0.00	0.00
	gBridge	0.90	0.00	3.01	0.00	0.25	0.92	0.00	3.05	0.00	0.24	0.89	0.00	3.01	0.00	0.25
	gMCP	0.24	0.00	0.89	0.00	0.00	0.01	0.00	0.06	0.00	0.00	0.00	0.00	0.01	0.00	0.00
0.9	Lasso	0.60	0.00	2.39	0.03	0.12	0.57	0.00	2.26	0.00	0.08	0.59	0.00	2.28	0.00	0.09
	gLasso	0.21	0.00	0.63	0.00	0.00	0.08	0.00	0.24	0.00	0.00	0.05	0.00	0.15	0.00	0.00
	gBridge	0.90	0.00	2.80	0.01	0.07	0.93	0.00	2.88	0.00	0.06	0.90	0.00	2.81	0.00	0.07
	gMCP	0.39	0.00	1.36	0.00	0.00	0.02	0.00	0.13	0.00	0.00	0.01	0.00	0.04	0.00	0.00

Table A8. *Model complexity of the selected model in terms of correct deletion, erroneous deletion, and proportion of correct model for scenario IV*

		N =20					N = 50					N = 100				
		group		variable			group		variable			group		variable		
ρ	method	C	E	C	E	P	C	E	C	E	P	C	E	C	E	P
0	Lasso	0.26	0.00	1.67	0.00	0.00	0.32	0.00	1.93	0.00	0.00	0.24	0.00	1.62	0.00	0.00
	gLasso	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
	gBridge	0.89	0.00	5.38	0.00	0.85	0.92	0.00	5.59	0.00	0.90	0.93	0.00	5.60	0.00	0.90
	gMCP	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	Lasso	0.34	0.00	2.22	0.00	0.01	0.33	0.00	2.03	0.00	0.00	0.32	0.00	2.04	0.00	0.00
	gLasso	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00
	gBridge	0.90	0.00	5.46	0.00	0.87	0.93	0.00	5.63	0.00	0.91	0.91	0.00	5.48	0.00	0.86
	gMCP	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.3	Lasso	0.36	0.00	2.31	0.00	0.01	0.41	0.00	2.68	0.00	0.01	0.34	0.00	2.26	0.00	0.00
	gLasso	0.02	0.00	0.11	0.00	0.02	0.01	0.00	0.05	0.00	0.01	0.02	0.00	0.12	0.00	0.02
	gBridge	0.86	0.00	5.26	0.00	0.81	0.95	0.00	5.74	0.00	0.94	0.96	0.00	5.75	0.00	0.93
	gMCP	0.01	0.00	0.07	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.5	Lasso	0.46	0.00	3.11	0.00	0.03	0.48	0.00	3.28	0.00	0.03	0.40	0.00	2.96	0.00	0.02
	gLasso	0.02	0.00	0.14	0.00	0.02	0.05	0.00	0.29	0.00	0.05	0.06	0.00	0.37	0.00	0.06
	gBridge	0.92	0.00	5.59	0.00	0.89	0.93	0.00	5.64	0.00	0.90	0.97	0.00	5.87	0.00	0.96
	gMCP	0.02	0.00	0.17	0.00	0.00	0.01	0.00	0.07	0.00	0.00	0.00	0.00	0.03	0.00	0.00
0.7	Lasso	0.46	0.00	3.43	0.00	0.03	0.48	0.00	3.64	0.00	0.06	0.52	0.00	3.75	0.00	0.08
	gLasso	0.10	0.00	0.62	0.00	0.10	0.06	0.00	0.36	0.00	0.06	0.08	0.00	0.49	0.00	0.08
	gBridge	0.96	0.00	5.78	0.00	0.94	0.96	0.00	5.81	0.00	0.94	0.95	0.00	5.79	0.00	0.94
	gMCP	0.11	0.00	0.88	0.00	0.00	0.02	0.00	0.24	0.00	0.00	0.01	0.00	0.10	0.00	0.00
0.9	Lasso	0.56	0.00	4.20	0.03	0.12	0.61	0.00	4.40	0.00	0.14	0.61	0.00	4.51	0.00	0.19
	gLasso	0.20	0.00	1.18	0.00	0.20	0.15	0.00	0.87	0.00	0.15	0.10	0.00	0.57	0.00	0.10
	gBridge	0.95	0.00	5.79	0.00	0.93	0.97	0.00	5.86	0.00	0.95	0.97	0.00	5.87	0.00	0.95
	gMCP	0.22	0.00	2.13	0.00	0.00	0.13	0.00	1.30	0.00	0.00	0.06	0.00	0.53	0.00	0.00

Table A9. *Model complexity of the selected model in terms of correct deletion, erroneous deletion, and proportion of correct model for scenario V*

		N =20					N = 50					N = 100				
		group		variable			group		variable			group		variable		
ρ	method	C	E	C	E	P	C	E	C	E	P	C	E	C	E	P
0	Lasso	1.53	0.00	5.82	0.00	0.00	1.46	0.00	5.73	0.00	0.00	1.63	0.00	6.28	0.00	0.00
	gLasso	0.15	0.00	0.44	0.00	0.00	0.11	0.00	0.34	0.00	0.00	0.08	0.00	0.23	0.00	0.00
	gBridge	3.61	0.00	11.73	0.00	0.02	3.61	0.00	11.56	0.00	0.01	3.86	0.00	12.35	0.00	0.01
	gMCP	0.02	0.00	0.12	0.00	0.02	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.01	0.00	0.01
0.1	Lasso	1.50	0.00	5.81	0.00	0.00	1.78	0.00	6.93	0.00	0.00	1.61	0.00	6.42	0.00	0.00
	gLasso	0.13	0.00	0.39	0.00	0.00	0.13	0.00	0.40	0.00	0.00	0.09	0.00	0.28	0.00	0.00
	gBridge	3.58	0.00	11.59	0.00	0.01	3.74	0.00	12.15	0.00	0.02	3.81	0.00	12.23	0.00	0.02
	gMCP	0.02	0.00	0.12	0.00	0.01	0.00	0.00	0.03	0.00	0.02	0.00	0.00	0.01	0.00	0.02
0.3	Lasso	1.76	0.00	7.01	0.00	0.00	1.75	0.00	6.99	0.00	0.00	1.61	0.00	6.44	0.00	0.00
	gLasso	0.23	0.00	0.68	0.00	0.00	0.13	0.00	0.38	0.00	0.00	0.10	0.00	0.31	0.00	0.00
	gBridge	3.46	0.00	11.35	0.00	0.01	3.69	0.00	12.01	0.00	0.01	3.74	0.00	12.05	0.00	0.02
	gMCP	0.02	0.00	0.12	0.00	0.01	0.01	0.00	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.02
0.5	Lasso	1.77	0.00	7.32	0.00	0.00	1.74	0.00	7.24	0.00	0.00	1.57	0.00	6.58	0.00	0.00
	gLasso	0.24	0.00	0.71	0.00	0.00	0.15	0.00	0.45	0.00	0.00	0.10	0.00	0.29	0.00	0.00
	gBridge	3.35	0.00	11.13	0.00	0.01	3.55	0.00	11.66	0.00	0.01	3.59	0.00	11.69	0.00	0.01
	gMCP	0.04	0.00	0.20	0.00	0.01	0.01	0.00	0.02	0.00	0.01	0.00	0.00	0.01	0.00	0.01
0.7	Lasso	1.64	0.00	6.98	0.00	0.00	1.60	0.00	6.83	0.00	0.00	1.67	0.00	7.21	0.00	0.00
	gLasso	0.19	0.00	0.58	0.00	0.00	0.15	0.00	0.45	0.00	0.00	0.09	0.00	0.27	0.00	0.00
	gBridge	3.22	0.00	10.89	0.00	0.01	3.43	0.00	11.45	0.00	0.01	3.64	0.00	12.12	0.00	0.03
	gMCP	0.06	0.00	0.29	0.00	0.01	0.01	0.00	0.05	0.00	0.01	0.00	0.00	0.02	0.00	0.03
0.9	Lasso	1.71	0.01	7.75	0.12	0.00	1.53	0.00	6.96	0.00	0.00	1.58	0.00	7.32	0.00	0.00
	gLasso	0.27	0.00	0.81	0.00	0.00	0.09	0.00	0.26	0.00	0.00	0.07	0.00	0.20	0.00	0.00
	gBridge	2.87	0.01	10.04	0.08	0.00	3.07	0.00	10.49	0.00	0.00	3.30	0.00	11.20	0.00	0.01
	gMCP	0.09	0.00	0.43	0.00	0.00	0.02	0.00	0.11	0.00	0.00	0.00	0.00	0.04	0.00	0.01

Table A10. *Model complexity of the selected model in terms of correct deletion, erroneous deletion, and proportion of correct model for scenario VI*

