- LUDWIG MAXIMILIANS-UNIVERSITÄT MÜNCHEN -
- Institut für Statistik -

Masterarbeit

Measurement Error in LASSO - Analytical Results and a Simulation Study

Wissenschaftliche Arbeit zur Erlangung des akademischen Grades Master of Science
eingereicht am Institut für Statistik

Betreuer: Prof. Dr. Christian Heumann

Autorin: Kathrin Jahreis
M.Sc. Biostatistik (PO 2010)

München, den 05.11.2015

Abstract

The LASSO (Tibshirani, 1996) is a powerful method which due to the fact that is uses a l_1-penalty allows for the estimation of regression coefficients and variable selection at the same time. An important property of the LASSO is that it can be applied even when the number of covariates p is larger than the number of observations n. This differs the LASSO from the popular OLS method which requires $p < n$. An assumption that is common to the vast majority of studies on the LASSO is that the design matrix $\mathbf{X}$ passed to the LASSO for performing linear regression contains perfect covariate measurements that do not suffer from additive measurement error. However, in practice where data corrupted by measurement errors or errors-in-variable data are rather the norm than the exception, this assumption does not meet the truth. In this work, we studied the LASSO in the presence of additive measurement error in the design matrix. In doing so, we allowed for analytical results on the estimation and variable selection consistency of both the LASSO with perfect design and the naive LASSO with additive covariate measurement error. We performed a Monte Carlo simulation study to assess the finite sample performance of the OLS and the LASSO under matrix uncertainty. Thereby, we also computed the corresponding corrected estimates. In particular, we used the well-known reliability ratio (Fuller, 1987) for the OLS estimates and a reliability ratio-like factor according to Sørensen et al. (2014) for the naive LASSO estimates in the presence of measurement error.

In summary, we found that the MSE values of both the naive LASSO and the naive OLS increase with growing measurement error variance and covariate correlation. With respect to the corrected LASSO and OLS estimates, our results suggest that there does not exist any overall evidence of the efficacy of the applied measurement error correction factors. Especially, the MSE values of the corrected estimates tend to be larger than the ones for the naive estimators. However, we found that the empirical averages of the MSE values were inflated by a few outliers and that the occurrence of outliers was due to the bad conditioning of a matrix which contributes to the correction factor. Moreover, our simulation results suggest that the coincidence of high covariate correlation and large measurement error variance leads the LASSO to be more or less unable to differ between important and unimportant covariates. This finding is also supported by theory. Finally, we also examined the distributions of the tuning parameter constituted by 500 runs of the simulation. We found that the general level of penalization measured by the median of the λ-distributions rises with increasing noise level and that the considered distributions are far from being Gaussian.

Contents

1 Introduction 1

2 Background and definition of the LASSO for high-dimensional data 5

3 Theory for the LASSO 10
- 3.1 Assumptions and corresponding notation . . . 11
- 3.2 Literature review on the LASSO for estimation and variable selection . . . 14
- 3.3 Theoretical properties of the LASSO for estimation and variable selection . 18
  - 3.3.1 Conditions for consistency of prediction and estimation . . . 19
  - 3.3.2 Conditions for consistency of covariate selection . . . 24
- 3.4 A short note on tuning parameter selection for the LASSO . . . 33

4 Multiple linear regression with additive measurement error in the covariates 34
- 4.1 Measurement error model setup . . . 35
- 4.2 Methods to correct for additive measurement error in the covariates . . . 42
  - 4.2.1 Error correction using the method of moments . . . 43
  - 4.2.2 Error correction using replicate data . . . 44

5 Analytical results for the linear LASSO with additive measurement error in the covariates 46
- 5.1 Model Setup . . . 47
- 5.2 Prediction and estimation accuracy . . . 50
- 5.3 Variable selection accuracy . . . 52
- 5.4 Correction methods for the LASSO with additive measurement error in the covariates . . . 55

6 Simulation study for the LASSO with additive measurement error in the covariates 57
- 6.1 Simulation model setup and methods . . . 57
- 6.2 Estimating LASSO in R with the packages `GLMNET` and `penalized` . . . 60
- 6.3 Simulation results . . . 62
  - 6.3.1 Results with respect to parameter estimation consistency . . . 63
  - 6.3.2 Results with respect to variable selection consistency . . . 71
  - 6.3.3 Results with respect to tuning parameter selection . . . 75

7 Discussion and conclusion 84

8 References 90

A Appendix 95
A.1 Some important results from matrix algebra 95
A.2 Simulation results 97
A.2.1 Tables containing MSE, bias, TPF and FPF (n=100) 97
A.2.2 Density plots for the tuning parameter λ (n=100) 109
A.2.3 Tables containing MSE, bias, TPF and FPF (n=500) 116
A.2.4 Density plots for the tuning parameter λ (n=500) 128
A.2.5 Comparison of $LASSO_{\texttt{cv.glmnet}}$ and $LASSO_{\texttt{OptL1}}$ 133

List of Figures

1 Density plots for the tuning parameters $\lambda_{\texttt{OptL1}}$and $\lambda_{\texttt{cv.glmnet}}$ ($n = 100$) . . . 78
2 Density plots for the tuning parameters $\lambda_{\texttt{OptL1}}$and $\lambda_{\texttt{cv.glmnet}}$ ($n = 500$) . . . 79
3 Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=10) 80
4 Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=50) 81
5 Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=500, p=50) 82
6 Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=150) 83
7 Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=100, p=10) 110
8 Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=10) 111
9 Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=100, p=50) 112
10 Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=50) 113
11 Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=100, p=150) 114
12 Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=150) 115
13 Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=500, p=50) 129
14 Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=500, p=50) 130
15 Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=500, p=250) 131
16 Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=500, p=250) 132

List of Tables

1 Composition of the parameter vectors $\boldsymbol{\beta}_k$ ($k = 1, 2, 3$) . . . 58
2 Parameter combinations involved in the MC simulation study . . . 61
3 Bias of naive OLS estimator for $\boldsymbol{\beta}_3$ . . . 66
4 Bias of naive $LASSO_{\texttt{cv.glmnet}}$ estimator for $\boldsymbol{\beta}_3$ . . . 66
5 MSE of naive OLS estimator for $\boldsymbol{\beta}_3$ . . . 66
6 MSE of naive $LASSO_{\texttt{cv.glmnet}}$ estimator for $\boldsymbol{\beta}_3$ . . . 67
7 Bias of OLS estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ . . . 68
8 Bias of $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ . . . 68
9 MSE of OLS estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ . . . 69
10 MSE of $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ . . . 69
11 Bias of OLS estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ . . . 69
12 Bias of $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ . . . 69
13 MSE of OLS estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ . . . 69
14 MSE of $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ . . . 70
15 TPFs of naive $LASSO_{\texttt{cv.glmnet}}$ estimator for $\boldsymbol{\beta}_3$ . . . 73
16 FPFs of naive $LASSO_{\texttt{cv.glmnet}}$ estimator for $\boldsymbol{\beta}_3$ . . . 73
17 TPFs of $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ . . . 74
18 FPFs of $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ . . . 74
19 TPFs of $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ . . . 74
20 FPFs of $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ . . . 74
21 Summary statistics for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ . . . 77
22 Summary statistics for the tuning parameter $\lambda_{\texttt{OptL1}}$ . . . 77
23 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_1$ (n=100, $\rho = 0.9$, $\epsilon = 1$) . . . 97
24 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_2$ (n=100, $\rho = 0.9$, $\epsilon = 1$) . . . 98
25 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_3$ (n=100, $\rho = 0.9$, $\epsilon = 1$) . . . 99
26 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_1$ (n=100, $\rho = 0.5$, $\epsilon = 1$) . . . 100
27 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_2$ (n=100, $\rho = 0.5$, $\epsilon = 1$) . . . 101
28 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_3$ (n=100, $\rho = 0.5$, $\epsilon = 1$) . . . 102
29 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_1$ (n=100, $\rho = 0.1$, $\epsilon = 1$) . . . 103
30 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_2$ (n=100, $\rho = 0.1$, $\epsilon = 1$) . . . 104
31 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_3$ (n=100, $\rho = 0.1$, $\epsilon = 1$) . . . 105
32 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_1$ (n=100, $\rho = 0.5$, $\epsilon = 2.5$) . . . 106
33 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_2$ (n=100, $\rho = 0.5$, $\epsilon = 2.5$) . . . 107
34 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_3$ (n=100, $\rho = 0.5$, $\epsilon = 2.5$) . . . 108
35 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_1$ (n=500, $\rho = 0.9$, $\epsilon = 1$) . . . 116
36 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_2$ (n=500, $\rho = 0.9$, $\epsilon = 1$) . . . 117
37 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_3$ (n=500, $\rho = 0.9$, $\epsilon = 1$) . . . 118
38 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_1$ (n=500, $\rho = 0.5$, $\epsilon = 1$) . . . 119
39 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_2$ (n=500, $\rho = 0.5$, $\epsilon = 1$) . . . 120
40 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_3$ (n=500, $\rho = 0.5$, $\epsilon = 1$) . . . 121
41 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_1$ (n=500, $\rho = 0.1$, $\epsilon = 1$) . . . 122
42 MSE, bias, TPF and FPF for $\boldsymbol{\beta}_2$ (n=500, $\rho = 0.1$, $\epsilon = 1$) . . . 123

43 MSE, bias, TPF and FPF for β_3 (n=500, $\rho = 0.1$, $\epsilon = 1$) 124
44 MSE, bias, TPF and FPF for β_1 (n=500, $\rho = 0.5$, $\epsilon = 2.5$) 125
45 MSE, bias, TPF and FPF for β_2 (n=500, $\rho = 0.5$, $\epsilon = 2.5$) 126
46 MSE, bias, TPF and FPF for β_3 (n=500, $\rho = 0.5$, $\epsilon = 2.5$) 127
47 Mean absolute difference between $LASSO_{\texttt{cv.glmnet}}$ and $LASSO_{\texttt{penalized}}$ for β_1 133
48 Mean absolute difference between $LASSO_{\texttt{cv.glmnet}}$ and $LASSO_{\texttt{penalized}}$ for β_2 134
49 Mean absolute difference between $LASSO_{\texttt{cv.glmnet}}$ and $LASSO_{\texttt{penalized}}$ for β_3 135

List of abbreviations

Abbreviation	Meaning
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BLUE	Best Linear Unbiased Estimator
CC	Compatibility Condition
DGP	Data generating process
EMEV	Estimated measurement error variance
FPF	False positive fraction
GSC	General Sign Consistency
IC	Irrepresentable Condition
IC-ME	Irrepresentable Condition with Measurement Error
i.i.d.	independent and identically distributed
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LS	Least Squares
MC	Monte Carlo
MEC	Measurement Error Condition
ML	Maximum Likelihood
MSE	Mean squared error
MU selectors	Matrix Uncertainty selectors
OLS	Ordinary Least Squares
RCL	Regularized Corrected LASSO
RE	Restricted eigenvalue
SIC	Strong Irrepresentable Condition
SNR	Signal-to-noise ratio
SSC	Strong Sign Consistency
TMEV	True measurement error variance
TPF	True positive fraction
WIC	Weak Irrepresentable Condition

1 Introduction

For various reasons, the popular and well-known *Ordinary Least Squares (OLS)* estimator represents the method of choice in the classical linear regression setting where the number of observations n exceeds or at least equals the number of model covariates p. However, in practice arise potentially many situations that deviate from the classical regression setting and thus lead the OLS solutions to be unstable or unidentifiable. For instance, in situations where some of the covariates that are assumed to be independent from each other are de facto perfectly correlated, the problem of multi-collinearity arises and entails rank deficiencies in the design matrix. The latter lead the OLS estimator to collapse (in the extreme case of perfect correlation) or at least to be unstable in the sense that it exhibits a very large variance. Another reason for rank deficiencies occurs in high-dimensional data settings where the number of model covariates p is typically larger than the sample size n. Given such "large p and small n problem", the model parameters cannot be uniquely defined by the OLS estimator. However, as stated by Raskutti et al. (2010), various fields in modern science and engineering such as computational biology, astrophysics, medical imaging, natural language processing and remote sensing, involve collecting data sets whose dimension p exceeds by far the sample size n. In such cases, strong variable selection is desirable in order to obtain interpretable prediction rules and shrinkage is desirable to prevent over-fit (Goeman, 2010). An important property of the *Least Absolute Shrinkage and Selection Operator* (hereinafter referred to as the *LASSO*) proposed by Tibshirani (1996) is that it is applicable even when the number of covariates exceeds the number of observations. This property makes the LASSO a suitable regression method for high-dimensional data settings with $p > n$ or $p \gg n$. Besides the well-known Ridge estimator (Hoerl and Kennard, 1970), the latter can be considered as being the most important shrinkage or regularization method in linear regression.