		N =20					N = 50					N = 100				
		group		variable			group		variable			group		variable		
ρ	method	C	E	C	E	P	C	E	C	E	P	C	E	C	E	P
0	Lasso	2.35	0.00	38.84	0.00	0.00	2.24	0.00	37.62	0.00	0.00	2.43	0.00	40.32	0.00	0.00
	gLasso	0.06	0.00	0.60	0.00	0.00	0.01	0.00	0.05	0.00	0.00	0.01	0.00	0.13	0.00	0.00
	gBridge	3.84	0.00	48.80	0.00	0.00	3.94	0.00	49.67	0.00	0.00	3.94	0.00	48.64	0.00	0.00
	gMCP	1.55	0.00	24.58	0.00	0.00	0.02	0.00	0.68	0.00	0.00	0.00	0.00	0.12	0.00	0.00
0.1	Lasso	2.34	0.00	38.91	0.00	0.00	2.24	0.00	37.39	0.00	0.00	2.41	0.00	40.00	0.00	0.00
	gLasso	0.05	0.00	0.46	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.01	0.00	0.13	0.00	0.00
	gBridge	3.79	0.00	48.60	0.00	0.00	3.92	0.00	49.54	0.00	0.00	3.93	0.00	48.67	0.00	0.00
	gMCP	1.55	0.00	24.68	0.00	0.00	0.03	0.00	0.80	0.00	0.00	0.00	0.00	0.12	0.00	0.00
0.3	Lasso	2.04	0.00	34.46	0.01	0.00	1.87	0.00	32.73	0.00	0.00	2.03	0.00	34.73	0.00	0.00
	gLasso	0.05	0.00	0.45	0.00	0.00	0.01	0.00	0.05	0.00	0.00	0.01	0.00	0.05	0.00	0.00
	gBridge	3.75	0.00	48.65	0.00	0.00	3.89	0.00	49.79	0.00	0.00	3.89	0.00	49.02	0.00	0.00
	gMCP	1.58	0.00	25.48	0.00	0.00	0.04	0.00	0.85	0.00	0.00	0.00	0.00	0.15	0.00	0.00
0.5	Lasso	1.97	0.00	36.79	0.05	0.00	1.58	0.00	30.91	0.00	0.00	1.70	0.00	32.87	0.00	0.00
	gLasso	0.06	0.00	0.64	0.00	0.00	0.01	0.00	0.05	0.00	0.00	0.01	0.00	0.08	0.00	0.00
	gBridge	3.70	0.00	49.34	0.00	0.00	3.84	0.00	50.54	0.00	0.00	3.85	0.00	49.89	0.00	0.00
	gMCP	1.06	0.00	19.48	0.01	0.00	0.03	0.00	0.99	0.00	0.00	0.00	0.00	0.16	0.00	0.00
0.7	Lasso	1.58	0.00	33.41	0.22	0.00	1.27	0.00	28.35	0.00	0.00	1.34	0.00	30.37	0.00	0.00
	gLasso	0.04	0.00	0.36	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.00	0.00
	gBridge	3.42	0.00	47.46	0.06	0.00	3.59	0.00	49.21	0.00	0.00	3.63	0.00	48.83	0.00	0.00
	gMCP	1.24	0.00	23.27	0.11	0.00	0.04	0.00	1.19	0.00	0.00	0.00	0.00	0.23	0.00	0.00
0.9	Lasso	2.73	0.00	51.58	5.60	0.00	1.05	0.00	27.83	0.11	0.00	0.91	0.00	25.76	0.00	0.00
	gLasso	0.06	0.00	0.58	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	gBridge	3.20	0.00	48.76	1.64	0.00	2.76	0.00	44.17	0.07	0.00	2.78	0.00	44.03	0.00	0.00
	gMCP	1.49	0.00	29.34	1.14	0.00	0.07	0.00	1.76	0.00	0.00	0.01	0.00	0.31	0.00	0.00

Table A11. *Count of zero parameter over 800 simulated datasets for scenario I*

			Tuning Parameter (λ)											
N	**ρ**	**Par.**	**0.000**	**0.002**	**0.005**	**0.008**	**0.014**	**0.024**	**0.041**	**0.070**	**0.203**	**0.346**	**0.588**	**1.000**
20	0	β_1	0	15	34	54	111	164	263	410	747	798	800	800
		β_2	0	0	0	0	0	0	0	0	1	29	524	800
		β_3	0	10	25	51	76	139	247	383	755	799	800	800
	0.1	β_1	2	11	37	59	91	150	250	401	744	798	800	800
		β_2	0	0	0	0	0	0	0	0	1	40	526	800
		β_3	0	12	38	55	89	134	236	383	748	799	800	800
	0.3	β_1	0	14	30	46	68	131	252	383	727	787	799	800
		β_2	0	0	0	0	0	0	0	0	3	57	484	800
		β_3	2	16	38	56	90	158	249	392	720	790	800	800
	0.5	β_1	1	10	30	55	102	176	284	421	684	775	800	800
		β_2	0	0	0	0	0	0	0	0	8	73	497	798
		β_3	2	15	48	78	118	181	293	426	673	757	797	800
	0.7	β_1	0	10	36	65	118	212	339	455	586	687	770	799
		β_2	0	0	0	0	0	0	0	0	21	110	467	788
		β_3	2	10	29	57	97	191	294	415	593	694	780	800
	0.9	β_1	0	24	76	130	218	322	414	454	512	584	697	787
		β_2	0	1	10	20	24	29	36	42	83	211	504	758
		β_3	1	28	87	138	214	309	389	427	490	573	693	793
50	0	β_1	1	19	48	78	128	208	367	553	797	800	800	800
		β_2	0	0	0	0	0	0	0	0	0	2	572	800
		β_3	1	13	41	77	131	227	369	574	798	800	800	800
	0.1	β_1	1	23	52	77	134	214	356	560	798	800	800	800
		β_2	0	0	0	0	0	0	0	0	0	3	558	800
		β_3	0	16	44	82	143	224	376	557	796	800	800	800
	0.3	β_1	1	15	47	83	131	217	342	535	776	800	800	800
		β_2	0	0	0	0	0	0	0	0	0	10	543	800
		β_3	1	12	47	82	150	229	360	535	787	800	800	800
	0.5	β_1	0	17	52	88	151	242	394	543	750	797	800	800
		β_2	0	0	0	0	0	0	0	0	0	17	527	800
		β_3	0	18	50	87	151	247	382	543	763	795	800	800
	0.7	β_1	0	23	65	124	199	292	415	509	681	777	799	800
		β_2	0	0	0	0	0	0	0	0	1	30	502	800
		β_3	1	22	69	118	185	283	408	499	681	768	799	800
	0.9	β_1	0	45	111	180	269	362	416	437	516	586	738	799
		β_2	0	0	0	0	0	1	1	1	7	88	497	797
		β_3	0	47	114	185	286	384	431	444	522	600	741	800
100	0	β_1	1	22	71	107	179	290	458	675	800	800	800	800
		β_2	0	0	0	0	0	0	0	0	0	0	611	800
		β_3	1	28	66	117	210	333	487	677	800	800	800	800
	0.1	β_1	2	22	63	105	192	310	496	697	800	800	800	800
		β_2	0	0	0	0	0	0	0	0	0	0	604	800
		β_3	2	28	66	112	190	323	502	689	800	800	800	800
	0.3	β_1	3	29	69	111	191	305	508	668	800	800	800	800
		β_2	0	0	0	0	0	0	0	0	0	1	604	800
		β_3	2	23	71	118	194	316	495	653	798	800	800	800
	0.5	β_1	3	35	84	131	219	341	503	628	789	799	800	800
		β_2	0	0	0	0	0	0	0	0	0	1	582	800
		β_3	2	36	92	141	239	360	499	627	790	800	800	800
	0.7	β_1	1	31	77	137	222	361	470	555	749	793	800	800
		β_2	0	0	0	0	0	0	0	0	0	5	535	800
		β_3	0	41	93	158	252	368	481	564	741	790	800	800
	0.9	β_1	0	53	146	248	345	403	418	440	550	638	770	800
		β_2	0	0	0	0	0	0	0	0	0	25	496	800
		β_3	0	67	153	240	349	420	436	455	540	612	772	800

Table A12. *Bias (standard error) for scenario III from the generating data with 50 subjects*

Par.	Method	ρ = 0	ρ = 0.1	ρ = 0.5	ρ = 0.9
β_{11}	Lasso	-0.0287(0.0015)	-0.0262(0.0017)	-0.0314(0.0019)	-0.0472(0.0033)
	gLasso	-0.0243(0.0015)	-0.0219(0.0016)	-0.0203(0.0018)	-0.0605(0.0029)
	gBridge	-0.0215(0.0015)	-0.0192(0.0016)	-0.0263(0.0018)	-0.0601(0.0030)
	gMCP	-0.0081(0.0015)	-0.0074(0.0016)	-0.0084(0.0018)	-0.0407(0.0031)
	GEE	-0.0024(0.0015)	0.0002(0.0016)	0.0016(0.0018)	-0.0029(0.0036)
β_{12}	Lasso	0.0002(0.0009)	0.0025(0.0009)	0.0121(0.0008)	0.0423(0.0022)
	gLasso	0.0008(0.0015)	0.0052(0.0015)	0.0254(0.0015)	0.1394(0.0026)
	gBridge	0.0006(0.0012)	0.0031(0.0012)	0.0207(0.0011)	0.1112(0.0025)
	gMCP	0.0008(0.0015)	0.0017(0.0015)	0.0112(0.0016)	0.0921(0.0029)
	GEE	0.0008(0.0015)	-0.00003(0.0015)	-0.0010(0.0016)	0.0039(0.0037)
β_{13}	Lasso	-0.0274(0.0016)	-0.0261(0.0016)	-0.0337(0.0018)	-0.0456(0.0033)
	gLasso	-0.0341(0.0016)	-0.0330(0.0015)	-0.0335(0.0018)	-0.1009(0.0030)
	gBridge	-0.0242(0.0016)	-0.0229(0.0015)	-0.0347(0.0018)	-0.0903(0.0030)
	gMCP	-0.0098(0.0016)	-0.0111(0.0015)	-0.0155(0.0018)	-0.0642(0.0031)
	GEE	-0.0011(0.0016)	0.0003(0.0015)	-0.0008(0.0018)	-0.0010(0.0036)
β_{21}	Lasso	-0.0007(0.0010)	-0.0004(0.0009)	0.0006(0.0010)	0.0031(0.0014)
	gLasso	-0.0008(0.0011)	-0.0001(0.0011)	0.0004(0.0013)	0.0052(0.0018)
	gBridge	-0.0007(0.0005)	-0.0001(0.0005)	0.0009(0.0005)	0.0009(0.0010)
	gMCP	-0.0011(0.0016)	-0.0007(0.0015)	0.0003(0.0018)	0.0082(0.0032)
	GEE	-0.0013(0.0016)	-0.0008(0.0016)	0.0003(0.0018)	0.0084(0.0036)
β_{22}	Lasso	-0.0005(0.0010)	0.0013(0.0009)	0.0011(0.0009)	0.0004(0.0009)
	gLasso	-0.0005(0.0012)	0.0014(0.0011)	0.0015(0.0012)	-0.0002(0.0012)
	gBridge	-0.0004(0.0005)	0.0002(0.0004)	0.0011(0.0006)	0.0004(0.0005)
	gMCP	-0.0007(0.0016)	0.0025(0.0015)	0.0022(0.0017)	-0.0034(0.0028)
	GEE	-0.0008(0.0017)	0.0024(0.0015)	0.0024(0.0017)	-0.0047(0.0036)
β_{23}	Lasso	0.0002(0.0010)	0.0008(0.0009)	-0.0015(0.0010)	0.0014(0.0013)
	gLasso	0.0001(0.0011)	0.0008(0.0011)	-0.0009(0.0013)	0.0015(0.0018)
	gBridge	-0.0004(0.0004)	-0.0003(0.0005)	-0.0003(0.0005)	0.0008(0.0009)
	gMCP	0.0005(0.0016)	0.0009(0.0015)	-0.0010(0.0018)	0.0021(0.0032)
	GEE	0.0003(0.0016)	0.0008(0.0016)	-0.0011(0.0018)	0.0028(0.0036)

Appendix B

Figure B1. *ME boxplot for scenario I: (1) Lasso, (2) GEE*

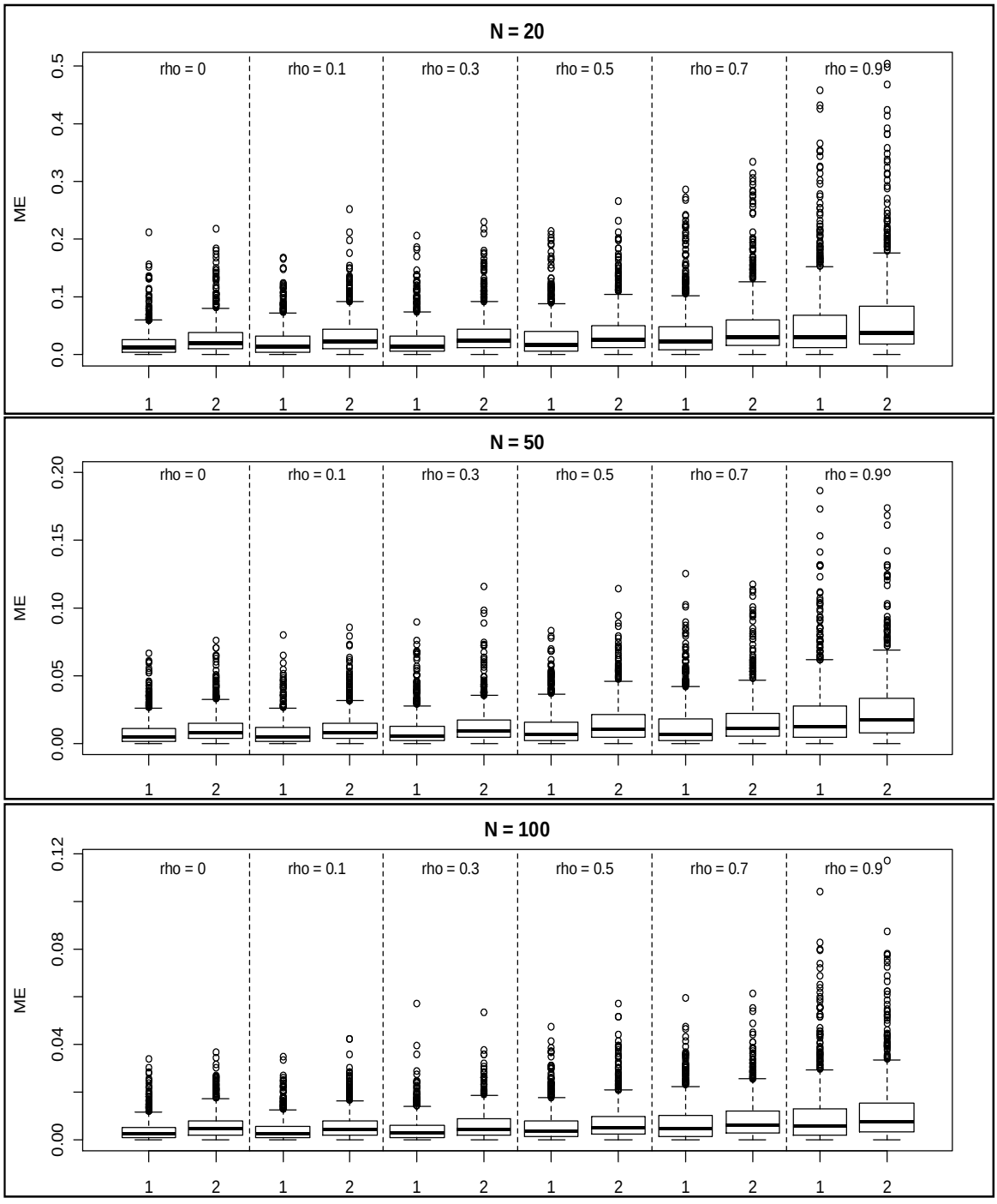


Figure B2. *ME boxplot for scenario II: (1) Lasso, (2) GEE*

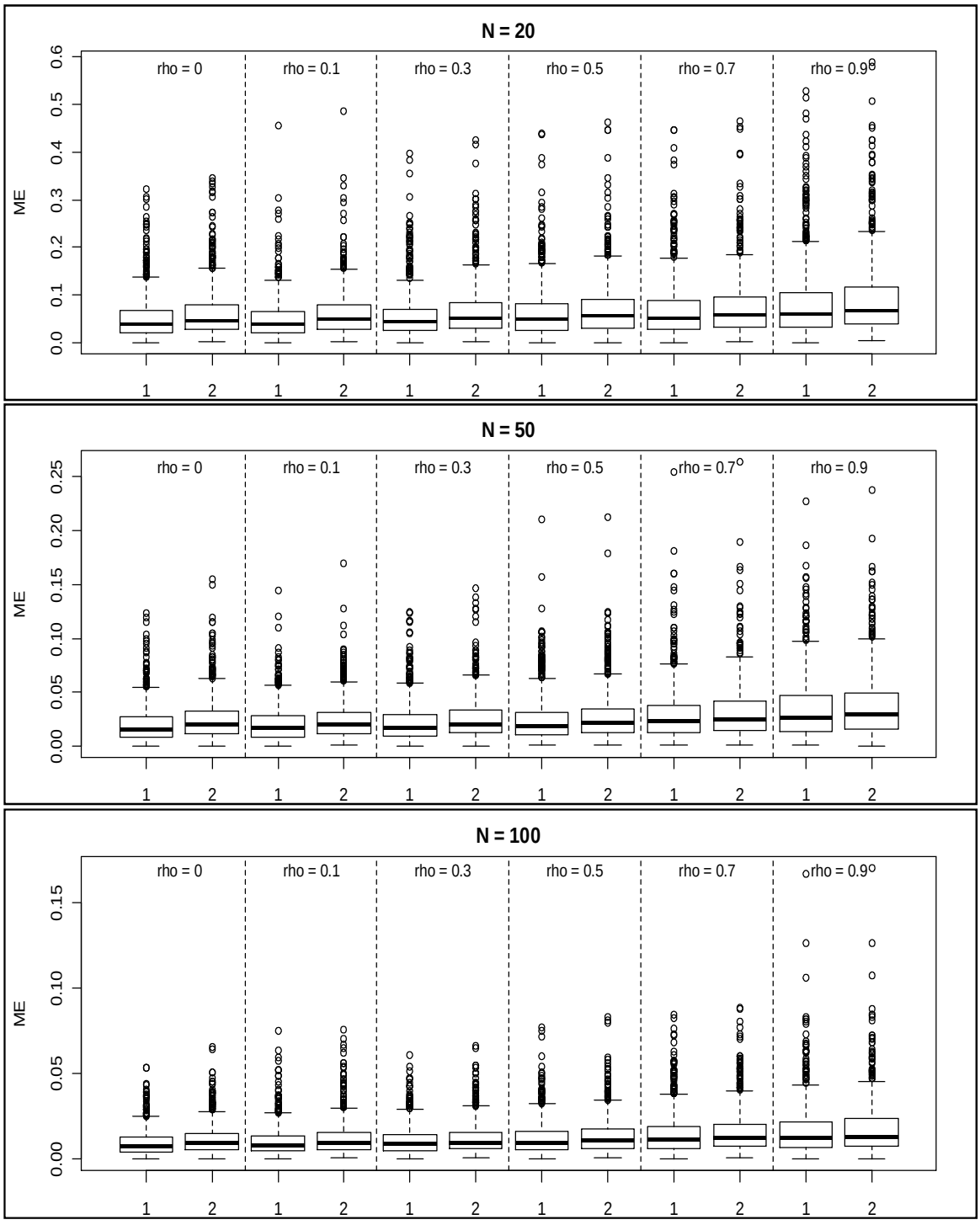


Figure B3. *ME boxplot for scenario III: (1)Lasso, (2) gLasso, (3) gBridge, (4) gMCP, (5) GEE*