The LASSO minimizes the sum of squared empirical errors subject to an l_1-penalty for complexity regularization. In this view, the resulting LASSO estimates can be regarded as shrunken LS estimator. Basically, the LASSO seeks to identify a model that not only fits well, but that is also "simple" to avoid large variation which typically occurs in the estimation of complex models with a huge amount of covariates (Zhao and Yu, 2006). With its l_1-geometry, the LASSO has an exceptional position among estimators using a l_q-norm for regularization. The reason therefore is that $q = 1$ is the only value of q for which variable selection takes place, while the optimization problem is still convex and

hence feasible for high-dimensional problems (see Meinshausen and Bühlmann, 2006). This implies that the LASSO does variable selection in the sense that depending on the respective choice of the regularization parameter, the coefficients estimated can exactly be zero for some variables. Its variable selection capability represents the main feature that distinguishes the LASSO from the Ridge estimator (Hoerl and Kennard, 1970).[1] With regard to variable selection, the LASSO meets the requirements of statistical accuracy and computational feasibility also for $p \gg n$ (Bunea et al., 2007). The latter property differs it from conventional variable selection methods, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) which both are infeasible to compute when the number of covariates p is of medium or large size (Bühlmann and Van de Geer, 2011). Furthermore, by doing variable selection for high-dimensional data, the LASSO reflects the assumption that in many applications, such as studies involving microarray or mass spectrum data, the total number of covariates p is large or even much larger than n, while the number of important covariates is typically smaller than n (see Zhang and Huang, 2008). This notion describes the main rationale behind the sparsity assumption inherent to the LASSO. Sparsity plays a role of utmost importance when the LASSO's performance is considered from a theoretical point of view (see section 3). From a practical perspective, the sparsity assumption in the LASSO leads to sparse or "simple" models that are easy to interpret. More precisely, by potentially setting some or even the majority of covariate coefficients exactly to zero, the LASSO produces sparse solutions and thus facilitates model interpretation when the number of covariates exceeds the number of observations.

Due to the fact that the LASSO has the ability to simultaneously select covariates and estimate parameters, it has become one of the most important and widely used models for regularization in high-dimensional linear regression settings. Thus, it gave rise to an extensive and fast growing body of literature over the last decade. Within the LASSO literature, there is one strand of literature that mainly focuses on the properties of the LASSO with regard to prediction and estimation (see Knight and Fu, 2000; Zhang and Huang, 2008; Meinshausen and Bühlmann, 2006; Meinshausen and Yu, 2009). In this context, one of the main results is the *Compatibility Condition (CC)* discussed in subsubsection 3.3.1. The latter allows to establish so-called *oracle results* for prediction and estimation which

1 Note that the Ridge estimator has the advantage of handling multicollinearity much better than the LASSO (see, e.g., Tibshirani, 1996).

concern the LASSO's ability to nearly achieve the risk of infeasible optimal selection in canonical regression (Hansen, 2013). On the other hand, its model selection properties and in particular its variable selection accuracy have been studied by - amongst others - Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Zou (2006) and Bickel et al. (2009). Meinshausen and Bühlmann (2006) introduced the *neighborhood stability condition* which is sufficient and *essentially* necessary (in a sense to be specified) for the variable selection consistency of the LASSO. At least for the classical case where $n > p$, the latter is equivalent to the *Irrepresentable Condition (IC)* which has been introduced simultaneously, but nevertheless independently by Zhao and Yu (2006) and Zou (2006). The IC is sufficient and *essentially* necessary for consistent model selection with the LASSO and it constitutes the central argument regarding its variable selection consistency. Bühlmann and Van de Geer (2009, 2011) formally showed that the IC for variable consistency implies the CC for prediction and estimation accuracy. This indicates that using the LASSO for variable selection constitutes a more sophisticated problem than using it for parameter estimation and prediction (Bühlmann and Van de Geer, 2011).

An assumption that is common to all of the aforementioned studies on the LASSO is that the design matrix $\mathbf{X}$ passed to the LASSO for performing linear regression contains true covariate measurements that do not suffer from additive measurement error. However, in practice where data corrupted by measurement errors or errors-in-variable data are rather the norm than the exception, this assumption does not meet the truth. In fact, error-prone covariate measurements represent a challenge for the majority of conventional statistical models. Moreover, the practitioner might not be aware of the existence or the precise extent of measurement error in certain situations. Therefore, a legitimate and interesting research question is how the LASSO performs with regard to estimation, prediction and variable selection given the presence of measurement error in the design matrix $\mathbf{X}$.

We would like to dedicate our work to the above question and hence focus on the LASSO for linear models in cases where the design matrix passed on to the LASSO suffers from measurement error. In recent literature, this situation has also been denominated by *matrix uncertainty* (Rosenbaum and Tsybakov, 2010, 2013). In that sense, our work differs from the large majority of literature on the LASSO where the authors implicitly assume perfectly measured covariates. We would like to contribute to a relatively recent and emerging strand of literature and study the LASSO under matrix uncertainty from a theoretical and

simulation-based point of view. Although the main rationale behind the LASSO and also its application based on suitable software packages, such as the `GLMNET` or `penalized` package in `R` (R Core Team, 2013), seem straightforward and do not pose major problems, the theory on the LASSO is very technical and - in our view - a very demanding and complex field of research. It is thus sensible to start with some central theoretical results regarding crucial properties of the LASSO under the assumption of perfect covariate measurements (see section 3). We would like to point out that in literature there have been established numerous different, but largely congruent conditions that imply that the LASSO is consistent for estimation and prediction or variable selection, respectively. Given the vast amount of different conditions and approaches, She (2010) summarizes that using the LASSO the design matrix cannot be too far from orthogonal to reach meaningful conclusions. He also brings forward that the many different assumptions used to guarantee the properties of the LASSO make it difficult to compare the theoretical results with each other and that they also seem to be restrictive in applications. We share this perspective adopted by She (2010). Thus, besides giving a rough outline of the different theoretical perspectives in subsection 3.2, we focus on a few main results regarding the estimation and prediction and variable selection accuracy of the LASSO (see section 3). For the sake of brevity we will not get into detail with respect to the theory of measurement error models, which constitutes itself a complex field of research. There is, however, a vast amount of insightful literature that deals with it and we refer the interested reader to the latter and the references therein (see, e.g., Fuller, 1987; Cheng and Van Ness, 1999; Carroll et al., 2006). Our introduction to measurement error models given in section 4 is rather tightly tailored towards the specific measurement error context examined in the simulation study on the LASSO discussed in section 6 of this work.

The organization of the work is as follows. We start in section 2 by giving some background information and a detailed description of the LASSO estimator for high-dimensional linear regression. The most important theoretical results on the LASSO assuming that the design matrix contains true covariate measurements will be presented in section 3. In section 4, we will shortly introduce the basics of measurement error theory for linear models and outline potential correction methods before considering the LASSO under matrix uncertainty from an analytical point of view in section 5. In the final part of this work (section 6), we perform a Monte Carlo (MC) simulation study to assess the finite sample performance of the LASSO under matrix uncertainty. To conclude, section 7 draws a balance and discusses

some limitations of our work. Appendix A contains some helpful results with regard to matrix algebra and provides additional simulation results.

2 Background and definition of the LASSO for high-dimensional data

In the following, we consider a situation where the observed data $(y_i, \mathbf{x}_i)$ $(i = 1, ..., n)$ are realizations of the random variables $(Y_i, \mathbf{X}_i)$. Given a p-dimensional vector of predictor variables or covariates, $\mathbf{x}_i \in \mathbb{R}^p$, and a univariate continuous output $y_i \in \mathbb{R}$ for individual or observation i, in the classical multiple linear regression model the response y_i is linked to the covariate vector $\mathbf{x}_i$ by means of the linear relation $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$, where ϵ_i is a scalar observation noise and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is the underlying true coefficient vector. If we allow for a set of n such observations, the above model can be written in the matrix-vector form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{n \times 1}$ is the response vector, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix containing true covariate values without measurement error (in which row $\mathbf{x}_i \in \mathbb{R}^{1 \times p}$ represents the covariates for the i^{th} observation), $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ is the true, but unknown regression coefficient vector and $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times 1}$ is a vector of observation noise. We assume that the model errors $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$ are i.i.d. (independent and identically distributed) normally distributed with mean zero and variance σ^2, i.e., $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. Given this setting, our goal typically is to quantify the unknown vector of regression coefficients $\boldsymbol{\beta}$ by means of the estimate $\hat{\boldsymbol{\beta}}$. Thereby, we implicitly assume that the model in Equation 1 is correct, i.e., we imply that the relation that underlies our data is truly linear. This assumption is crucial, since in practice we do not have notice of the true underlying model and thus need to differ between the true parameter $\boldsymbol{\beta}$ and the best approximating parameter $\boldsymbol{\beta}^*$. Note that throughout this work, we write vectors and matrices in boldface. In particular, upper case bold letters such as $\mathbf{X}$ are used for matrices. Estimates are denoted by placing a hat over the corresponding letter for the true parameter, e.g., $\hat{\boldsymbol{\beta}}$ is an estimator for $\boldsymbol{\beta}$. Finally, we use capital letters to denote random variables as compared to their corresponding realized values which we denote by small letters.

The most widely-used estimation method to solve the linear model in Equation 1 for the vector of regression coefficients $\boldsymbol{\beta}$ is the *Least Squares (LS)* estimation method (see, e.g.,

Fahrmeir et al., 2013; Hastie et al., 2009) which results in the popular OLS estimator given by

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2}$$

If the design matrix $\mathbf{X}$ has full column rank, the Gram matrix $\mathbf{X}^T\mathbf{X}$ is positive definite (see subsection A.1) which implies that the OLS estimator given by Equation 2 is the unique solution to Equation 1. Note that for the full rank case the OLS estimator can be shown to be the *best linear unbiased estimator (BLUE)*. This means that the LS estimates for the coefficients β_j $(j = 1, ..., p)$ have the smallest variance among all linear unbiased estimates which in turn implies that they also exhibit the smallest mean squared error (MSE) among all linear unbiased estimators (see Gauss-Markov Theorem in, e.g., Hastie et al. (2009, Chapter 3)).Then However, if the columns of $\mathbf{X}$ are linearly dependent, $\mathbf{X}$ has not full column rank and $\mathbf{X}^T\mathbf{X}$ results in being singular. In this case the OLS coefficients in $\hat{\boldsymbol{\beta}}_{OLS}$ given by Equation 2 are not uniquely defined and thus unidentifiable. As a consequence, the OLS estimator does no longer produce meaningful estimations of the true underlying parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$.

In their study on Ridge regression Hoerl and Kennard (1970) consider the behaviour of the OLS estimator in a setting where due to non-orthogonal problems caused by correlations between covariates, the Gram matrix $\mathbf{X}^T\mathbf{X}$ is not nearly a unit matrix. To demonstrate the impact of $\mathbf{X}^T\mathbf{X}$ being ill-conditioned on $\hat{\boldsymbol{\beta}}_{OLS}$, they introduce the squared distance from $\hat{\boldsymbol{\beta}}_{OLS}$ to the true coefficient vector $\boldsymbol{\beta}$, L_1^2 (see Equation 3). The expectation and the variance of L_1^2 are given in Equation 4 and Equation 5, respectively. Thereby, $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p = \lambda_{min} > 0$ denote the eigenvalues of the Gram matrix $\mathbf{X}^T\mathbf{X}$ in descending order.

$$L_1^2 = (\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \tag{3}$$

$$\mathbb{E}(L_1^2) = \sigma^2 \sum_{k=1}^{p} (\frac{1}{\lambda_k}) \tag{4}$$

$$\mathbb{V}(L_1^2) = 2\sigma^4 \sum_{k=1}^{p} (\frac{1}{\lambda_k})^2 \tag{5}$$

Lower bounds for the expectation and the variance are σ^2/λ_{min} and $2\sigma^4/\lambda_{min}^2$, respectively. As long as $\lambda_{min} \neq 0$, it holds that $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{(p\times p)}$ is regular (see subsection A.1) which implies that the inverse of $\mathbf{X}^T\mathbf{X}$ exists and is unique (see subsection A.1). However, it can easily be seen from Equation 4 and Equation 5 that $\hat{\boldsymbol{\beta}}_{OLS}$ becomes more and more unstable in the sense that the expected distance between the true coefficient vector $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_{OLS}$ and also the variance of this distance will tend to be large given that $\mathbf{X}^T\mathbf{X}$ has one or more small eigenvalues λ_k $(k = 1, ..., p)$. The above considerations show the uncertainty in $\hat{\boldsymbol{\beta}}_{OLS}$ when $\mathbf{X}^T\mathbf{X}$ moves from a unit matrix to an ill-conditioned one (Hoerl and Kennard, 1970). In section 3, we will study the importance of the eigenvalues of the Gram matrix $\mathbf{X}^T\mathbf{X}$ more deeply and show that they play a role of utmost importance with regard to the performance of the *LASSO*.

One potential reason that leads to linear dependencies among columns in $\mathbf{X}$ and thus to $\mathbf{X}^T\mathbf{X}$ having one or more small eigenvalues is given by (multi-)collinearity among covariates. The latter arises in situations where some of the covariates that are assumed to be independent from each other are de facto (perfectly) correlated. Hastie et al. (2009) state that the non-full-rank case occurs mostly if one or more qualitative inputs are coded in a redundant fashion. A further reason for rank deficiencies entailing small eigenvalues of $\mathbf{X}^T\mathbf{X}$ typically occurs in high-dimensional regression settings where the number of model covariates p is (considerably) larger than the sample size n ($p > n$ or $p \gg n$, hereinafter also referred to as *large p small n case*). More in particular, the rank of $\mathbf{X}$ cannot be larger than n if $p > n$. This leads $\mathbf{X}^T\mathbf{X}$ to have - at a maximum - n eigenvalues being non-zero or equivalently, to have at least $(p - n)$ eigenvalues equal to zero.
However, in scientific fields, such as computational biology, astrophysics or signal and (medical) image analysis, there are application areas where the large p small n case is inherent to potential research problems (Raskutti et al., 2010). Since the researcher is typically unaware of the variables playing a role for the considered response, there is no natural way to resolve the non-unique representation problem of OLS estimation by, e.g., recoding or dropping redundant columns in $\mathbf{X}$ (see Hastie et al., 2009). In view of the aforementioned deliberations, we can summarize that given high-dimensional regression settings with $p > n$ or given the presence of perfect collinearity among covariates, the OLS

solutions $\hat{\boldsymbol{\beta}}_{OLS}$ for Equation 1 are unidentifiable.