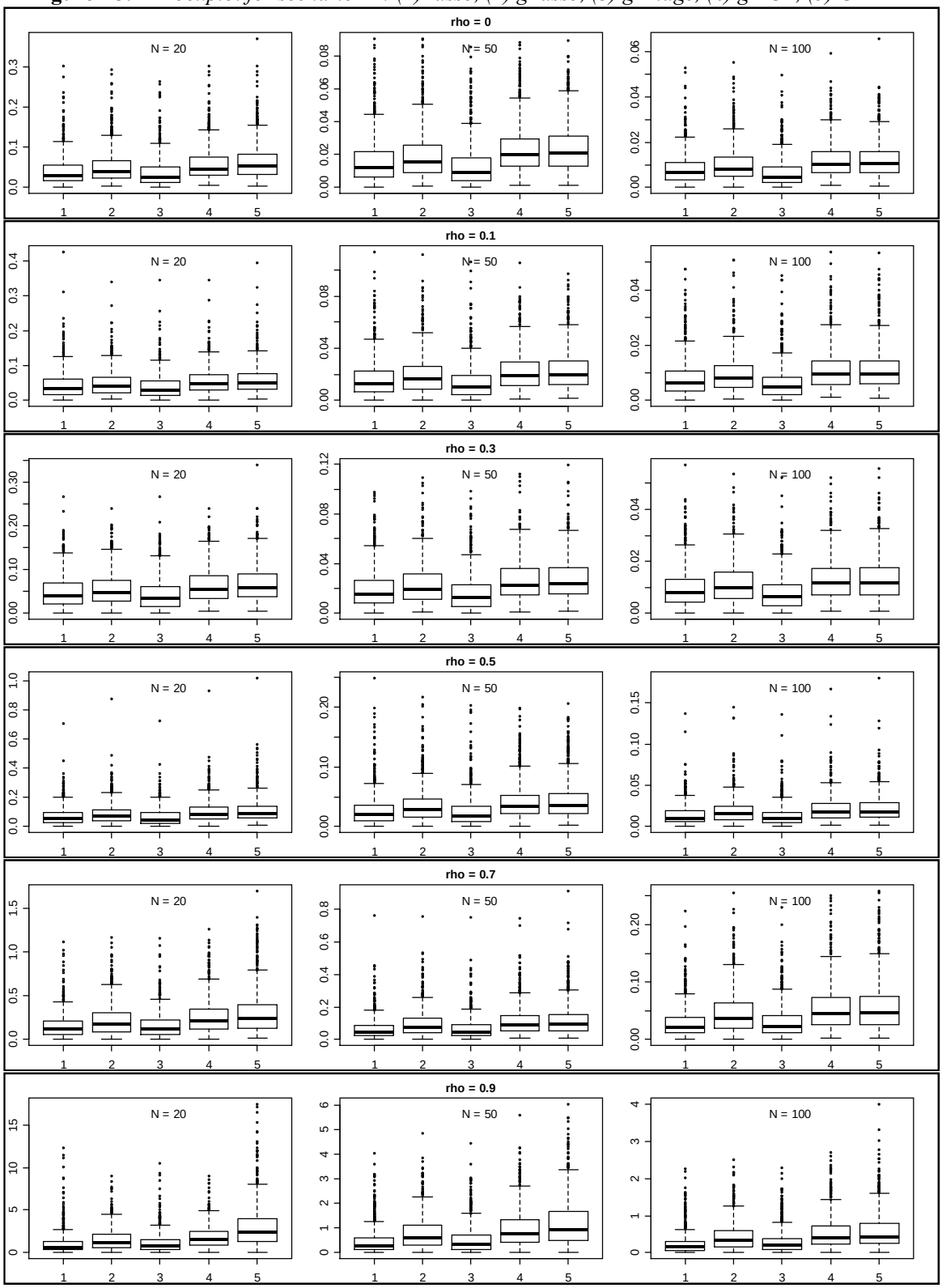


Figure B4. *ME boxplot for scenario IV: (1)Lasso, (2) gLasso, (3) gBridge, (4) gMCP, (5) GEE*

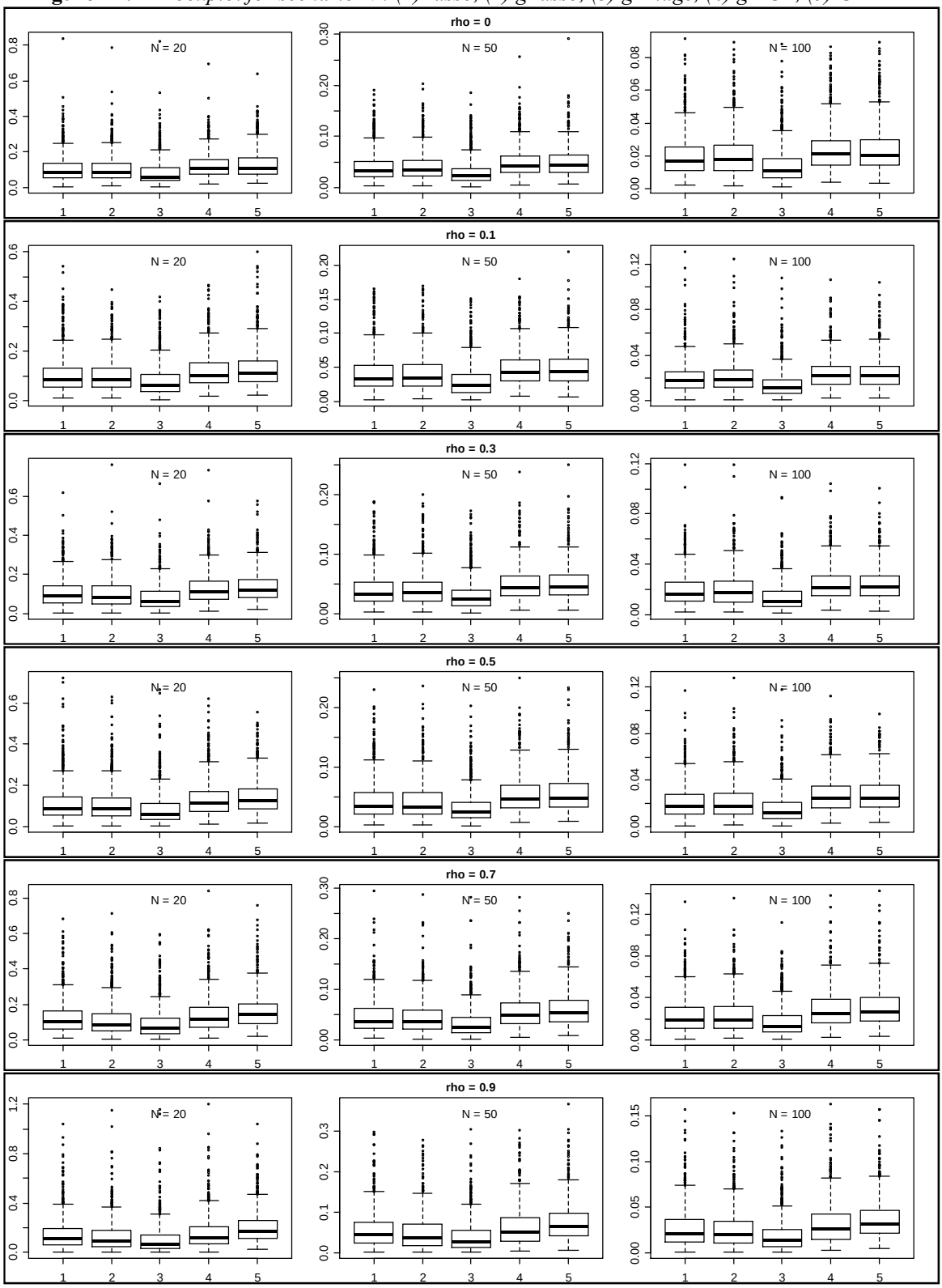


Figure B5. *ME boxplot for scenario V: (1)Lasso, (2) gLasso, (3) gBridge, (4) gMCP, (5) GEE*

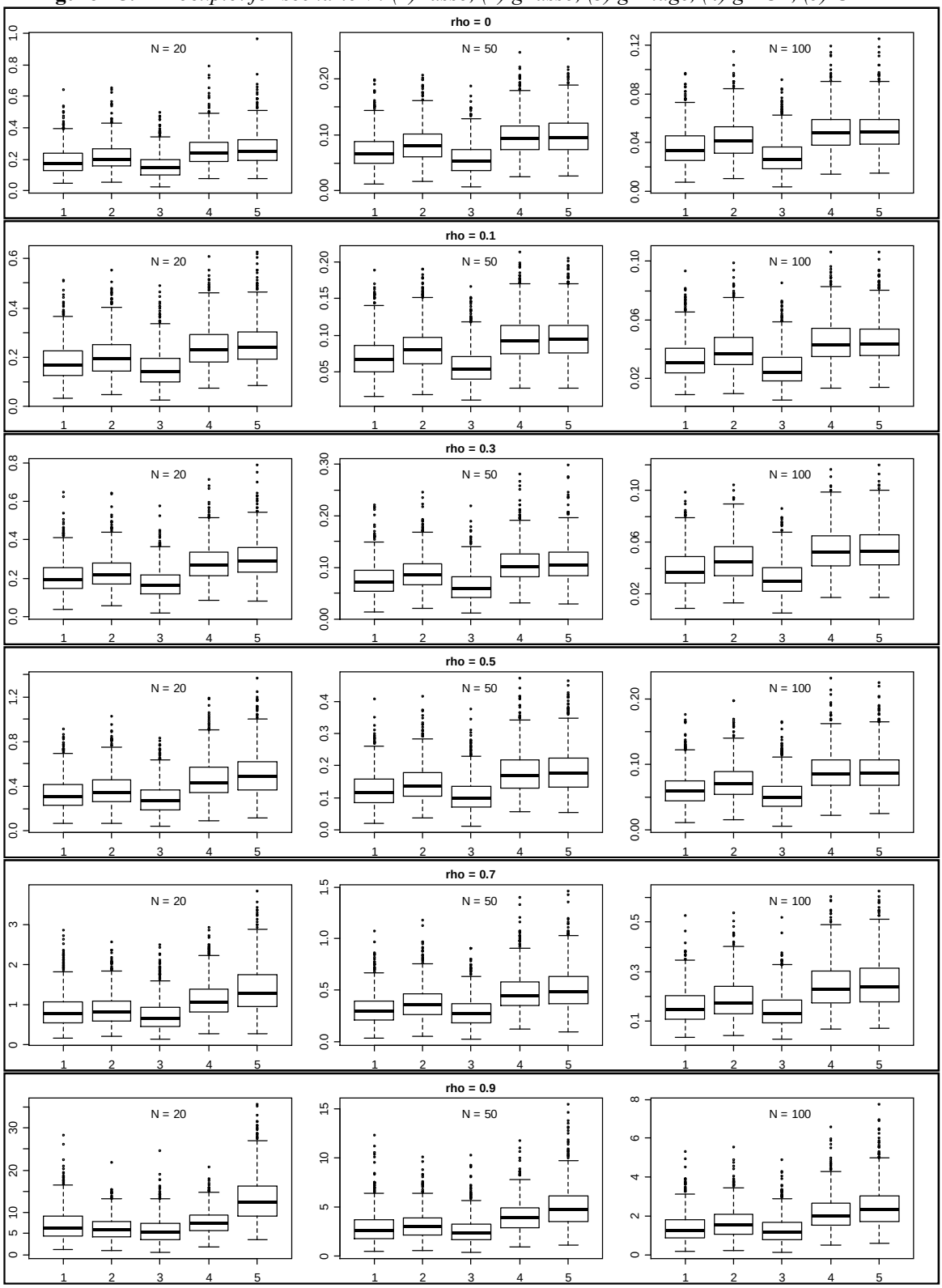


Figure B6. *ME boxplot for scenario VI: (1)Lasso, (2) gLasso, (3) gBridge, (4) gMCP, (5) GEE*