Against this background, in high-dimensional data settings it is a sensible approach to allow for the class of shrinkage estimators which in fact are biased estimators, but typically have a smaller variance than OLS estimators. This means that they sacrifice a little bias to achieve a larger reduction in variance of the predicted values and thus improve the overall prediction accuracy in terms of exhibiting low MSE values. According to Hastie et al. (2009), the MSE of an estimator $\hat{\boldsymbol{\theta}}$ in estimating $\boldsymbol{\theta}$ is defined by

$$MSE(\hat{\boldsymbol{\theta}}) = \mathbb{E}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^2 = \mathbb{V}ar(\hat{\boldsymbol{\theta}}) + \left[\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\right]^2. \tag{6}$$

Besides the well-known Ridge estimator (Hoerl and Kennard, 1970), the LASSO proposed by Tibshirani (1996) can be considered as being the most important shrinkage or regularization method in linear regression. The LASSO estimate is defined by

$$\hat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left(1/n \, \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1\right), \tag{7}$$

where $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{n}(y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2$ and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|$ which equals the sum of absolute values of the vector's entries. Note that $\lambda \geq 0$ is a *penalty parameter* (also *regularization parameter* or *tuning parameter*). The LASSO estimator given in Equation 7 minimizes the sum of squared empirical errors subject to a l_1-penalty for complexity regularization. In this view, the resulting LASSO estimates $\hat{\beta}_{LASSO,j}(\lambda)$ $(j = 1, ..., p)$ can be regarded as shrunken or regularized LS estimator, a fact that is directly reflected in the LASSO's name (Bühlmann and Van de Geer, 2011). The LASSO seeks to identify a model that not only fits well, but that is also "simple" - in the sense of sparse - to avoid large variation which typically occurs in the estimation of complex models that involve a vast amount of covariates (Zhao and Yu, 2006).

The optimization problem in Equation 7 is convex which allows the LASSO to identify the solution and also facilitates efficient computation of the LASSO estimates even if $p \gg n$ (Bühlmann and Van de Geer, 2011). The convexity of optimization problems implies that any locally optimal point is also (globally) optimal (Boyd and Vandenberghe, 2004), while non-convexity entails the presence of local minima. As mentioned before, the fact that its program is convex differs the LASSO from popular model selection methods, such as AIC or BIC which both use a l_0-penalty and therefore are non-convex functions in the parameter

estimate $\boldsymbol{\beta}$. Note that AIC and BIC are infeasible to compute when p is of medium or large size (Bühlmann and Vande Geer, 2011). In contrast, the l_1-penalty used by the LASSO is symmetric, convex on $(0, \text{inf})$ and has singularities at the origin which is necessary for a method to produce sparse solutions (Fan and Li, 2001). However, note that unlike for the Ridge estimator there does not exist any analytic solution for the LASSO. The reason therefore is that the LASSO coefficients are non-linear in the response $\mathbf{y}$. Efron et al. (2004) proposed the *Least Angle Regression (LARS)* algorithm which is a fast iterative algorithm whose regularization path is close to that of the LASSO. The LARS algorithm can be used for computing the entire LASSO regularization path for linear regression models. In doing so, it makes use of the fact that the coefficient profiles are piecewise linear which leads to an algorithm with the same computational cost as the full LS fit on the data (Goeman, 2010).

The amount of regularization that is applied to the LASSO estimate is controlled by the penalty parameter $\lambda \geq 0$. For $\lambda = 0$, the LASSO estimator given by Equation 7 equals the OLS estimator which minimizes the un-regularized empirical loss. In contrast, a very large λ ($\lambda \to \infty$) will in the limit shrink the entire coefficient vector $\hat{\boldsymbol{\beta}}_{LASSO}$ to zero and thus result in the null model. Moderate values of λ will in general result in shrinkage of the solutions towards zero where some coefficients may be estimated as being exactly zero. However, the above described shrinkage behaviour of the LASSO simultaneously implies one of its major advantages and disadvantages. In fact, as stated by Zhang (2009) the l_1-regularization used by the LASSO has two important properties: it shrinks estimated coefficients $\hat{\beta}_j$ corresponding to irrelevant features towards zero, but it also shrinks estimated coefficients corresponding to relevant features towards zero. This implies that the LASSO leads to biased estimates for large covariate coefficients and thus could be suboptimal in terms of estimation risk (Zou, 2006). The aforementioned considerations of Zhang (2009) and Zou (2006) are in line with the results of a study on the asymptotic behaviour of the LASSO estimates (Knight and Fu, 2000). More in particular, Knight and Fu (2000, Theorem 2) show that the non-zero parameters are estimated with some asymptotic bias if $\lambda > 0$. Besides producing biased estimates for covariates with a reasonable large effect on the response, the LASSO also shows poor performance for nearly singular designs. Note that nearly singular designs are designs whose Gram matrix is non-singular but may have one or more small eigenvalues (Knight and Fu, 2000). Generally, nearly singular designs indicate the presence of (multi-)collinearity among the covariates that are part of the model. Tibshirani (1996) showed that given the presence of (multi-)collinearity among

covariates, the LASSO is inferior in terms of prediction performance compared to the Ridge estimator. The Ridge regression shrinks the coefficients of correlated covariates towards each other whereas the LASSO tends to somehow ignore the correlation by picking only one of the correlated covariates and disregarding the remaining ones (Friedman et al., 2010). Hence, if there are several covariates with high pairwise correlation the LASSO only selects one of the correlated covariates (Zou and Hastie, 2005). Finally, as stated by Bühlmann and Van de Geer (2011), it follows from the analysis of the LARS algorithm (Efron et al., 2004) that every model estimated by the LASSO has a cardinality smaller than or equal to $\min(n, p)$. This property of the LASSO might be beneficial for high-dimensional datasets with $p \gg n$, since it allows the LASSO per se to achieve a certain dimensionality reduction (see also subsubsection 3.3.2). On the other hand, this property can also be considered as a caveat of the LASSO, since in cases where more than $\min(n, p)$ covariates have a de facto impact on the response, it leads to a wrongful reduction of dimensionality.

3 Theory for the LASSO

Before considering the performance of the LASSO for linear models in the presence of measurement error in the design, it is sensible to examine its performance under the assumption of the availability of an ideal $\mathbf{X}$ containing perfect covariate measurements. In this section, we thus focus on LASSO without measurement error in the design. In doing so, we will derive that the latter exhibits good theoretical properties in the sense that its prediction error is of the same order of magnitude as the prediction error resulting if the correct sub-model was known in advance (Bühlmann and Van de Geer, 2011; Fan and Li, 2001). We account for the performance of the LASSO in terms of its ability to correctly estimate the coefficients β_j $(j = 1, ..., p)$ and to select the influential covariates by setting the coefficients of unimportant covariates to zero. While the latter performance component refers to the LASSO's consistency in variable selection, the former relates to its estimation consistency. The theoretical results presented in the following thus allow us to assess under which ideal conditions the LASSO will be consistent in estimation and prediction and variable selection given a perfect design matrix. As shown below, such conditions are mostly conditions on the design matrix $\mathbf{X}$ which already suggests that the presence of measurement error in the latter might have an impact on the performance of the LASSO. In our work, we will restrict ourselves to fixed design. However, Bunea et al. (2007) give interesting insights into oracle inequalities for the LASSO with

random design. Therefore, the interested reader is referred to the latter and the references therein.

3.1 Assumptions and corresponding notation

According to the majority of the prevalent literature on the LASSO, we generally assume throughout this work that the true coefficient vector $\boldsymbol{\beta}$ is sparse. This means that we hypothesize that the majority of coefficients $\beta_{LASSO,j}$ is exactly zero. Note that the assumption of sparsity refers to the fact that in high-dimensional data settings the number of covariates p is potentially growing as any power of the number of observations n if $n \to \infty$, while the number of variables de facto impacting the response is growing at most slightly slower than n (Meinshausen and Bühlmann, 2006). The sparsity assumption will be formalized below. Besides sparsity in $\boldsymbol{\beta}$, we hypothesize for simplicity and without loss of generality that the intercept of the considered linear model is zero and that all covariates are centred and measured on the same scale. These assumptions are literally made in all theoretical studies on the LASSO and they are also of practical relevance. They can be achieved approximately by empirical mean centring of the covariate vectors $\mathbf{X}^{(j)}$ $(j = 1, ..., p)$ and the response vector $\mathbf{Y}$ and scaling with the corresponding empirical standard deviation $\hat{\sigma}_j$. Such standardization procedure guarantees that each covariate is affected more or less equally by the penalization (Bühlmann and Van de Geer, 2011). Bühlmann and Van de Geer (2011) point out that the normalization with the standard deviation is quite natural and commonly used in practice. After these steps, the standardized data to which the LASSO is applied, satisfies $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i = 0$ and $\hat{\sigma}_j^2 = n^{-1} \sum_{i=1}^{n} (X_i^{(j)} - \bar{X}^{(j)})^2 = 1$ for all j. Note that with having only ones on the diagonal the Gram matrix for the covariates $\mathbf{C}_{xx}$ corresponds rather to a correlation matrix than to a covariance matrix. An alternative approach to the above procedure is to keep an intercept, but leave the latter unpenalized (Bühlmann and Van de Geer, 2011).We now introduce the general notation used throughout our work. Basically, we follow the notation used in Sørensen et al. (2014) as well as in Zhao and Yu (2006). However, the dependence on n is neglected notationally and we also make some small notational adjustments where necessary. Some notational refinements, which are necessary for the examination of the LASSO with measurement error, will be given in section 4 and section 5.

Without loss of generality, we assume that

$$\boldsymbol{\beta} = (\beta_1, ..., \beta_q, \beta_{q+1}, ..., \beta_p)^T$$

is the true, but unknown coefficient vector where $\beta_j \neq 0$ for $j = 1, ..., q$ and $\beta_j = 0$ for $j = q+1, ..., p$. Given the sparsity inherent to $\boldsymbol{\beta}$, we partition the set of existing covariates by defining $S_0 = \{1, ..., q\}$ and $S_0^c = \{q+1, ..., p\}$. As a consequence, one has $|S_0| = q$ and $|S_0^c| = (p-q)$, where $|\cdot|$ denotes the cardinality of a set. We will call S_0 the *active set* and q the corresponding *sparsity index* of $\boldsymbol{\beta}$. Furthermore, we partition $\boldsymbol{\beta}$ according to the active set S_0 by defining $\boldsymbol{\beta}_{S_0} = (\beta_1, ..., \beta_q)^T \in \mathbb{R}^q$ and $\boldsymbol{\beta}_{S_0^c} = (\beta_{q+1}, ..., \beta_p)^T \in \mathbb{R}^{(p-q)}$. Thus, $\boldsymbol{\beta}_{S_0}$ is the part of the true coefficient vector that contains non-zero coefficients belonging to influential covariates, while $\boldsymbol{\beta}_{S_0^c}$ is the part of the true coefficient vector that contains zero coefficients which belong to unimportant or non-relevant covariates. It is obvious that if we knew the active set S_0, we could simply neglect all unimportant variables $\mathbf{X}^{(j)}$ with $j \notin S_0$. The overall squared accuracy would then be $\frac{\sigma^2}{n} \times q$ (see Bühlmann and Van de Geer, 2011). However, in practice S_0 is typically unknown.

It follows from the assumption of sparsity that $\boldsymbol{\beta} = (\boldsymbol{\beta}_{S_0}^T, \boldsymbol{\beta}_{S_0^c}^T)^T$ and that $\boldsymbol{\beta}_{S_0^c} = \mathbf{0}$. In line with the above definitions, we introduce the partitioning of the design matrix $\mathbf{X} = (\mathbf{X}_{S_0}, \mathbf{X}_{S_0^c})$, where $\mathbf{X}_{S_0} \in \mathbb{R}^{n \times q}$ and $\mathbf{X}_{S_0^c} \in \mathbb{R}^{n \times (p-q)}$ are the first q and the last $(p-q)$ columns of $\mathbf{X} \in \mathbb{R}^{n \times p}$, respectively. Thus, $\mathbf{X}_{S_0}$ contains the n perfect measurements of the q influential covariates, while $\mathbf{X}_{S_0^c}$ contains the n perfect measurements of the $(p-q)$ unimportant covariates.

Throughout this work, sample covariance matrices are denoted by $\mathbf{C}$ and subscripts show which covariates are involved. Thus, let $\mathbf{C}_{xx} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ be the empirical covariance of the covariate measurements or equally the scaled Gram matrix of the design matrix $\mathbf{X}$. Defining the covariance of the q influential covariates as $\mathbf{C}_{xx}(S_0, S_0) = \frac{1}{n}\mathbf{X}_{S_0}^T\mathbf{X}_{S_0} \in \mathbf{R}^{q \times q}$, the covariance of the $p-q$ unimportant covariates as $\mathbf{C}_{xx}(S_0^c, S_0^c) = \frac{1}{n}\mathbf{X}_{S_0^c}^T\mathbf{X}_{S_0^c} \in \mathbf{R}^{(p-q)\times(p-q)}$ and the covariance of the influential covariates with the unimportant covariates as $\mathbf{C}_{xx}(S_0, S_0^c) = \frac{1}{n}\mathbf{X}_{S_0}^T\mathbf{X}_{S_0^c} = \mathbf{C}_{xx}(S_0^c, S_0)^T = (\frac{1}{n}\mathbf{X}_{S_0^c}^T\mathbf{X}_{S_0})^T \in \mathbf{R}^{q\times(p-q)}$, respectively, one can express the scaled Gram matrix $\mathbf{C}_{xx}$ in terms of a block-wise form:

$$\mathbf{C}_{xx} = \begin{pmatrix} \mathbf{C}_{xx}(S_0, S_0) & \mathbf{C}_{xx}(S_0, S_0^c) \\ \mathbf{C}_{xx}(S_0^c, S_0) & \mathbf{C}_{xx}(S_0^c, S_0^c) \end{pmatrix}. \tag{8}$$

Population covariance matrices are denoted by $\boldsymbol{\Sigma}$ and indexed by subscripts and superscripts according to the way described for sample covariance matrices. Thus, we

have

$$\mathbf{\Sigma}_{xx} = \begin{pmatrix} \mathbf{\Sigma}_{xx}(S_0, S_0) & \mathbf{\Sigma}_{xx}(S_0, S_0^c) \\ \mathbf{\Sigma}_{xx}(S_0^c, S_0) & \mathbf{\Sigma}xx(S_0^c, S_0^c) \end{pmatrix}, \tag{9}$$

where $\mathbf{\Sigma}_{xx}(S_0, S_0) \in \mathbb{R}^{q \times q}$, $\mathbf{\Sigma}_{xx}(S_0^c, S_0) \in \mathbb{R}^{(p-q) \times q}$, $\mathbf{\Sigma}_{xx}(S_0, S_0^c) \in \mathbb{R}^{q \times (p-q)}$ and $\mathbf{\Sigma}_{xx}(S_0^c, S_0^c) \in \mathbb{R}^{(p-q) \times (p-q)}$.

Unless otherwise specified, we assume fixed true covariates that satisfy

$$(1/n)\ \mathbf{X}^T\mathbf{X} = \mathbf{C}_{xx} \to \mathbf{\Sigma}_{xx}\ , \text{as } n \to \infty \tag{10}$$

and

$$(1/n)\ \max_{1 \le i \le n} (\mathbf{x}_i^T \mathbf{x}_i) \to 0\ , \text{as } n \to \infty, \tag{11}$$

where $\mathbf{\Sigma}_{xx}$ is a positive definite matrix. As mentioned before, we typically assume that the covariates are scaled so that the diagonal elements of $\mathbf{C}_{xx}$ are all identical 1. The convergences in Equation 10 and Equation 11 are deterministic. [2]

The LASSO estimates are partitioned according to the same pattern as $\boldsymbol{\beta}$. This means that we have $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{S_0}^T, \hat{\boldsymbol{\beta}}_{S_0^c}^T)^T = \hat{\boldsymbol{\beta}}_{Lasso}(\lambda)$ where $\hat{\boldsymbol{\beta}}_{S_0} \in \mathbb{R}^q$ and $\hat{\boldsymbol{\beta}}_{S_0^c} \in \mathbb{R}^{(p-q)}$. Note that according to the above definition, the dependence of $\hat{\boldsymbol{\beta}}$ on λ is implicit. Furthermore, we would like to point out that since the LASSO does not necessarily provide correct estimations of the true $\boldsymbol{\beta}$, the elements of $\hat{\boldsymbol{\beta}}_{S_0}$ are not necessarily non-zero, neither are the elements of $\hat{\boldsymbol{\beta}}_{S_0^c}$ necessarily zero. To differentiate the true active set S_0 from the active set of the LASSO, we define for any $\lambda \ge 0$, $\hat{S}(\lambda) = \{j : \hat{\beta}_{LASSO,j}(\lambda) \neq 0\}$ being the active set of the LASSO. $\hat{S}(\lambda)$ contains all non-zero LASSO coefficient estimates, i.e., the $\hat{\beta}_{LASSO,j}$'s for all covariates that the LASSO deems to have an impact on the response.

In the following, we use the notation $\|\cdot\|_q$ for the l_q-norm of a vector. The l_q-norm of a vector $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_q = \sqrt[q]{\sum_i |x_i|^q},\ \text{where } q \in \mathbb{R}. \tag{12}$$

2 The above regularity conditions have been introduced in the literature by Knight and Fu (2000). However, they have also been used by Zhao and Yu (2006) and Sørensen (2014), among others.

In line with the above definition, it holds for the l_1-norm of $\mathbf{x}$ that $\|\mathbf{x}\|_1 = \sum_i |x_i|$. This implies that the l_1-norm of $\mathbf{x}$ equals the sum of absolute values of the vector's entries. The l_2-norm of $\mathbf{x}$ is defined as $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$.

3.2 Literature review on the LASSO for estimation and variable selection

Although directly after its publication in Tibshirani (1996) the LASSO received only very little attention (Tibshirani, 2011), it has become a very popular and widely used model for regularization in the context of high-dimensional regression during the last years. One of the reasons for its gain in popularity is undoubtedly its ability to simultaneously select covariates and estimate parameters. Given an appropriate amount of regularization, the LASSO was further shown to have *oracle procedure properties*. The term *oracle procedure* refers to the fact that in terms of estimating the zero and the non-zero components of $\boldsymbol{\beta}$, the LASSO estimator works as well as if the correct sub-model was known in advance (Fan and Li ,2001). Given the aforementioned reasons, the LASSO gave rise to an extensive and still growing body of literature over the last decade. In view of the large number of theoretical studies which lead to a broad variety of partly competing findings and results, it is reasonable to first get an overview of the most important LASSO studies and the results presented therein. Note that the vast majority of LASSO studies focuses either on the estimation and prediction accuracy (see, e.g., Bunea et al., 2007; Zhang and Huang, 2008; Bickel et al., 2009; Meinshausen and Yu, 2009) or on the variable selection capacity of the LASSO (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Zou, 2006). However, note that the LASSO's oracle capacity simultaneously implies the accuracy for correctly estimating the zero coefficients (variable selection) and the accuracy for correctly assessing the non-zero coefficients (Fan and Li, 2001). The reason therefore simply is that identifying the significant predictors will also enhance the prediction performance of the fitted model (Zou, 2006).

Knight and Fu (2000) were the first to thoroughly study the asymptotic properties of LASSO-type estimates in the low-dimensional setting where the number of parameters p is fixed and smaller than the sample size n. In doing so, they showed that under appropriate conditions for the design matrix $\mathbf{X}$, the LASSO estimates are consistent for the coefficients β_j $(j = 1, ..., p)$. Moreover, they derived that with an appropriate choice for the tuning

parameter ($\lambda_n \propto n^{1/2}$ as $n \to \infty$), the limiting distributions of the estimated coefficients can have positive probability mass at 0 if $\beta_j = 0$. This implies that given an appropriate amount of regularization there is a non-vanishing positive probability for the LASSO to select the true model. However, Zhang and Huang (2008) state that on closer inspection of the results presented by Knight and Fu (2000), the positive probability mass at zero is less than one in the limit for certain configurations of the covariates and regression coefficients. This means that the LASSO does not perform consistent variable selection without proper assumptions.

Meinshausen and Bühlmann (2006) studied the LASSO for neighbourhood selection in Gaussian graphical models. They showed that given the sparsity of $\boldsymbol{\beta}$, the LASSO is computationally very efficient and consistent even in the estimation of high-dimensional graphs if a so-called *neighbourhood stability condition* is satisfied. They also insisted on the importance of suitably choosing the penalty parameter λ for consistency of the LASSO. In particular, they demonstrated that an prediction-optimal oracle penalty obtained through a cross-validated choice λ_{cv}, leads the LASSO to include many noise variables in the neighbourhood estimate and therefore does not allow for consistent neighbourhood estimation. In fact, the probability of including noise variables using λ_{cv} does not even vanish asymptotically for a fixed number of variables p (Meinshausen and Bühlmann, 2006, Proposition 1). To achieve consistency in neighbourhood selection, λ must be chosen larger than the prediction-optimal value. In addition, certain regularity conditions must be satisfied for the covariate covariance matrices. More specifically, the latter must be non-singular, show a common empirical variance and satisfy the neighbourhood stability assumption. Meinshausen and Bühlmann (2006) demonstrated that under the aforementioned conditions the LASSO is consistent in estimating the dependency between Gaussian variables even when p grows faster than n.

Zhao and Yu (2006) and Zou (2006) studied the LASSO's variable selection consistency and independently from each other obtained the same results. They formalized the *neighbourhood stability condition* according to Bühlmann and Meinshausen (2006) in the context of linear regression as so-called *irrepresentable condition (IC)*. In doing so, they showed that given the lC and certain other regularity conditions on the design matrix, the LASSO is sign-consistent with the convention $sign(0) \stackrel{!}{=} 0$ and selects exactly the set of non-zero regression coefficients, provided that these are bounded away from zero at a certain rate.

Under the above conditions, the LASSO is consistent for variable selection even when the number of covariates p is as large as $\exp(n^a)$ for some $0 < a < 1$. For being the central argument in favour of the LASSO's variable selection consistency, we will perform an in-depth analysis of the IC according to Zhao and Yu (2006) in subsubsection 3.3.2.

A l_2-norm parameter estimation error bound for the LASSO together with the corresponding convergence rate has been derived in studies conducted by Zhang and Huang (2008) and Meinshausen and Yu (2009). In their work, Zhang and Huang (2008) altered the prevalent model assumptions by using a different definition of sparsity. The latter implies that the regression coefficients outside the true model are small in the sense that the sum of their absolute values is below a certain level, but that they are not necessarily zero. This clearly differs the analysis performed by Zhang and Huang (2008) from the study on the l_2-consistency of the LASSO proposed by Meinshausen and Bühlmann (2006) and also from the above studies on the LASSO's variable selection consistency. As opposed to Zhang and Huang (2008), the aforementioned studies imply that the coefficients of the covariates outside the true model are exactly zero and that all non-zero coefficients are uniformly bounded away from zero at a certain rate. Given the particular sparsity assumption used in Zhang and Huang (2008), variable selection no longer corresponds to distinguishing between non-zero and zero coefficients. Instead, the objective is to select a sparse model which fits the mean vector $\mathbf{X}\boldsymbol{\beta}$ well and thus ideally includes all variables with large coefficients $|\beta_j|$. Against the background of this objective, Zhang and Huang (2008) established a set of sufficient conditions which manages without the IC proposed by Zhao and Yu (2006) and Zou (2006). Instead, they proposed a *sparse Riesz condition* on the correlation of the design variables which limits the range of the eigenvalues of the covariance matrices of all subsets of a fixed number of covariate vectors $\mathbf{x}_j$. Zhang and Huang (2008) demonstrated that under the latter, the LASSO selects a model with the correct order of dimension whose bias is bounded. Note that all variables with coefficients above a pre-defined threshold level are selected regardless of the values of the other coefficients. Zhang and Huang (2008) also showed that the LASSO is rate-consistent in the sparsity and the bias even in high-dimensional regression settings.[3] They concluded that with regard to rate-consistency in model selection, the performance of the LASSO for correlated designs under the sparse Riesz condition is comparable to its performance in the much simpler orthogonal designs. It is worth noting that the sparse Riesz condition used by Zhang and

3 For the corresponding definition of rate consistency we refer to Zhang and Huang (2008, p. 1571).

Huang (2008) and the *Strong Irrepresentable Condition (SIC)* derived by Zhao and Yu (2006) and Zou (2006) do not imply each other in general.
An alternative key condition on the covariate covariance matrix which is similar to the sparse Riesz condition introduced in Zhang and Huang (2008) is given by the *Restricted Eigenvalue assumption (RE assumption)* proposed in a simultaneous study on the LASSO and the Dantzig selector conducted by Bickel et al. (2009). Basically, the RE assumption constitutes a condition on the Gram matrix that - similar to the sparse Riesz condition proposed by Zhang and Huang (2008) - reflects that requiring the entire Gram matrix to be positive definite might often be too restrictive. Under the RE assumption it is hence sufficient that small sub-matrices of the Gram matrix are non-singular. Raskutti et al. (2010) comment on the RE assumption introduced by Bickel et al. (2009) that it is one of the weakest known sufficient conditions presented so far for bounding the l_2-error of the LASSO. Note that the definition of sub-matrices depends on the sparsity of the true $\boldsymbol{\beta}$. However, in practice the sparsity set S_0 and also its cardinality $q = |S_0|$ are unknown. This implies that the RE assumption itself requires an assumption about the sparsity of the considered model. For a more detailed description of the RE assumption, we refer to Bickel et al. (2009). We would like to point out here that the general notion of the RE assumption is largely reflected in the IC proposed by Zhao and Yu (2006) and Zou (2006). The latter will be discussed more in detail in subsubsection 3.3.2.

Meinshausen and Yu (2009) argument that the IC, which is the central condition for consistent variable selection of the LASSO (see subsubsection 3.3.2), can easily be violated in the presence of correlation among the covariates of a model. Therefore, they study the behaviour of the LASSO if the IC is relaxed. The relaxed condition used by Meinshausen and Yu (2009) regards the number of non-zero components of $\boldsymbol{\beta}$ and also the minimal singular values of the design sub-matrices that contain small subsets of covariates. In the prevalent LASSO literature, the relaxed condition is commonly referred to as the *condition of bounded minimal and maximal sparse eigenvalues*. If the latter is satisfied and at the same time an appropriate amount of shrinkage is applied, Meinshausen and Yu (2009) show that the LASSO is still consistent in the l_2-norm sense for fixed designs even if the IC is violated.[4] The corresponding rate of convergence can even be considered as optimal in the sense that - apart from a logarithmic factor in p and n - it corresponds to the rate

4 Note that according to Meinshausen and Yu (2009), an estimator is said to be l_2-consistent if $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{l_2} \to 0$ as $n \to \infty$.

that could be achieved if the true sparse model was known (Meinshausen and Yu, 2009). However, the LASSO is unable to recover the correct sparsity pattern of the model under the relaxed conditions. More precisely, it tends to select the non-zero entries of $\boldsymbol{\beta}$ and, in addition, some zero entries. Meinshausen and Yu (2009) emphasize in this context that the non-zero entries of $\boldsymbol{\beta}$ are in any case included in the model selected by the LASSO as long as the model covariates are linearly independent. It should be noted that in view of linear dependencies among the covariates, even the relaxed condition proposed by Meinshausen and Yu (2009) does not hold.