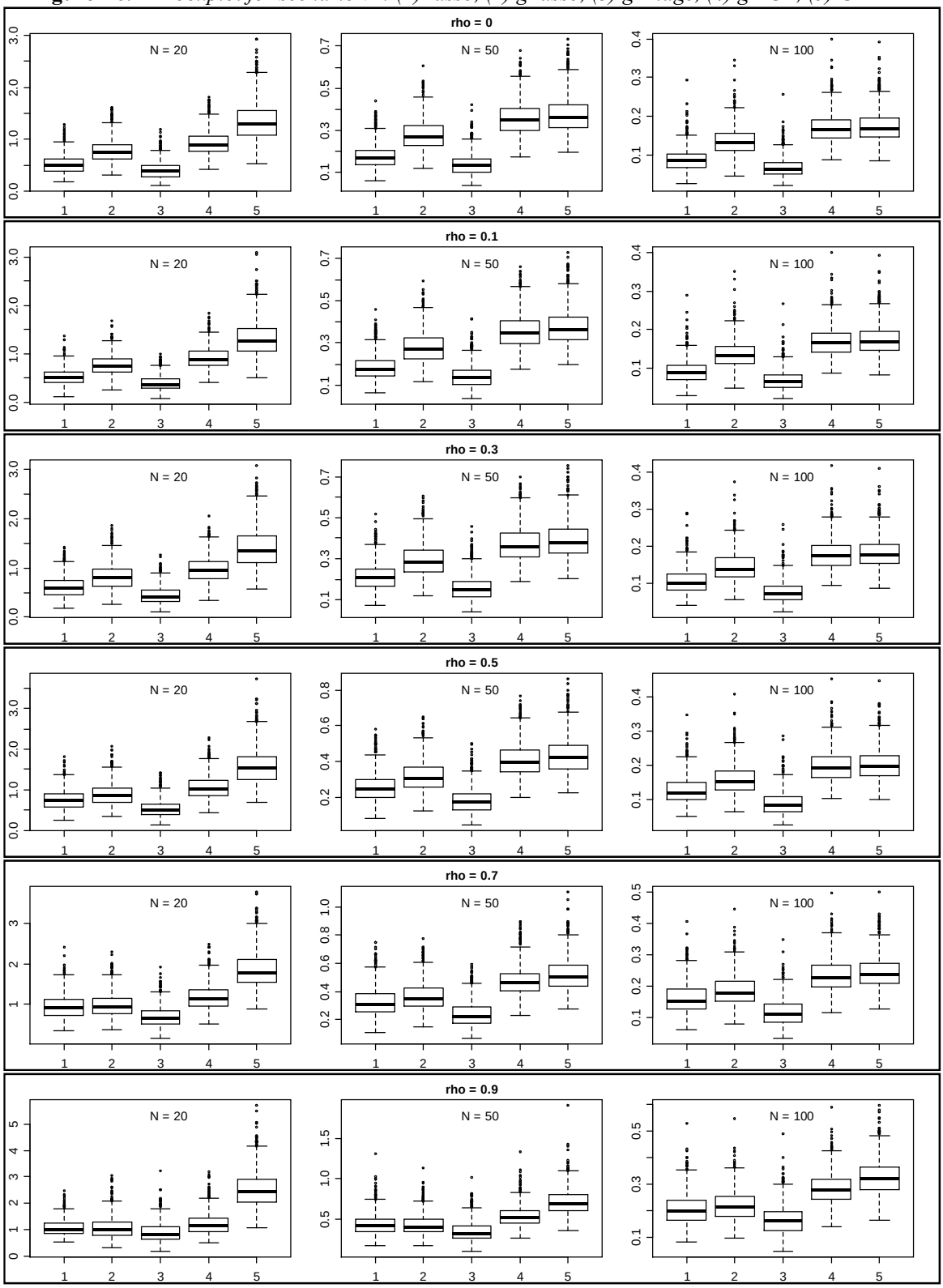


Figure B7. *Regularization path for scenario I*

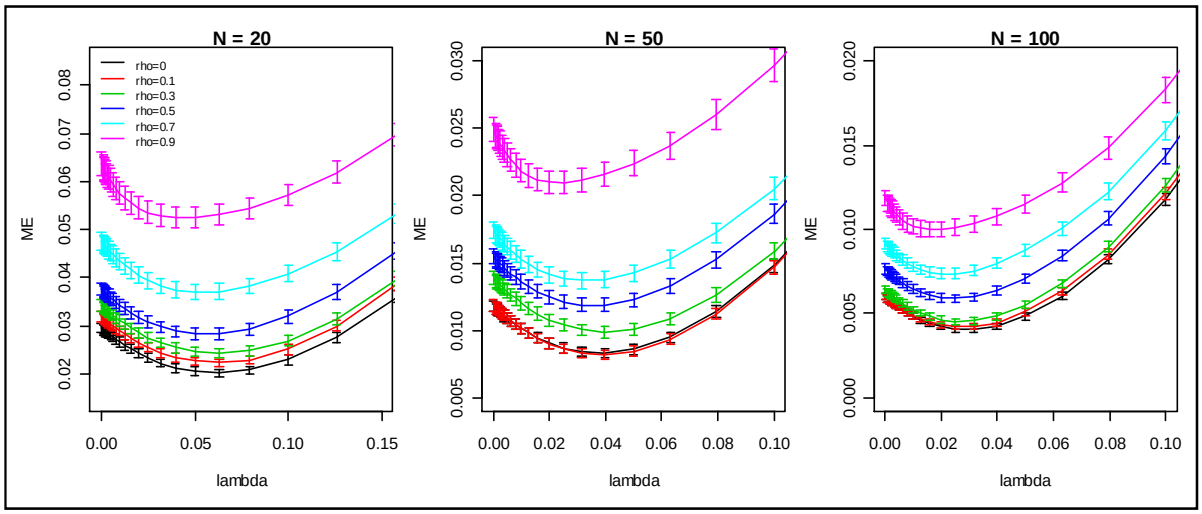


Figure B8. *Regularization path for scenario II*

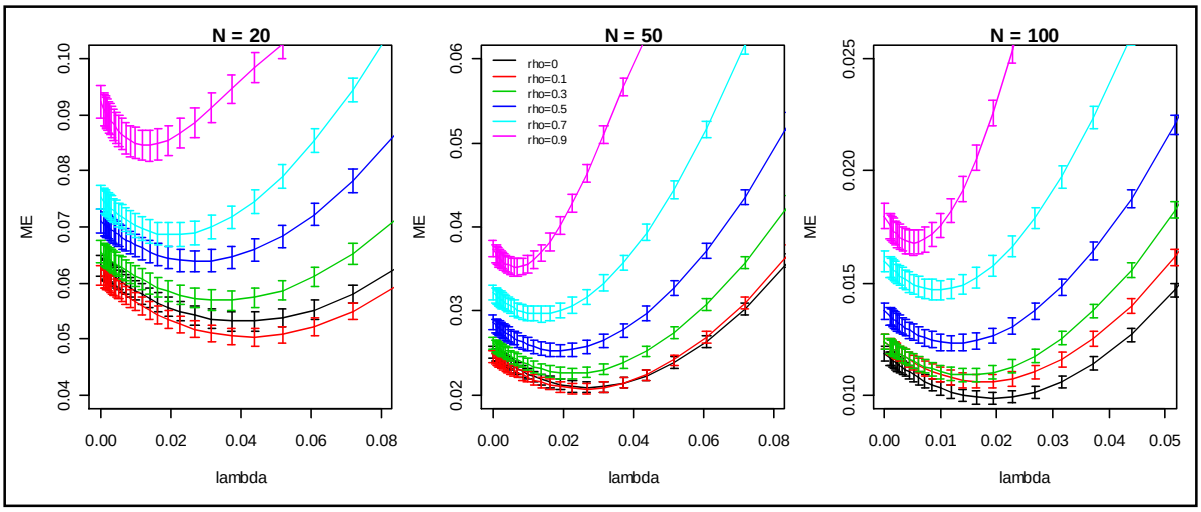


Figure B9. *Model error path for scenario III: gLasso (red line), gBridge (green line), gMCP (blue line), the Lasso (black line)*

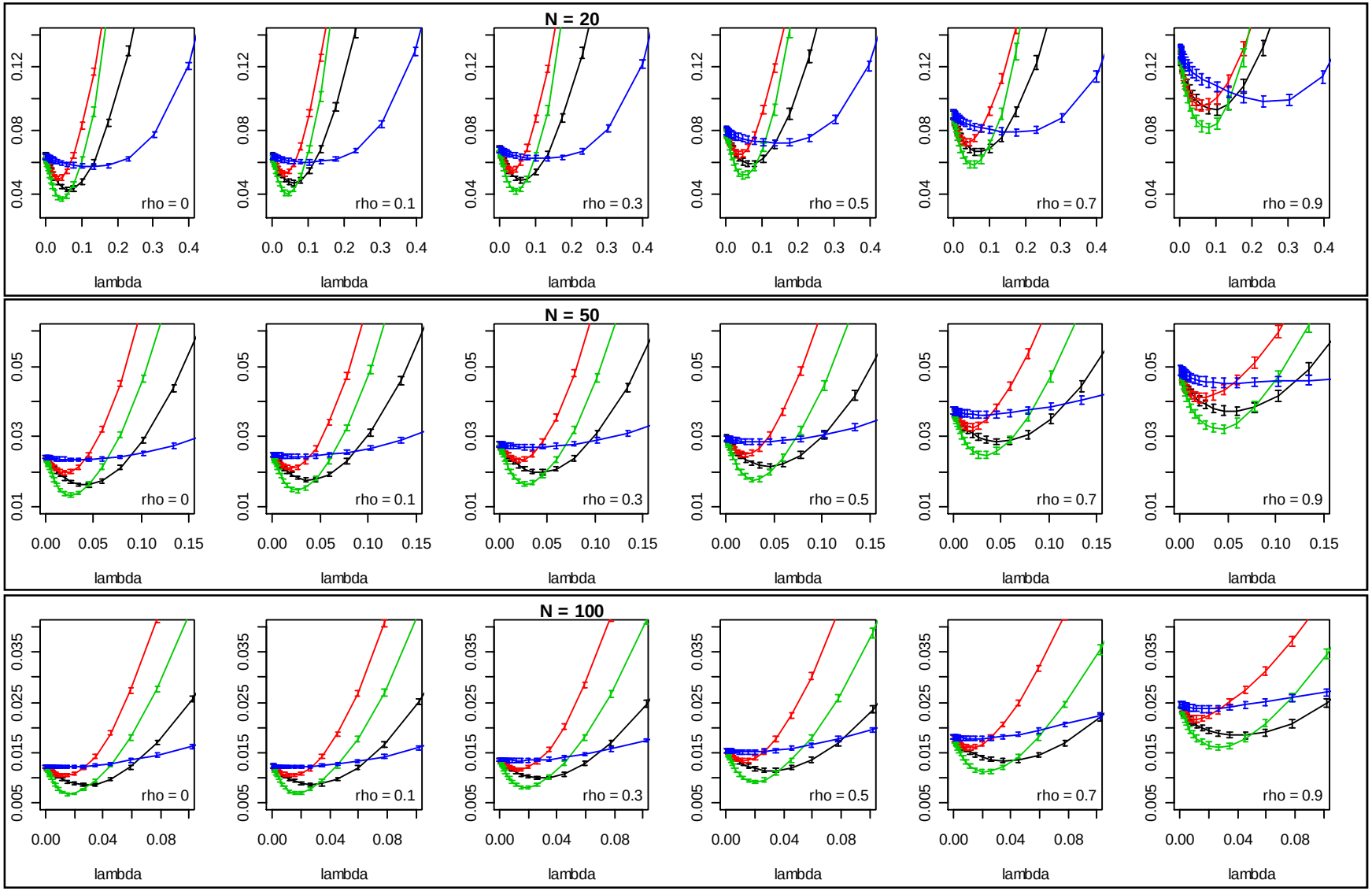


Figure B10. *Model error path for scenario IV: gLasso (red line), gBridge (green line), gMCP (blue line), the Lasso (black line)*

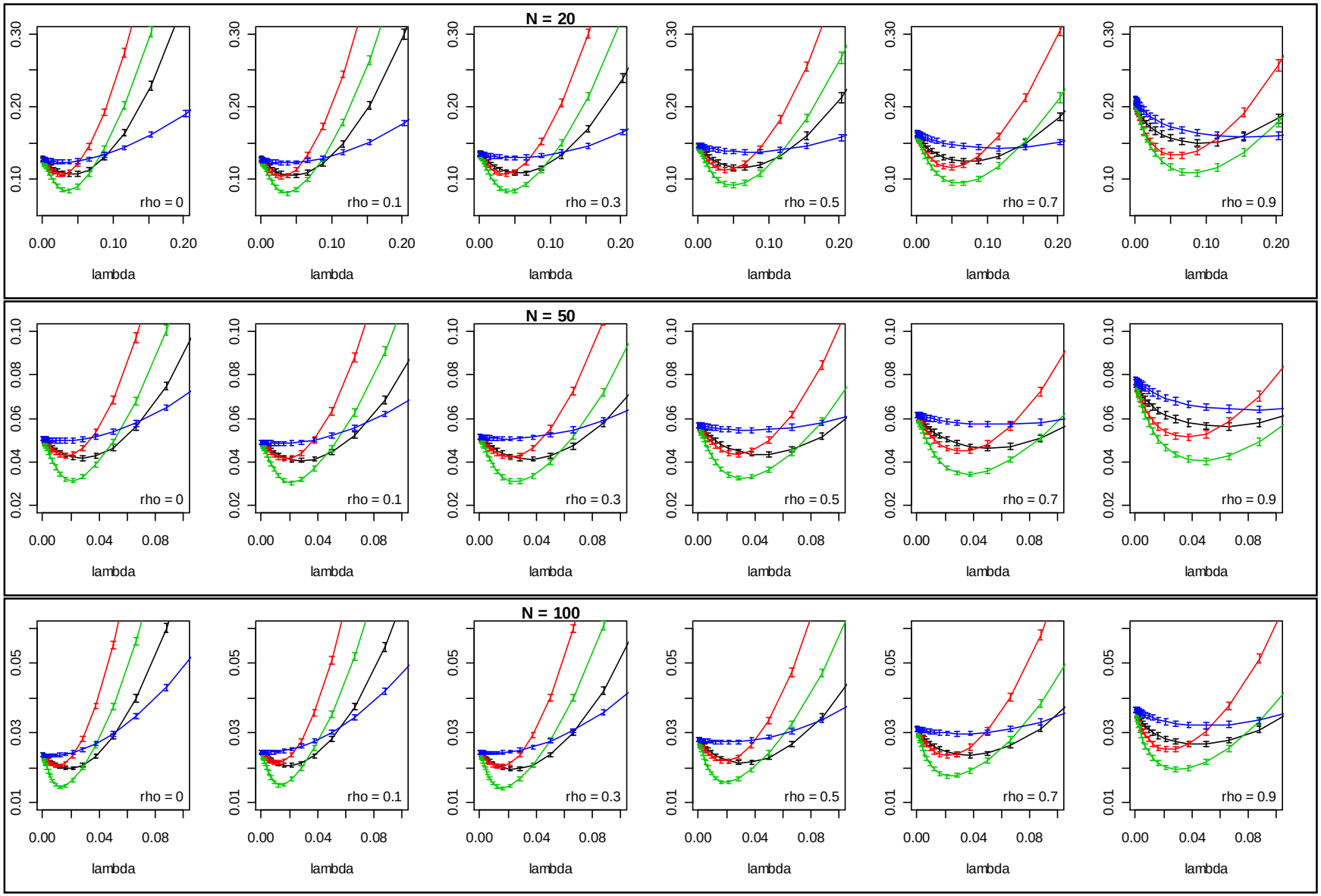


Figure B11. *Model error path for scenario V: gLasso (red line), gBridge (green line), gMCP (blue line), the Lasso (black line)*

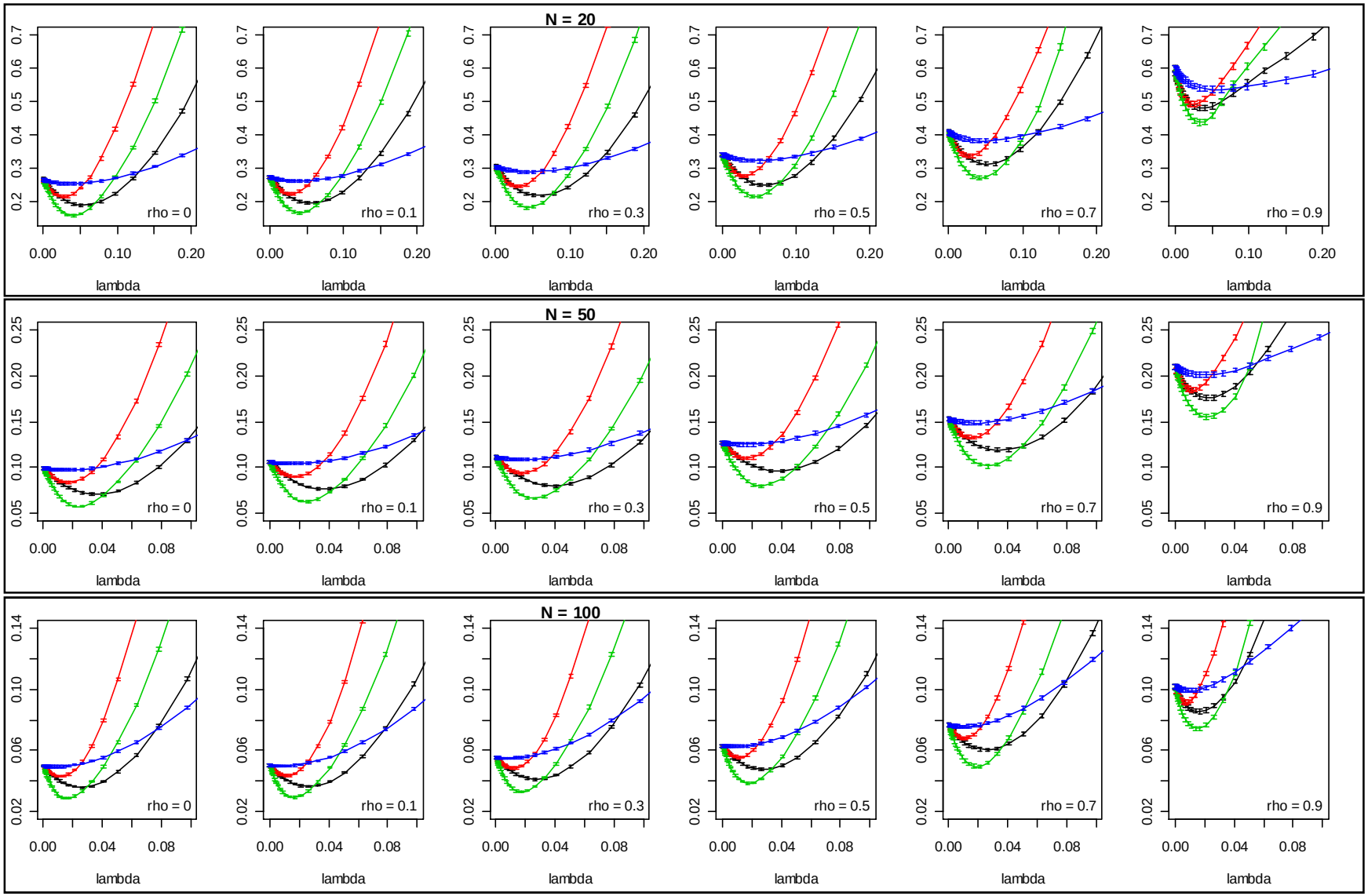


Figure B12. *Model error path for scenario VI: gLasso (red line), gBridge (green line), gMCP (blue line), the Lasso (black line)*