Van de Geer and Bühlmann (2009) revisit some sufficient conditions (together with their interrelations) for oracle inequalities for the LASSO in regression. They argue that the different sufficient conditions for oracle inequalities hold in fairly general situations and thus also allow for a fairly general class of design matrices. However, at large all of the partly very technical theories roughly outlined above have in common that they are based on some rather strong working assumptions with respect to the Gram matrix $\mathbf{X}^T\mathbf{X}$. She (2010) states that - in some sense - the design matrix simply cannot be too far from orthogonal to reach meaningful conclusions when using the LASSO. Yet, the multitude of different assumptions makes it difficult to compare the theoretical results with each other. Furthermore, such assumptions seem to be applicable in practice only to a limited extent. We share the perspective adopted by She (2010) and thus focus on two central results regarding the consistency of the LASSO in estimation and variable selection. The actual application of the LASSO will be deferred to section 6 where we present the results of our simulation study for the LASSO with measurement error in the design matrix.

3.3 Theoretical properties of the LASSO for estimation and variable selection

Under the assumption of sparsity for the true parameter vector $\boldsymbol{\beta}$, we can use the LASSO to simultaneously select covariates and estimate their corresponding coefficients β_j. For both of these features, we require the LASSO to exhibit good theoretical properties. In particular, we would like the LASSO to be consistent in parameter estimation, i.e., we stipulate that $(\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}^n) \xrightarrow{p} 0$, as $n \to \infty$. At the same time, we want the LASSO to be consistent in variable selection which implies that $P(\{i : \hat{\beta}_i^n \neq 0\} = \{i : \beta_i^n \neq 0\}) \to 1$, as $n \to \infty$.

Ideally - of course - we would like the LASSO to achieve both types of consistency at the same time. Unfortunately it is known from practice that an estimate which is consistent in terms of parameter estimation does not necessarily consistently select the true model and vice versa (Zhao and Yu, 2006). Given this background and also the fact that the prevalent literature mostly treats both features of the LASSO separately, we divide this subsection in an estimation consistency (subsubsection 3.3.1) and a variable selection consistency (subsubsection 3.3.2) part. In subsubsection 3.3.1, we state the CC (Bühlmann and Van de Geer, 2011) on the design matrix. The latter is sufficient and necessary for the LASSO to be consistent in estimation. As aforementioned, the CC does not imply the variable selection consistency of the LASSO which means that under only a CC, the LASSO does not perform consistent variable selection (Bühlmann and Van de Geer, 2011). In subsubsection 3.3.2, we present the SIC (Zhao and Yu, 2006) which is sufficient and *essentially* necessary (in a sense to be specified) for the LASSO to perform consistent variable selection.

3.3.1 Conditions for consistency of prediction and estimation

It can be shown that the LASSO enjoys good theoretical properties in the sense that its prediction error is about the same magnitude as the prediction error one would have if one knew a priori which covariates have an influence on the response compared to those that are not part of the true model (Bühlmann and Van de Geer, 2011). This notion directly leads to the aforementioned oracle inequality for the LASSO. Basically, the latter constitutes a probability inequality for the LASSO's error term. In the following, we consider the case of squared error loss with fixed design and present some of the main theoretical arguments for establishing such oracle results for the LASSO. Together with the CC on the design matrix $\mathbf{X}$, the latter provide the main argument that evidence the consistency of the LASSO for estimating the parameter vector $\boldsymbol{\beta}$. We explicitly assume that the linear model in Equation 1 holds exactly. The parameter vector $\boldsymbol{\beta}$ comprises the true, but unknown covariate coefficients. Generally, the results presented below are quoted from Bühlmann and Van de Geer (2011, Chapter 6.2.1 - 6.2.2). Note that for the sake of consistent notation throughout the entire work, we adjusted them notationally where necessary. In practice - of course - we are not able to assure if our assumption of the linear model holding exactly is truly correct. Thus, for corresponding oracle results in cases where the assumed model only represents a linear approximation of the truth, the interested reader is referred to Bühlmann and Van de Geer (2011, Chapter 6.2.3) and the references therein.

The technical base for all derivations concerning the estimation and prediction error of the LASSO is constituted by the *basic inequality* shown in Equation 13. The latter exploits the fact that the LASSO estimator is a penalized empirical risk minimizer which implies that its penalized empirical risk is less than or equal to the penalized empirical risk of any other parameter choice (Bühlmann and Van de Geer, 2011). Note that Equation 14 can be obtained by simply rewriting the basic inequality given by Equation 13.

$$\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LASSO}\|_2^2/n + \lambda\|\hat{\boldsymbol{\beta}}_{LASSO}\|_1 \leq \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/n + \lambda\|\boldsymbol{\beta}\|_1 \tag{13}$$

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{LASSO} - \boldsymbol{\beta})\|_2^2/n + \lambda\|\hat{\boldsymbol{\beta}}_{LASSO}\|_1 \leq 2\boldsymbol{\epsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_{LASSO} - \boldsymbol{\beta})/n + \lambda\|\boldsymbol{\beta}\|_1 \tag{14}$$

$$2|\boldsymbol{\epsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_{LASSO} - \boldsymbol{\beta})| \leq \left(\max_{1\leq j\leq p} 2|\boldsymbol{\epsilon}^T\mathbf{X}^{(j)}|\right)\|\hat{\boldsymbol{\beta}}_{LASSO} - \boldsymbol{\beta}\|_1 \tag{15}$$

$$\mathcal{J} := \{\max_{1\leq j\leq p} 2|\boldsymbol{\epsilon}^T\mathbf{X}^{(j)}|/n \leq \lambda_0\} \tag{16}$$

For the case of quadratic loss, the term $2\boldsymbol{\epsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_{LASSO} - \boldsymbol{\beta})/n$ on the right hand side of Equation 14 is referred to as the *empirical process part* of the problem. The reason therefore simply is that it constitutes the term where the random noise vector $\boldsymbol{\epsilon}$ plays a role. According to Equation 15, we can bound the empirical process part for quadratic loss in terms of the l_1-norm of the parameters involved. It is clear that when using the LASSO, we require its inherent penalization mechanism to correct for the effects caused by the empirical process part, since we are clearly interested only in that portion of the difference between $\hat{\boldsymbol{\beta}}_{LASSO}$ and $\boldsymbol{\beta}$ which is caused by systemic properties of the LASSO. Given the upper bound of the empirical process part quantified in Equation 15, one can define the set $\mathcal{J}$ in Equation 16. The latter contains all regularization parameters λ_0 that are at least as large as the empirical process and that are therefore supposed to "overrule" the random part. According to Bühlmann and Van de Geer (2011), we conservatively assume for all further derivations that the tuning parameter λ satisfies $\lambda \geq 2\ \lambda_0$. This assumption is arbitrary to some extent, but serves the purpose of assuring that the set $\mathcal{J}$ is adjusted for the random influence. It can be shown that with Gaussian errors and λ_0 being chosen of order $\sqrt{\log(p)/n}$, the set $\mathcal{J}$ has a large probability (see Bühlmann and Van de Geer, 2011,

Lemma 6.2.).

Building on Equation 14 and applying a chain of several mathematical statements and derivations to the latter,[5] one obtains the CC on the design matrix $\mathbf{X}$. The CC is a crucial condition needed to establish results with respect to the estimation and prediction consistency of the LASSO. It is met for the set S_0, if for some $\phi_0 > 0$ and for all $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}_{S_0^c}\|_1 \leq 3\|\boldsymbol{\beta}_{S_0}\|_1$, it holds that

$$\|\boldsymbol{\beta}_{S_0}\|_1^2 \leq \frac{\left(\boldsymbol{\beta}^T \mathbf{C}_{xx}\, \boldsymbol{\beta}\right) q}{\phi_0^2}. \tag{17}$$

Note that ϕ_0^2 is a *compatibility constant* that equals a lower bound for the RE of the matrix $\mathbf{C}_{xx}$ (see Bickel et al. (2009) presented in subsection 3.2). At best, ϕ_0^2 is bounded from below by a positive constant, since very small compatibility constants or REs imply that the design $\mathbf{X}$ contains correlated covariates.[6]

The above presented CC can be interpreted as identifiability assumption in terms of the l_1-norm of the influential coefficients in the model. If one replaces $\|\boldsymbol{\beta}_{S_0}\|_1^2$ in Equation 17 by its upper bound $q\|\boldsymbol{\beta}_{S_0}\|_2^2$, the CC resembles a condition on the smallest eigenvalue of the Gram matrix $\mathbf{C}_{xx}$ (Bühlmann and Van de Geer, 2011). As illustrated in section 2, the smallest eigenvalue of $\mathbf{C}_{xx}$ must be non-zero when working with the OLS method, since otherwise $\mathbf{X}^T\mathbf{X}$ is singular and the inverse of $\mathbf{X}^T\mathbf{X}$ does not exist (see subsection A.1). However, in Equation 17 the restriction $\|\boldsymbol{\beta}_{S_0^c}\|_1 \leq 3\|\boldsymbol{\beta}_{S_0}\|_1$ limits the set of coefficient vectors $\boldsymbol{\beta}$ for which Equation 17 is de facto required to hold. Hence, the CC is actually weaker than imposing non-zero eigenvalues on $\mathbf{C}_{xx}$ (Bühlmann and Van de Geer, 2011). Since in practice one has no information about the nature of S_0, Equation 17 is de facto not applicable. Note that if at least the sparsity index $q = |S_0|$ was known, it would be sufficient to check the above inequalities for all sets $S \subset \{1, ..., p\}$ with cardinality q. This approach directly refers to the aforementioned RE assumption which was proposed by Bickel et al. (2009).

5 For being very technical, those derivations will not be discussed in this work. The interested reader is referred to Bühlmann and Van de Geer (2011, Chapter 6.2.2.)

6 It should be stated that the constant 3 in the term $\|\boldsymbol{\beta}_{S_0^c}\|_1 \leq 3\|\boldsymbol{\beta}_{S_0}\|_1$ is - at least to a certain extent - arbitrary and can thus be replaced by any constant bigger than 1 provided that some other constants (in particular, the lower bound for λ) get also adjusted (see Bühlmann and Van de Geer (2011, Chapter 6.2.2)).

Allowing for the CC and the conditions that are needed for $\mathcal{J}$ to have large probability, it can be shown that the following probabilistic statement holds for the oracle inequality (see Equation 20) of the LASSO with Gaussian errors (Bühlmann and Van de Geer, 2011, Corollary 6.2.):

Oracle results for the LASSO with respect to estimation and prediction
Assume that all diagonal elements of the Gram matrix $\mathbf{C}_{xx}$ correspond to 1 (i.e., the covariates are scaled) and that the CC holds for S_0, with $\mathbf{C}_{xx}$ normalized in this way. For some constant $t > 0$, let the regularization parameter be

$$\lambda = 4\,\hat{\sigma}\sqrt{\frac{t^2 + 2\,log(p)}{n}}, \tag{18}$$

where $\hat{\sigma}^2$ is an estimator of the noise variance σ^2. Then with probability at least $1 - \alpha$, where

$$\alpha = 2\exp\left[-t^2/2\right] + \mathbb{P}\left(\hat{\sigma} \leq \sigma\right), \tag{19}$$

one has

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{LASSO} - \boldsymbol{\beta})\|_2^2/n + \lambda\|\hat{\boldsymbol{\beta}}_{LASSO} - \boldsymbol{\beta}\|_1 \leq \frac{4\,\lambda^2 q}{\phi_0^2}. \tag{20}$$

In summary, the above statement constitutes a probability inequality for the error term of the LASSO with Gaussian errors. It shows that if λ is chosen within a range of the order $\hat{\sigma}\sqrt{\log(p)/n}$, the *oracle inequality* for fixed design (see Equation 20) holds with high probability. In particular, the oracle inequality for the LASSO combines two major results. Fristly, it provides the bound

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{LASSO} - \boldsymbol{\beta})\|_2^2/n \leq \frac{4\,\lambda^2 q}{\phi_0^2} \tag{21}$$

for the prediction error of the LASSO. And secondly, it limits the l_1-error of the LASSO by establishing the bound

$$\|\hat{\boldsymbol{\beta}}_{LASSO} - \boldsymbol{\beta}\|_1 \leq \frac{4\,\lambda q}{\phi_0^2}. \tag{22}$$

Either bound depends on the unknown sparsity index q which reflects the number of

influential covariates in the model, on the compatibility constant ϕ_0^2 which depends on p or n only through the scaled Gram matrix and on the tuning parameter λ. Note that both of the above bounds hold on $\mathcal{J}$ and that given

$$\lambda = 4\,\hat{\sigma}\sqrt{\frac{t^2 + 2\,log(p)}{n}}, \tag{23}$$

the amount of regularization λ satisfies $\lambda \geq 2\,\lambda_0$. This implies that λ is large enough to overrule the noise constituted by the empirical process part in Equation 14. Clearly, the lower bound for $(1-\alpha)$ which determines the probability that Equation 20 holds given an appropriate λ, diminishes if we use an inappropriate estimator for σ. Note that we call an estimator $\hat{\sigma}$ *inappropriate* if it tends to underestimate the true standard deviation of the noise. This is due to the fact that α increases with $\mathbb{P}\left(\hat{\sigma} \leq \sigma\right)$. Generally, we need $\hat{\sigma}$ to be well-calibrated, i.e., neither too small nor too large. According to Bühlmann and Van de Geer (2011), the estimator $\hat{\sigma}^2 = \mathbf{Y}^T\mathbf{Y}/n$ satisfies $\sigma \leq \hat{\sigma} \leq const.\,\sigma$, with the constant (*const.*) well under control for any reasonable signal-to-noise ratio.[7]

If one substitutes λ on the right hand side of Equation 20 by the definition of λ proposed in Equation 23, one obtains

$$\frac{64q\,\hat{\sigma}^2}{n\,\phi_0^2}(t^2 + 2\log\left(p\right)) \propto const.\frac{\hat{\sigma}^2\ \log\left(p\right)}{n}q \tag{24}$$

as (approximate) bound for the term on the left hand side of Equation 20. Note that the additional $\log\left(p\right)$-factor in the transformed bound $const.\frac{\hat{\sigma}^2\ \log\left(p\right)}{n}q$ can be seen as a kind of fine that one has to pay for not knowing a priori the active set q. Recall also that $(\sigma^2/n) \times q$ is the squared accuracy for the prediction error of LS estimation in a situation where the sparsity index q would be known (see Bühlmann and Van de Geer, 2011). Given the above results, we can conclude that - up to the $\log\left(p\right)$-factor and the compatibility constant ϕ_0^2 - the mean-squared prediction error of the LASSO is of the same order as if one knew a priori which covariates are relevant and thus calculated the OLS estimator for the model including the q relevant variables only. In this regard, the rate in Equation 20 is optimal up to the factor $\log\left(p\right)$ and the inverse compatibility constant $1/\phi^2$ (see Bühlmann and Van de Geer, 2011).

7 The signal-to-noise ratio (SNR) is defined by $SNR = \frac{\|\mathbf{X}\boldsymbol{\beta}\|_2}{\sqrt{n}\sigma}$ (Bühlmann and Van de Geer, 2011).

We can conclude that Equation 20 constitutes a probability inequality for the statistical error of the LASSO with fixed design and Gaussian errors. Given an appropriate choice for the tuning parameter λ, Equation 20 holds with high probability. In particular, it provides bounds for the prediction and estimation error of the LASSO. Against the background of Equation 18 to Equation 20, one can make the following statement about the prediction and estimation consistency of the LASSO with true design matrix and Gaussian errors: it follows from the oracle result given by Equation 20 that as long as the number of observations n tends to infinity with a rate faster than $q \log(p)$, the LASSO will be consistent for prediction (see Equation 21). With regard to its estimation consistency, it holds that as long as the number of observations n tends to infinity faster than $q^2 \log(p)$, the LASSO will also be consistent for estimation in the l_1-norm (see Equation 22) (see also Sørensen et al., 2014). Note that the oracle inequality given by Equation 20 can be considered as suboptimal if the linear model contains a lot of non-zero coefficients $|\beta_j|$ that are smaller than the noise level $\sqrt{\sigma^2/n}$. For such models, an oracle bound which is proportional to the number of significantly non-zero β_j times $\sigma^2 log\ p/n$ would be better suited (Bühlmann and Van de Geer, 2011). However, as stated by Bühlmann and Van de Geer (2011), this extension is mathematically of the same nature as the extension where the linear model is not assumed to hold exactly. For the above problem we refer the interested reader to Bühlmann and Van de Geer (2011, Chapter 6.2.3) and the references therein.

3.3.2 Conditions for consistency of covariate selection

In subsubsection 3.3.1, we showed that by imposing a lower bound on the RE of the Gram matrix $\mathbf{C}_{xx}$, the CC leads the LASSO to be consistent in prediction and estimation if additional conditions regarding the noise vector $\boldsymbol{\epsilon}$ and the tuning parameter λ are satisfied. We now present the IC according to Zhao and Yu (2006) which is sufficient and *essentially* necessary (in a sense specified below) for the LASSO to select the true underlying model both in the classical fixed p setting and in the large p setting where p increases with rising n (Zhao and Yu, 2006). It can be proved that the IC implies the CC and that the former is thus always stronger than the latter (see Bühlmann and Van de Geer, 2009; Bühlmann and Van de Geer, 2011, Chapter 7.2).

In their study on the model selection consistency of the LASSO, Zhao and Yu (2006) investigated how well the sparse linear models given by the LASSO relate to the underlying true models. In doing so, they allowed for a high-dimensional data setting with $p \gg n$. The high-dimensional data were basically generated according to the linear model considered in Equation 1, with the difference, however, that the data and the coefficient vector $\boldsymbol{\beta}$ are indexed by n which accounts for the fact that they are allowed to change with rising sample size. Note that by using the notation presented in subsection 3.1, we notationally neglect the dependence of the data and $\boldsymbol{\beta}$ on n.

As in subsubsection 3.3.1, we assume in the following that the true coefficient vector $\boldsymbol{\beta}$ is sparse in the sense that some of the regression coefficients β_j are exactly zero. This assumption accounts for the fact that in practice we tend to include covariates in the model which de facto have no influence on the response. Solely focusing on the model selection capacity of the LASSO, Zhao and Yu (2006) used some dedicated definitions of sign consistency to separate the variable selection consistency aspect of the LASSO from its parameter estimation and prediction capacity aspect. Generally, they assumed that an estimate $\hat{\boldsymbol{\beta}}$ is equal in sign with the true model $\boldsymbol{\beta}$ which is written $\hat{\boldsymbol{\beta}} =_s \boldsymbol{\beta}$, if and only if $sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta})$. Note that $sign(.)$ maps a positive entry to 1, a negative entry to -1 and zero to 0. Hence, sign consistency is assumed to prevail if the estimator $\hat{\boldsymbol{\beta}}$ matches the zeros and signs of the true model $\boldsymbol{\beta}$. The above definition of sign consistency is stronger than the usual selection consistency which requires the zeros of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ to be matched, but not the signs. However, as indicated by Zhao and Yu (2006), such stronger definition of sign consistency is needed to prove the necessity of the IC and to avoid situations where a model is estimated with matching zeros, but reversed signs. Besides the aforementioned global definition of sign consistency, Zhao and Yu (2006, Definition 2 and 3) introduced two specific kinds of sign consistency which refer to the way how the amount of regularization for the LASSO has been determined. More precisely, a distinction is made between *Strong Sign Consistency (SSC)* given by Equation 25 and *General Sign Consistency (GSC)* defined by Equation 26. SSC implies the usage of an a priori established penalization parameter that ensures consistent model selection via the LASSO. On the other side, GSC implies that for a random realization of the tuning parameter there exists a correct amount of shrinkage λ that identifies the true model (Zhao and Yu, 2006). Zhao and Yu (2006) show that the two kinds of sign consistency are almost equivalent to one condition which is the IC. For this reason, we will not distinguish between both types of sign consistency when examining

the IC in the following.

Strong Sign Consistency: The LASSO is **strongly sign consistent** if there exists $\lambda_n = f(n)$, i.e., a function of n and independent of Y or $\mathbf{X}$ such that

$$\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}(\lambda_n) =_s \boldsymbol{\beta}) = 1. \tag{25}$$

General Sign Consistency: The LASSO is **general sign consistent** if

$$\lim_{n\to\infty} P(\exists\lambda \geq 0, \hat{\boldsymbol{\beta}}(\lambda) =_s \boldsymbol{\beta}) = 1. \tag{26}$$

Given the above definitions of sign consistency, we now present the SIC according to Zhao and Yu (2006). The SIC constitutes the central condition for the LASSO to be consistent in variable selection. However, we also need to have a glance at the *Weak Irrepresentable Condition (WIC)* to understand why the IC is said to be sufficient and *essentially* necessary for the LASSO to be consistent in variable selection. Assuming that $\mathbf{C}_{xx}(S_0, S_0)$ is invertible, the definitions of the SIC and the WIC are as follows:

Strong Irrepresentable Condition: There exists a positive constant vector $\boldsymbol{\eta}$, such that

$$|\boldsymbol{C}_{xx}(S_0^c, S_0)\boldsymbol{C}_{xx}(S_0, S_0)^{-1} sign(\boldsymbol{\beta}_{S_0})| \leq \mathbf{1} - \boldsymbol{\eta} \tag{27}$$

where $\mathbf{1}$ is a $(p-q)\times 1$ vector of 1's and the inequality holds element-wise.

Weak Irrepresentable Condition

$$|\boldsymbol{C}_{xx}(S_0^c, S_0)\boldsymbol{C}_{xx}(S_0, S_0)^{-1} sign(\boldsymbol{\beta}_{S_0})| < \mathbf{1}, \tag{28}$$

where, as above, the inequality holds element-wise.

As implied by the name, the WIC is slightly weaker than the SIC. This means that $\mathbf{C}_{xx}$ can converge in such way that the entries of $|\boldsymbol{C}_{xx}(S_0^c, S_0)\boldsymbol{C}_{xx}(S_0, S_0)^{-1} sign(\boldsymbol{\beta}_{S_0})|$ approach $\mathbf{1}$ from below so that the WIC holds but the SIC fails in the limit (Zhao and Yu, 2006). From

a technical point of view, the left hand sides of Equation 27 and Equation 28 are similar to a regularization constraint on the regression coefficients of the unimportant covariates $\mathbf{X}_{S_0^c}$ on the important ones $\mathbf{X}_{S_0}$. Note that in practice, the signs of the true $\boldsymbol{\beta}$ are not known so that Equation 27 and Equation 28 possibly should hold for all signs of $\boldsymbol{\beta}_{S_0}$. With respect to Equation 27, this means that we require Equation 29 to hold.

$$|(\mathbf{X}_{S_0}^T \mathbf{X}_{S_0})^{-1} \mathbf{X}_{S_0}^T \mathbf{X}_{S_0^c}| = |\boldsymbol{C}_{xx}(S_0, S_0)^{-1} \boldsymbol{C}_{xx}(S_0, S_0^c)| < \mathbf{1} - \boldsymbol{\eta} \tag{29}$$

We can derive from Equation 29 that for the SIC to hold for all possible signs of $\boldsymbol{\beta}_{S_0}$, it is necessary that the l_1-norms of the regression coefficients are smaller than 1. Generally, the SIC (and therefore also the WIC) mainly depends on the covariance of the predictor variables. It implies that the LASSO achieves consistent model selection according to the true model if and (almost) only if the covariates that are not part of the true model are *irrepresentable* by the covariates that are part of the true model. This requirement is met if $\boldsymbol{C}_{xx}(S_0, S_0^c)$ has small entries which implies that the covariance between pertinent and unimportant covariates is small. For Equation 29 to hold, the covariances between the pertinent covariates of the true model must also be small, since otherwise $\boldsymbol{C}_{xx}(S_0, S_0)$ may have small eigenvalues (nearly singular design) or even exhibit one or more zero eigenvalues (singular design) which also results in the violation of Equation 29. In their work, Zhao and Yu (2006, Proposition 1) quantitatively related the model selection capacity of LASSO and how well the SIC holds by deriving a lower bound for the probability of LASSO selecting the true model under the assumption that the SIC holds. In doing so, they also stressed the role of the tuning parameter λ which counterbalances the trade-off between sign consistency of the coefficient estimations for the influential covariates $\boldsymbol{\beta}_{S_0}$ and the shrinkage of the irrelevant coefficient estimations $\boldsymbol{\beta}_{S_0^c}$ towards zero. They concluded that it is easier for the LASSO to select the true model if the SIC in Equation 27 holds with a larger constant $\boldsymbol{\eta}$.

Basing on the SIC, Zhao and Yu (2006) established meaningful results with regard to the model selection consistency of the LASSO both for the classical small p and q case where q, p and $\boldsymbol{\beta}^n = \boldsymbol{\beta}$ are all fixed as $n \to \infty$ and for the large p and q case where $p = p_n$ and $q = q_n$ are allowed to grow with n.

Regularity conditions for model selection consistency (small p and q case)[8]

$$C^n \to C \text{ , as } n \to \infty, \tag{30}$$

where C is a positive definite matrix.

$$(1/n) \max_{1 \leq i \leq n} ((x_i^n)^T x_i^n) \to 0 \text{ , as } n \to \infty. \tag{31}$$

In particular, for the small p and q case they showed that under the assumption of the regularity conditions given by Equation 30 and Equation 31, the LASSO is *strongly sign consistent* if the SIC holds. Hence, if the SIC is satisfied, it holds for $\forall \lambda_n$ that satisfy $\lambda_n / n \to 0$ and $\lambda_n / n^{\frac{1+c}{2}} \to \infty$ with $0 \leq c < 1$ that

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1 - o(e^{-n^c}). \tag{32}$$

This implies that if the SIC holds in the classical setting where q, p and β^n are fixed as $n \to \infty$, the probability of the LASSO selecting the true model converges to 1 with an exponential rate while only the finite second moment of the noise terms is assumed (Zhao and Yu, 2006). Against the background of Equation 32 and the work of Knight and Fu (2000)[9], Zhao and Yu (2006) infer that the SIC enables consistent model selection and parameter estimation at the same time. The WIC is also necessary even for the weaker GSC (see Zhao and Yu, 2006, Theorem 2). Therefore, it holds that the SIC implies SSC which again implies GSC which in turn implies the WIC. This chain of implications is crucial since it constitutes the motivation for stating that the IC is *essentially* necessary and sufficient for both SSC and GSC if one ignores the minor technical difference between the WIC and the SIC. Recall that as necessary condition the WIC requires the relation ≤ 1 to hold, while the SIC as sufficient condition requires the relation $\leq 1 - \eta$ with $0 < \eta < 1$ to hold. Nevertheless, we will no longer distinguish between the SIC and the WIC in the following, but globally refer to the IC as being the *essentially* necessary and sufficient condition for consistent variable selection with the LASSO.