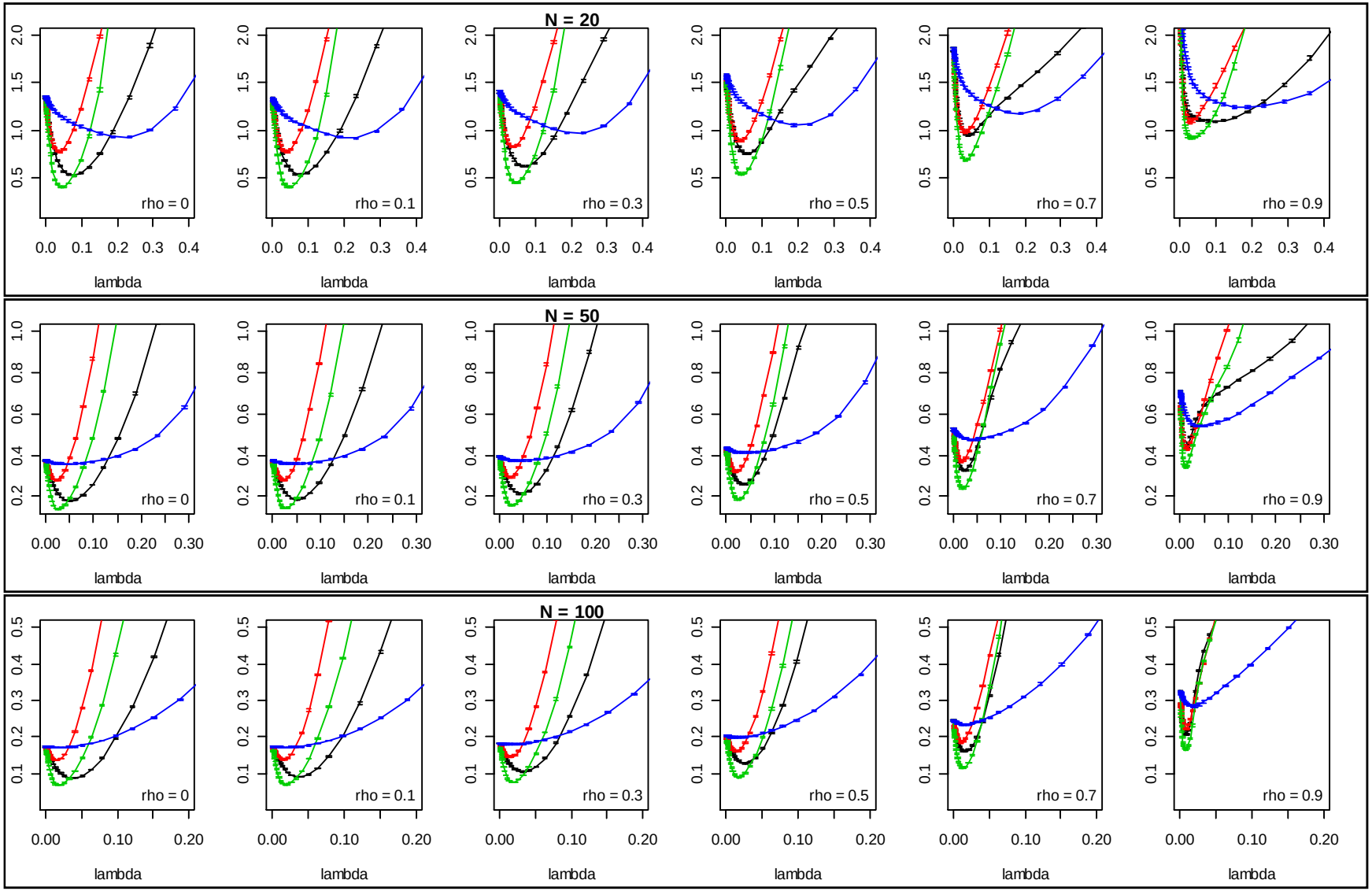


Figure B13. *Path of count of zero coefficients for scenario III*

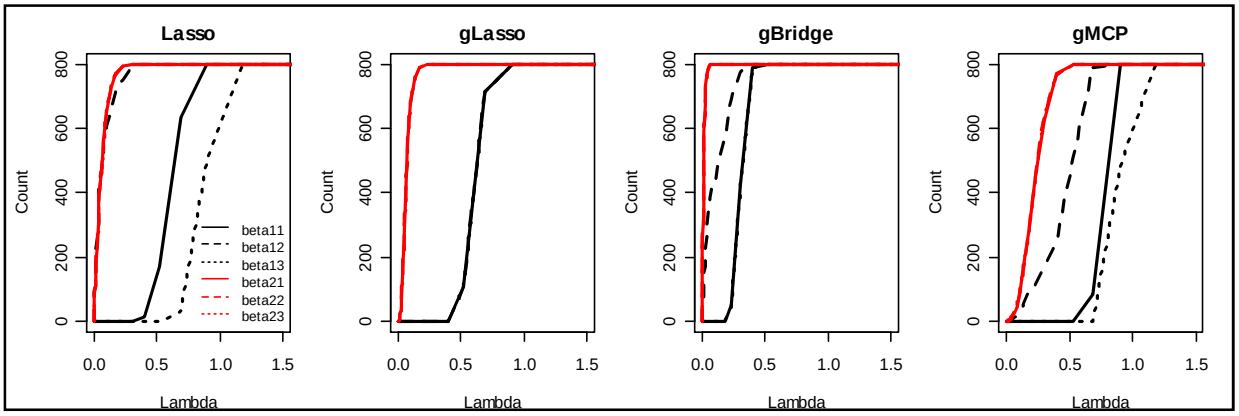


Figure B14. *Bias-variance tradeoff from Lasso:bias (red line); standard deviation (blue line)*

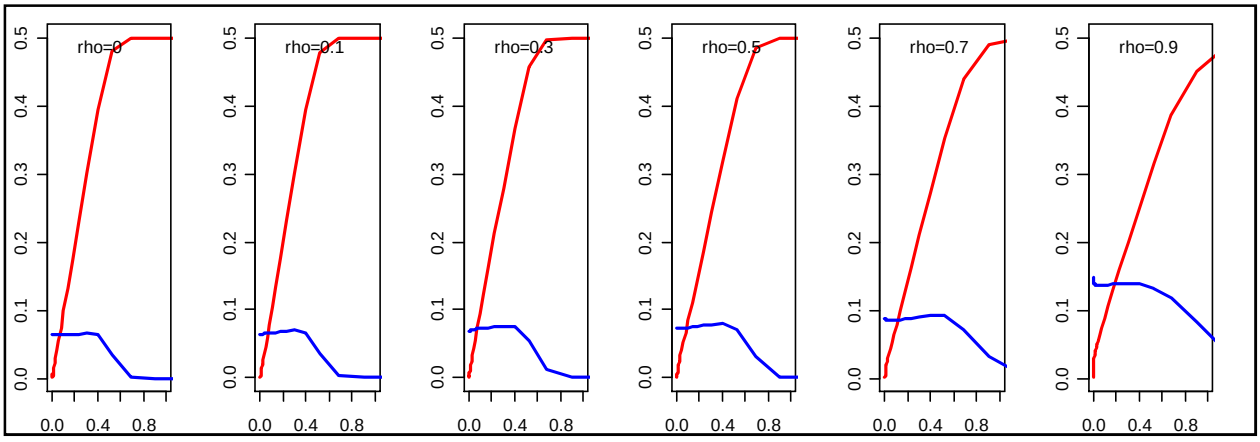


Figure B15. *Bias-variance tradeoff from group Lasso:bias (red line); standard deviation (blue line)*

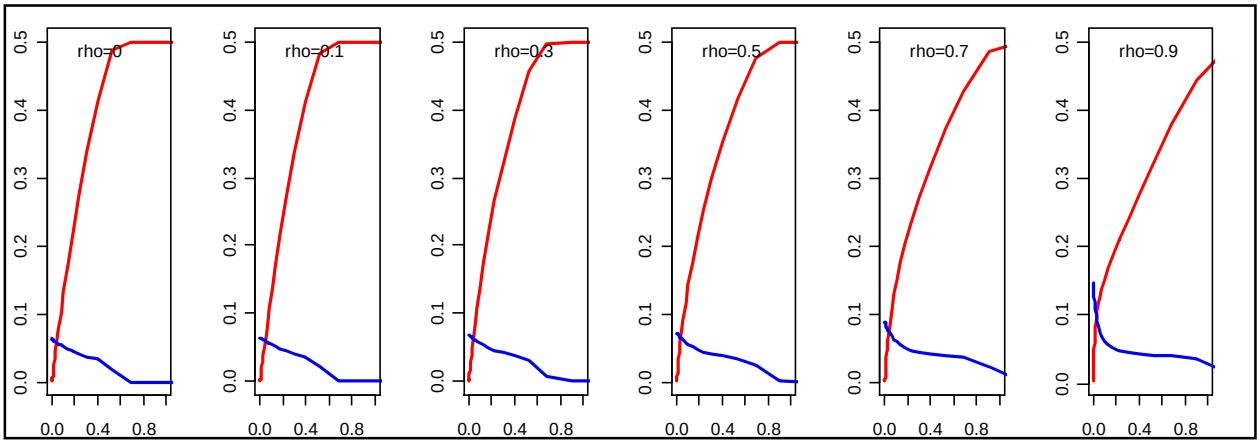


Figure B16. *Bias-variance tradeoff from group bridge:bias (red line); standard deviation (blue line)*

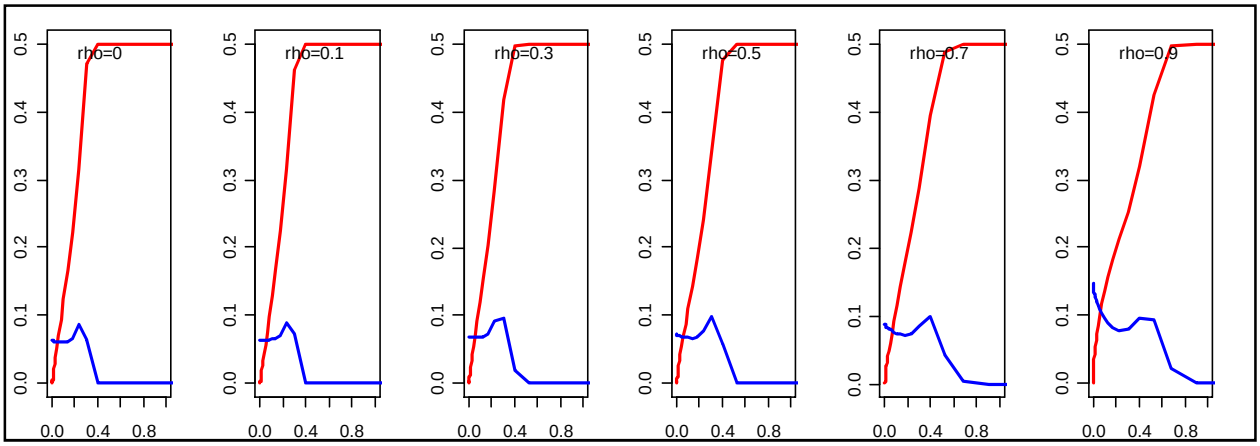
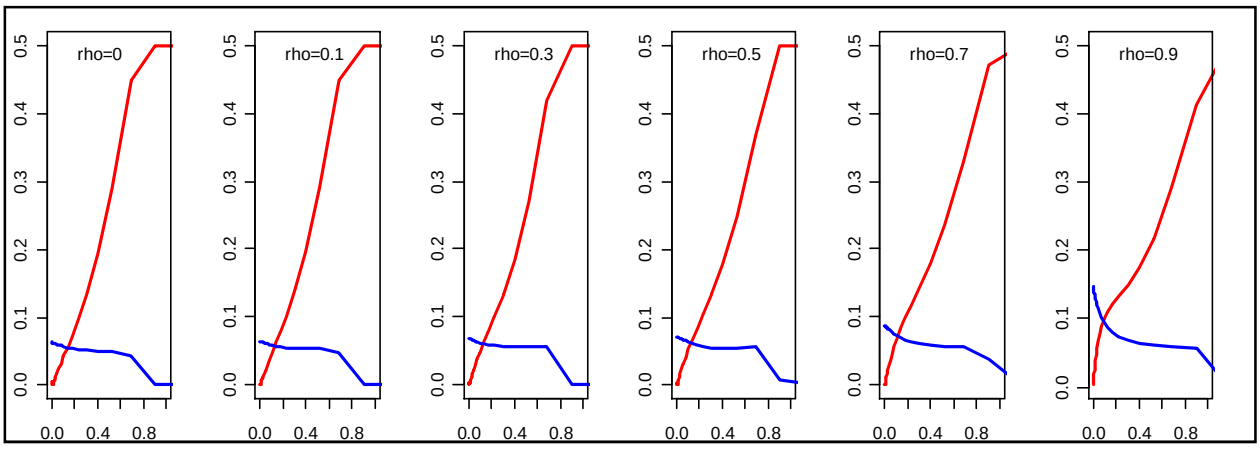


Figure B17. *Bias-variance tradeoff from group MCP:bias (red line); standard deviation (blue line)*



Appendix C

This Appendix presents the R 2.12 codes used to carry out the simulations and analyses in Chapter 3. The codes for generating the simulated data under the different scenarios are essentially the same so we only include the codes for scenario I.

Code to data simulation

```
rm(list=ls(all=TRUE));#clear workspace
library(geepack)
library(matlab)

#Number of subject
 N <- 20 #20/50/100#
#Number of observations for each
subject
 T <- 10
#Number of data simulated
 S <- 800
#model parameters
lambda0 <- 0 ; lambda1 <- 0 ; lambda2
<- 0.5; lambda3 <- 0;
#Rho
seq_rho <- c(0,seq(0.1,0.9,0.2))
R <- length(seq_rho)

SIMList <- as.list(1:S)
SIMLISTDIGGLE <- as.list(1:R)

set.seed(240685)
####################
###Data generation
####################
for (r in 1:R)
{
rho <- seq_rho[r]

for (s in 1:S)
{

#Error
b <- rnorm(n=N,0,1)
e <- rnorm(n=N*T,0,1)
e <- matrix(e,nrow=N,ncol=T)
e <- cbind(rep(NA,N),rep(NA,N),e)
eta <- matrix(nrow=N, ncol=T+2)
eta[,1] <- rnorm(n=N)
eta[,2:(T+2)] <- rnorm(n=N*(T+1), mean
= 0, sd =(sqrt(1-rho^2)))

#X and Y variables

X <- matrix(ncol=T+2,nrow=N)
Y <- matrix(ncol=T+2,nrow=N)

X[,1] <- eta[,1]

for (i in 1:N)
{
       for (t in 2:(T+2))
       {
       X[i,t] = rho*X[i,t-1] + eta[i,t]
       }
       for (t in 3:(T+2))
       {
       Y[i,t] = lambda0 + lambda1*X[i,t]
+ lambda2*X[i,t-1] + lambda3*X[i,t-2] +
b[i] + e[i,t]
       }
}

#dataframe for analysis
data <- NULL
#id <- sort(rep(1:N,T))
xt_2 <- X[,1:T]
xt_1 <- X[,1:T+1]
xt <- X[,1:T+2]
y <- Y[,3:(T+2)]
for ( j in 1:N)
{
data_per_id <- cbind
(rep(j,T),xt[j,],xt_1[j,],xt_2[j,],y[j,
])
data <- rbind(data,data_per_id)
}
data<-data.frame(data)
names(data) <-
c('id','xt_2','xt_1','xt','y')
SIMList[[s]] <- data

}
names(SIMList) <-
paste('SIM',1:S,sep="")
SIMLISTDIGGLE[[r]] <- SIMList

}

names(SIMLISTDIGGLE) <-
paste('rho',c(0,seq(0.1,0.9,0.2)),sep='
')
```

Code to approximate the lagged coefficients with B-splines basis and perform the group Lasso

```
rm(list=ls(all=TRUE));#clear workspace
library(matlab)
library(geepack)
library(grpreg)

lambda <- c(0,logspace(-3,0.6,39))
```

```
nlambda <- length(lambda)
df <- 4 #4/5/8#
ngroup <- 5
nbeta <- ngroup * df
index <- rep(c(1:ngroup), each = df)
id <- rep(c(1:N),each = T)
PE <- array(dim = c(N, nlambda, S))
BETA <- array(dim = c(nlambda,nbeta,S))
SIGMA <- matrix (1, nrow = T, ncol = T)
+ diag(1, T, T)
SIGMA.inv <- solve(SIGMA)
GAMMA <- array(dim = c(N*T, df*ngroup,
S))
lag <- c(0:11)
B <- bs(lag, df = df, intercept = TRUE)

for (s in 1:S) {
       SIM_dat <- SIMList1[[s]]
       X1 <- as.matrix(SIM_dat[,2:13])
       X2 <- as.matrix(SIM_dat[,14:25])
       X3 <- as.matrix(SIM_dat[,26:37])
       X4 <- as.matrix(SIM_dat[,38:49])
       X5 <- as.matrix(SIM_dat[,50:61])
       Xstar1 <- X1%*%B
       Xstar2 <- X2%*%B
       Xstar3 <- X3%*%B
       Xstar4 <- X4%*%B
       Xstar5 <- X5%*%B
       x <-
as.matrix(cbind(Xstar1,Xstar2,Xstar3,Xs
tar4,Xstar5))
       y <- as.vector(SIM_dat$y)
       fit.glasso <- grpreg(x, y, group
= index, family ="gaussian", penalty
="gLasso", nlambda = nlambda, lambda =
lambda)
       beta.glasso <-
t(fit.glasso$beta[-1,])
       BETA[,,s] <- beta.glasso
       for (l in 1:nlambda) {
              BETA_l <- BETA[l,,s]
              mu <- BETA_l%*%t(x)
              diff <- (y - mu)
              diff.id <-
cbind(id,c(diff))
              for (i in 1:N) {
                     diff.i <-
subset(diff.id, diff.id[,1] == i)
                     PE[i,l,s] <-
(diff.i[,2])%*%SIGMA.inv%*%(diff.i[,2])
              }
       }
}


PE.mean <- PE.se <- c(1:nlambda)
for (l in 1:nlambda) {
       PE.mean[l] <-
mean(apply(PE[,l,],2,mean))
       PE.se[l] <-
sd(apply(PE[,l,],2,mean)) / sqrt(S)
}
```

```
plot(lambda,PE.mean,type='l',col=1,lwd=
2,ylab="PE",xlab="Lambda",ylim=range(17
.17,17.2),xlim=range(0,0.01))


#calcuate test error
rm(list=ls(all=TRUE));#clear workspace
load("F:\\Biostats\\4th_semester\\Maste
r_Thesis\\Sim7\\databsplines7v4")

coeff <- apply(BETA[1,,],1,mean)
b1 <- B%*%coeff[1:df]
b2 <- B%*%coeff[(df*1+1):(df*2)]
b3 <- B%*%coeff[(df*2+1):(df*3)]
b4 <- B%*%coeff[(df*3+1):(df*4)]
b5 <- B%*%coeff[(df*4+1):(df*5)]

load("F:\\Biostats\\4th_semester\\Maste
r_Thesis\\Sim7\\datasim7testset")
beta.hat <- c(b1,b2,b3,b4,b5)
SIGMA <- matrix (1, nrow = T, ncol = T)
+ diag(1, T, T)
SIGMA.inv <- solve(SIGMA)
PE <- matrix(0, ncol = N, nrow = S)
for (s in 1:S) {
       SIM_test_dat <- SIMList1[[s]]
       x <- SIM_test_dat[,2:61]
       y <- SIM_test_dat$y
       id <- SIM_test_dat$id
       mu <- beta.hat%*%t(x)
       diff <- (y - mu)
       diff.id <- cbind(id,c(diff))
       for (i in 1:N) {
              diff.i <- subset(diff.id,
diff.id[,1] == i)
              PE[s,i] <-
(diff.i[,2])%*%SIGMA.inv%*%(diff.i[,2])
       }
}
PE.mean <- mean(apply(PE,1,mean))
PE.se <- sd(apply(PE,1,mean))/sqrt(800)
PE.mean;PE.se;
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Bi-level selection in Longitudinal Analysis with Time Dependent Covariates

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Wijaya, Madona

Datum: **12/09/2011**