8 Note that although the convergences in Equation 30 and Equation 31 are deterministic, the following results regarding the consistency of the LASSO for variable selection also hold quite generally for random designs (Zhao and Yu 2006).

9 Knight and Fu (2000) showed that for $\lambda_n = o(n)$ the LASSO is consistent in parameter estimation and exhibits asymptotic normality.

Since the LASSO is known to be a suitable regression technique for high-dimensional data, the implications of the SIC for the LASSO's variable selection performance in high-dimensional regression settings are of importance. In the large p and q setting, one has $p = p_n$ and $q = q_n$ which implies that the number of included covariates and the sparsity index are allowed to grow with the sample size. This means that the dimension of the Gram matrix $\mathbf{C}_{xx} = \mathbf{C}_{xx}^n$ and also the dimension of the coefficient vector $\boldsymbol{\beta} = \boldsymbol{\beta}_n$ are also growing with n. Against this background, the regularity conditions given by Equation 30 and Equation 31 become inapplicable, since $\mathbf{C}_{xx}^n$ does no longer converge to a fixed, positive definite matrix. Furthermore, $\boldsymbol{\beta}_n$ may alter as n grows. Consequently, Zhao and Yu (2006) proposed the following regularity conditions for covariate selection consistency of the LASSO in the *large p and q case*:

Regularity conditions for model selection consistency (large p and q case)
Assuming that there exists $0 \leq c_1 < c_2 \leq 1$ and $M_1, M_2, M_3, M_4 > 0$, it holds that:

$$1/n \ (\mathbf{X}_i^n)^T \mathbf{X}_i^n \leq M_1 for \ \forall i, \tag{33}$$

$$\boldsymbol{\alpha}^T \mathbf{C}_{xx}^n(S_0, S_0)\boldsymbol{\alpha} \geq M_2, for \forall \ \|\boldsymbol{\alpha}\|_2^2 = 1, \tag{34}$$

$$q_n = O(n^{c_1}), \tag{35}$$

$$n^{1-c_2/2} \min_{i=1,\ldots,q} |\beta_i^n| \geq M_3. \tag{36}$$

The restriction imposed by Equation 33 is generally satisfied if the covariates are normalized before the LASSO is applied (Zhao and Yu, 2006). Equation 34 imposes a lower bound on the eigenvalues of the design which comprises the important covariates $\mathbf{C}_{xx}^n(S_0, S_0)$. In doing so, Equation 34 ensures that $\mathbf{C}_{xx}^n(S_0, S_0)$ is positive definite and that hence its inverse exists. This is crucial since $\mathbf{C}_{xx}^n(S_0, S_0)$ needs to be invertible for both the SIC and the WIC to hold. Equation 35 and Equation 36 act as main conditions for the following results and stand in close relation to each other. Bühlmann and Van de Geer (2011, Ch. 7.4) call Equation 36 a *beta-min condition*, since it controls the size of the smallest entry of $\boldsymbol{\beta}_{S_0}^n$

and thus reflects the fact that important covariates must have a sufficiently large effect on the response to possibly be detected by the LASSO. More in particular, Equation 36 implies that there is a gap of size n^{c_2} between the smallest entry or decay rate of $\boldsymbol{\beta}^n_{S_0}$ and $n^{-\frac{1}{2}}$ to prevent the LASSO estimation to be dominated by the noise terms which aggregate at a rate of $n^{-\frac{1}{2}}$ (Zhao and Yu, 2006). On the other hand, Equation 35 defines a restriction for the sparsity index q_n which indicates the size of the true model. More precisely, its square root $\sqrt{q_n}$ is required to grow at a rate slower than the rate gap to keep the estimation bias of the LASSO solution from dominating the coefficient estimates (Zhao and Yu, 2006). Under the above regularity conditions Equation 33 to Equation 36 and assuming that the SIC holds, Zhao and Yu (2006, Theorem 3 and 4) establish variable selection consistency results for the LASSO in the large p and q case. For general noise with ϵ^n_i $(i = 1, ..., n)$ being i.i.d. random variables with finite $2k$'th moment $\mathbb{E}(\epsilon^n_i)^{2k} < \infty$ for an integer $k > 0$, they show that the SIC implies that the LASSO has SSC for $p_n = o(n^{(c_2-c_1)k})$. In particular, for $\forall\lambda_n$ that satisfies $\frac{\lambda_n}{\sqrt{n}} = o(n^{\frac{c_2-c_1}{2}})$ and $\frac{1}{p_n}(\frac{\lambda_n}{\sqrt{n}})^{2k} \to \infty$, it holds that

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - O(\frac{p_n n^k}{\lambda_n^{2k}}) \to 1 \text{ as } n \to \infty. \tag{37}$$

For the specific case of Gaussian noise (ϵ^n_i are i.i.d.Gaussian random variables), Zhao and Yu (2006, Theorem 4) derive that the SIC implies that the LASSO has SSC if there exists $0 \leq c_3 < c_2 - c_1$ for which $p_n = O(e^{n^{c_3}})$. In particular, for $\lambda_n \propto n^{\frac{1+c_4}{2}}$ with $c_3 < c_4 < c_2 - c_1$, it holds that

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - o(e^{n^{-c_3}}) \to 1 \text{ as } n \to \infty. \tag{38}$$

Generally, Equation 37 and Equation 38 relate to exactly the same context and only differ with regard to the assumptions made for the noise terms ϵ_i. For general noise, Equation 37 implies that the LASSO does consistent variable selection if the SIC holds and if additionally $\mathbb{E}(\epsilon^n_i)^{2k} < \infty$ for an integer $k > 0$ which means that the noise terms have some finite moments. If, e.g., $k = 1$ (only the second moment is assumed), then p is allowed to grow slower than $n^{c_2-c_1}$. If all moments of the noise exist ($k = \infty$), then the number of covariates p is allowed to increase at any polynomial rate while the probability of LASSO selecting the true model converges to 1 at a even faster rate (with rate $O(\frac{p_n n^k}{\lambda_n^{2k}})$). For models where ϵ_i follows an i.i.d. Gaussian distribution, the SIC implies that the probability of the LASSO selecting the correct model converges to 1 quite fast (with rate $o(e^{n^{-c_3}})$), even though p might increase up to exponentially fast compared to n ($p_n = O(e^{n^{c_3}})$) (see Equation 38). Note, however, that Zhao and Yu (2006) emphasize that a comparably high convergence

rate is not achievable for all noise distributions, since the tail probability of noise terms does not vanish quick enough to allow p to grow at higher degree polynomial rates if higher moments of the noise distribution do not exist.

In summary, the results stated in this subsubsection indicate that the IC is sufficient and essentially necessary for consistent variable selection with the LASSO both in the classical setting with fixed number of covariates p and in settings where p grows with the sample size n (Zhao and Yu, 2006). The IC basically postulates that the covariance matrix of the design may not exhibit too strong degrees of linear dependence within smaller sub-matrices, in particular within the covariance sub-matrix of the irrelevant and the truly important covariates $\mathbf{C}_{xx}(S_0^c, S_0)$. If the IC is violated, the LASSO is unable to consistently recover the underlying true model. The reason therefore lies in the fact that in order to produce sparse models the LASSO shrinks the coefficient estimates belonging to important covariates too heavily. Thus, if the IC fails, the irrelevant covariates are correlated with the important covariates enough to be selected by the LASSO to outweigh the over-shrinkage of the non-zero coefficient estimates (Zhao and Yu, 2006). From a technical point of view, the IC also reflects that for singular covariance matrices, edges are not uniquely defined by the distribution which makes nearly singular covariance matrices inappropriate for consistent variable selection. For statistical practice, this implies that one has to be aware of the fact that if unimportant variables are strongly correlated with covariates that are part of the true model, the LASSO is unlikely to select the true model. However, the problem is that in practice the active set S_0 and thus also the sparsity index q are unknown. In other words, we typically do not know which and how many covariates are part of the true model, since not knowing the relevant covariates might mostly be the precise reason for choosing the LASSO as regression method. This implies that for a practical guarantee, the IC should hold for all possible S_0. This finding is in line with Zhang and Huang (2008) who comment on the IC that without knowing S_0 it is not really obvious how to verify the SIC other than using simple bounds on the covariate correlation $\mathbf{x}_j^T \mathbf{x}_k$ for $j \neq k$ as in Zhao and Yu (2006, see Corollary 1).[10] Furthermore, for sufficiency of the

10 Zhao and Yu (2006) provide some sufficient conditions that ascertain that the SIC is satisfied in practice. It is, however, noticeable that for these conditions to hold, one must again make several assumptions. As an example, we refer to Corollary 1 where they consider the case of constant positive correlation r_n. They show that there exists some $c > 0$ such that the SIC holds for $0 < r_n \leq \frac{1}{1+cq}$. In practice the question would be how c should be determined. It is clear that if we use a reasonably large value for c, the upper bound for r_n becomes very low and we expect the SIC to hold. However, this simply corresponds to the fact that our design does not contain correlations between the covariates which is rather an unlikely case

SIC we explicitly assumed that the minimal non-zero coefficients of the true regression model are sufficiently large (Equation 36). Apart from the problem of how to practically define "sufficiently large", this assumption is in general unlikely to be satisfied (see also Bühlmann and Van de Geer, 2011). We can thus infer that in spite of the fact that the IC constitutes a clear condition which is also easy to interpret, it requires strong assumptions that are not verifiable in statistical practice. This is line with Bühlmann and Van de Geer (2011, p.184) who state that the LASSO for variable selection only works in a rather narrow range of problems excluding many cases where there are strong (empirical) correlations between the covariates. More in particular, the LASSO tends to select too many covariates. On the other hand, a certain amount of false negative selections cannot be avoided either if the absolute value of some β_j is below the LASSO's detection limit (Bühlmann and Van de Geer, 2011). However, it should be pointed out that although the LASSO might not be able to infer the correct set of covariates with non-zero coefficients from the data if the IC is violated, in practice it can nevertheless be used to find at least some covariates with substantial impact (in terms of their corresponding $|\beta_j|$) on the response. More specifically, it can be shown that it holds that

$$\mathbb{P}[\hat{S}(\lambda) \supset S_0^{relevant(C)}] \to 1 \text{ for } n \to \infty, \tag{39}$$

if the substantial covariates satisfy $S_0^{relevant(C)} = \{j : |\beta_j| \geq C, j = 1, ..., p\}$ for any fixed threshold value $0 < C < \infty$. Note that $\hat{S}(\lambda)$ is the sub-model chosen by the LASSO for a given λ and $\hat{\mathcal{S}} = \{\hat{S}(\lambda); all\ \lambda\}$ is the set containing all possible LASSO sub-models. In cases where $S_0^{relevant(C)} = S_0$ (which means that all non-zero coefficients have a minimal absolute value of C), one has $\hat{S}(\lambda) \supset S_0$ with high probability. The property of the LASSO which is reflected in Equation 39, is referred to as its *variable screening property*. The latter implies that the model estimated by the LASSO includes the substantial covariates with high probability (Bühlmann and Van de Geer, 2011). Finally, as stated by Bühlmann and Van de Geer (2011), it follows from the analysis of the LARS algorithm (Efron et al., 2004) that every model estimated by the LASSO has a cardinality smaller than or equal to $\min(n, p)$. Hence, for high-dimensional datasets with $p \gg n$, it follows that models estimated by the LASSO have the cardinality $\min(n, p) = n$. Given that n is a reasonably small number compared to p if $p \gg n$, the LASSO achieves per se a considerable dimensionality reduction with respect to the full amount of covariates contained in the dataset.[11]

in high-dimensional data settings.

11 For more theoretical details on variable screening with the LASSO, we refer to Bühlmann and Van de

3.4 A short note on tuning parameter selection for the LASSO

Up to this point, we considered the regularization parameter λ as kind of "black box" by simply assuming an a priori given, suitable amount of regularization. However, as mentioned before, the amount of regularization λ applied is crucial for the LASSO to be consistent in estimation and prediction as well as in variable selection. In this section, we only briefly sketch the issue of tuning parameter selection for the LASSO which - of course - constitutes a very interesting and important research question by its own.

Before we apply the LASSO in practice, we first need to determine an appropriate amount of regularization λ. To choose λ, we usually revert to some cross-validation scheme which aims at optimizing the prediction error. Hastie et al. (2009, p. 241) characterize cross-validation as the "probably [...] simplest and most widely used method for estimating prediction error". In the context of penalized linear regression with the LASSO, a common approach for estimating λ is to use k-fold cross-validation. Note that k-fold cross-validation randomly splits the available data into k equally sized, disjoint sets and then uses $(k-1)$ folds to fit or train the model and the kth fold to test it. The estimate of the prediction error corresponds to the mean of the totality of the k "left-out-fold" cases. Common choices for k are, e.g., $k = 5$ and $k = 10$ (Hastie et al., 2009). Using $k = n$ results in the so-called *leave-one-out cross-validation* scheme which is, however, computationally very expensive (Chand, 2012). An alternative consists in applying *generalized cross-validation* which uses an approximation to leave-one-out cross-validation and therefore requires less computational power. For a detailed discussion of the different cross-validation methods we refer to Hastie et al. (2009).

A problem that arises from the use of a cross-validated tuning parameter λ_{cv} for the LASSO is that prediction optimality does often not align with the goal of variable selection (Bühlmann and Van de Geer, 2011). More precisely, the use of λ_{cv} leads the LASSO to select too many variables, which results in an elevated false positive fraction (FPF). This implies that if the LASSO is applied with the aim of variable selection, it is advisable to employ a larger amount of regularization than the one suitable for good prediction (Bühlmann and Van de Geer, 2011).

Nevertheless, Bühlmann and Van de Geer (2011) point out that if we intend to apply the LASSO with the purpose of variable screening (see subsubsection 3.3.2), using $\hat{\lambda}_{cv}$

Geer (2011, Chapter 2 and 7).

is more appropriate. They state that there is an empirical evidence that $\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0$ or $\hat{S}(\hat{\lambda}_{CV}) \supseteq S^{relevant(C_n)}$ and that this empirical evidence is further supported by theory (see, e.g., Meinshausen and Bühlmann, 2006). Given the aforementioned dimensionality reduction inherent to the LASSO (see subsubsection 3.3.2), one can alternatively include all $\min(n, p)$ variables by using a value λ sufficiently close to zero which means that no regularization parameter needs to be chosen (Bühlmann and Van de Geer, 2011).

In summary, it can be stated that for the LASSO to be consistent in variable selection, the user has to apply a larger amount of regularization $\lambda = \lambda_n$ than for good or even optimal prediction (Bühlmann and Van de Geer, 2011). From a theoretical perspective, this means that λ should be chosen of a larger order than $\sqrt{\log(p)/n}$ (see subsubsection 3.3.1). This theory is also supported by empirical studies on the variable selection performance of LASSO. More in particular, Chand (2012) performed an empirical study and pointed out that cross-validation is not a reliable method if variable selection is the primary objective when using the LASSO. As an alternative, he proposed to use a BIC-type tuning parameter selector which he showed to facilitate consistent variable selection under certain conditions.

4 Multiple linear regression with additive measurement error in the covariates

An assumption that is common to all of the aforementioned studies on the LASSO is that the design matrix $\mathbf{X}$ passed to the LASSO for performing linear regression contains true covariate measurements that do not suffer from any type of measurement error. In practice, where data corrupted by measurement errors including errors-in-variable data are rather the norm than the exception, this assumption does, however, not meet the truth. Error-prone covariate measurements represent a challenge for the majority of conventional statistical models. Correspondingly, there is a well-developed body of literature discussing various theoretical aspects of different types of covariate measurement errors and their implications for, e.g., linear regression methods. Note that the field of measurement error theory is a wide and interesting, but also complex field of research. The topic-related complexity is further enhanced by the fact that the prevalent literature makes use of heterogeneous and partly very complex notational styles (see, e.g., Schneeweiß and Mittag, 1986; Fuller, 1987).

Before considering the LASSO under matrix uncertainty from an analytical point of view in section 5, we will now briefly outline the basics of measurement error theory for linear models and present potential correction methods. We consider additive covariate measurement error only. We will also not get into detail with regard to the theory of additive measurement error models which constitutes itself a complex field of research. There is a vast amount of insightful literature that deals with it and we refer the interested reader to the latter and the references therein (see, e.g., Fuller, 1987; Cheng and Van Ness, 1999; Carroll et al., 2006). Our introduction of additive measurement error models given in this section is rather tightly tailored towards the specific measurement error context examined in our simulation study of the LASSO in section 6. The notation used and also the majority of the results presented in the following are based on Buonaccorsi (2010, Chapter 5 and 6).

4.1 Measurement error model setup

To introduce the measurement error model framework applied in section 5 and section 6, we consider the multiple linear regression model defined in Equation 40. Note that $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ represents the vector of true covariate measurements for the i^{th} observation without the constant for the intercept β_0 and $\mathbf{x}_{i*} = (1, x_{i1}, ..., x_{ip})^T$ is the same vector of truly measured covariates, but also includes the constant for the intercept. Likewise, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_p) = (\beta_0, \boldsymbol{\beta}_1^T)$ is the coefficient vector including the intercept, while $\boldsymbol{\beta}_1^T = (\beta_1, ..., \beta_p)$ represents the coefficient vector without β_0. For a model without intercept β_0 (see Equation 1 in section 2), it holds that $\mathbf{x}_i = \mathbf{x}_{i*}$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_1$.

$$Y_i|\mathbf{x}_i = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_{ij} + \epsilon_i = \beta_0 + \boldsymbol{\beta}_1^{\mathbf{T}}\mathbf{x}_i + \epsilon_i = \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x}_{i*} + \epsilon_i \tag{40}$$

The corresponding model for all available observations i ($i = 1, ..., n$) in matrix form is given in Equation 41. Note that the only difference between Equation 41 and Equation 1 in section 2 is that the linear model in Equation 41 includes an intercept. Hence, Equation 41 involves the design matrix $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$, where row $\mathbf{x}_i \in \mathbb{R}^{(p+1)}$ contains the constant for the intercept and all covariate measurements for the i^{th} observation and the unknown coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^{(p+1)}$ to be estimated. In line with Equation 1, $\mathbf{Y} \in \mathbb{R}^n$ is the univariate response vector and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a vector of model errors whose components ϵ_i

($i = 1, ..., n$) are assumed to be i.i.d. normally distributed with mean zero and constant variance σ^2, i.e., $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. In line with Buonaccorsi (2010), we will refer to $\boldsymbol{\epsilon}$ as the *error in the equation* or *model error* to explicitly distinguish it from the measurement error in the response which is denoted by $\mathbf{q}$.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{41}$$

where

$$\mathbf{X} = \begin{bmatrix} x_{1*}^{\mathbf{T}} \\ x_{2*}^{\mathbf{T}} \\ \vdots \\ x_{n*}^{\mathbf{T}} \end{bmatrix} \tag{42}$$

The sample means of $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ and $\mathbf{Y} \in \mathbb{R}^n$ can be obtained by means of Equation 43 and Equation 44, respectively. Note that $\bar{\mathbf{X}} \in \mathbb{R}^{(p+1)}$ is a vector, while $\bar{Y} \in \mathbb{R}$ is a scalar. The corresponding sample covariances $\mathbf{C}_{xx} \in \mathbb{R}^{(p\times p)}$ and $\mathbf{C}_{xy} \in \mathbb{R}^p$ are defined in Equation 45 and Equation 46.

$$\bar{\mathbf{X}} = \frac{\sum_{i=1}^{n} \mathbf{X}_i}{n} \tag{43}$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{44}$$

$$\mathbf{C}_{xx} = \frac{\sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})^{\mathbf{T}}}{(n-1)} \tag{45}$$

$$\mathbf{C}_{xy} = \frac{\sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(Y_i - \bar{Y})}{n-1} \tag{46}$$

As mentioned above, in Equation 40 and Equation 41 we typically assume that the design matrix $\mathbf{X}$ contains error-free covariate values. Given such true design matrix and further assuming that the latter has full rank, we can use Equation 2 to infer $\boldsymbol{\beta}$ from the given data

$(y_i, \mathbf{x}_i)$ $(i = 1, ..., n)$. In doing so, we obtain the unbiased LS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ which is also BLUE. Instead of Equation 2, one can also use Equation 43 to Equation 46 to estimate $\boldsymbol{\beta}$ by means of the following relations:

$$\hat{\boldsymbol{\beta}}_{1,OLS} = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy} \tag{47}$$

$$\hat{\beta}_{0,OLS} = \bar{Y} - \hat{\boldsymbol{\beta}}_1^{\mathbf{T}}\bar{\mathbf{X}} \tag{48}$$

Before proceeding with prevalent additive measurement error models and their implications for the OLS estimator, it should be noted with regard to the covariate values in $\mathbf{X}$ that a basic distinction can be made between functional and structural settings. In fully structural settings, one assumes that all covariates are random and that the random covariate vectors $\mathbf{X}_i$ are i.i.d. normally distributed with mean μ_X and variance σ_X^2. By contrast, in functional settings the covariates are treated as fixed or realized values ($\mathbf{X}_i = \mathbf{x}_i$). The main reason for this distinction is that with random $\mathbf{X}_i$ one might also be interested in the correlation between $\mathbf{X}$ and $\mathbf{Y}$ (Buonaccorsi, 2010). However, for simplicity, we do not particularly differentiate between structural and functional settings in the following. Instead we use broad definitions for $\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_{xx}$ that potentially cover any combination of random and fixed covariates. According to Buonaccorsi (2010), we define

$$\boldsymbol{\mu}_x = \sum_{i=1}^{n} \mathbb{E}(\mathbf{X}_i)/n \tag{49}$$

and

$$\boldsymbol{\Sigma}_{xx} = \mathbb{E}(\mathbf{C}_{\mathbf{xx}}). \tag{50}$$

Thus, one has $\boldsymbol{\mu}_x = \mathbb{E}(\mathbf{X}_i)$ and $\boldsymbol{\Sigma}_{xx} = Cov(\mathbf{X}_i)$ in fully structural cases where the random covariate vectors $\mathbf{X}_i$ $(i = 1, ..., n)$ are i.i.d. normally distributed and $\boldsymbol{\mu}_x = \sum_{i=1}^{n} \mathbf{x}_i/n$ and $\boldsymbol{\Sigma}_{xx} = \mathbf{C}_{xx} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T/(n-1)$ in functional cases with fixed covariate vectors $\mathbf{x}_i$ $(i = 1, ..., n)$. Given a combination of fixed and random covariates, the expected value and the covariance of $\mathbf{X}_i$ are conditional on any fixed components and thus can change with i (see Buonaccorsi, 2010, chapter 5.8.2). Equation 49 and Equation 50 allow for the definition of approximate biases and correction methods in a general way, i.e., without

differing between structural and functional seetings.

We now introduce two prevalent measurement error models, more specifically the *additive Berkson model* and the *classical additive measurement error model*. The latter will be applied in our simulation study in section 6 where we examine the performance of the LASSO in the presence of additive measurement error in the design matrix $\mathbf{X}$. Consider the multiple linear model presented in Equation 41. Instead of the perfect covariate vectors $\mathbf{X}_i$ ($i = 1, ...n$) given in Equation 41, in practice we commonly observe inaccurate covariate values that suffer from additive measurement error. Depending on the particular context of data generation, one can distinguish between two basic measurement error models.

The *additive Berkson model* for observation i is defined as

$$\mathbf{X}_i = \mathbf{w}_i + \mathbf{e}_i, \tag{51}$$

where $\mathbb{E}(\mathbf{e}_i|\mathbf{w}_i) = \mathbf{0}$ and $\mathbb{C}ov(\mathbf{e}_i|\mathbf{w}_i) = \mathbf{\Sigma}_{\mathbf{e}_i}$, if we allow the measurement error variances to change with i. If there is also measurement error in the univariate response y_i, one has $D_i = y_i + q_i$ with $\mathbb{E}(q_i|y_i) = 0$ and $\mathbb{C}ov(q_i|y_i) = \sigma^2_{qi}$. As mentioned before, q_i is the error in the response and should not be confused with the noise term ϵ_i. We further assume that the error in the response q_i is independent of the error in the covariates $\mathbf{e}_i$. Without error in the response q_i, substituting $\mathbf{X}_i$ in Equation 40 by Equation 51 leads to

$$Y_i = \beta_0 + \boldsymbol{\beta}_1 \mathbf{w}_i + \eta_i, \tag{52}$$

where

$$\eta_i = \boldsymbol{\beta}_1^T \mathbf{e}_i + \epsilon_i, \tag{53}$$

with $\mathbb{E}(\eta_i) = 0$ and $\mathbb{C}ov(\eta_i) = \boldsymbol{\beta}_1^2 \mathbf{\Sigma}_{ei} + \sigma^2$.

It is characteristic for the above presented additive Berkson model that the inaccurate covariate value vector $\mathbf{w}_i$ is fixed, while the vector of true covariate values $\mathbf{X}_i$ is random. Hence, the Berkson model specifies an assumption about the distribution of the random, true covariate values $\mathbf{X}_i$ given the observed - or targeted - values $\mathbf{w}_i$. The Berkson model is suitable for experimental situations where $\mathbf{w}_i$ is a vector of fixed target doses or where $\mathbf{w}_i$

corresponds to fixed factor levels in general. In such experiments, the observed covariate vector $\mathbf{w}_i$ is controlled in the sense that it is pre-specified by the experimenter. However, in practice the true dose $\mathbf{X}_i$ will show more variability than the estimated or targeted dose $\mathbf{w}_i$. For detailed examples of experimental settings which involve Berkson error and applications of the Berkson error model we refer to Heid et al. (2004) and Fuller (1987, Chapter 1.6).
Generally, the naive OLS estimator given in Equation 2 can be used for parameter estimation in linear regression models with data subject to additive Berkson error. The reason therefore is that the covariate vector $\mathbf{w}_i$ is controlled so that the "error part" η_i of the linear regression is uncorrelated with $\mathbf{w}_i$, since any random variable is uncorrelated with a constant value. Thus, using the OLS estimator for inference of β_j $(j = 0, ..., p)$ in Equation 41 results in unbiased parameter estimates. Furthermore, if there is Berkson error in the covariates but no error in the response, the predictions of $\mathbf{Y}$ from $\mathbf{w}$ and corresponding prediction intervals are also correct (Buonaccorsi, 2010). It follows that one can ignore the error in the covariates by simply using the standard OLS estimator for parameter estimation in a linear regression model if $\mathbf{w}_i$ is controlled. Nevertheless, it is important that this conclusion only holds if the Berkson error is additive and if the underlying model is linear (Fuller, 1987; Buonaccorsi, 2010).

In the additive Berkson model, the true value $\mathbf{X}_i$ and the random measurement error $\mathbf{e}_i$ are correlated, while the observed covariate values $\mathbf{w}_i$ are controlled and thus independent of the random measurement error $\mathbf{e}_i$. By contrast, the *classical additive measurement error model* assumes that given y_i and $\mathbf{x}_i$ for observation i, it holds that

$$D_i = y_i + q_i \tag{54}$$

$$\mathbf{W}_i = \mathbf{x}_i + \mathbf{u}_i, \tag{55}$$

with

$$\mathbb{E}(q_i | y_i, \mathbf{x}_i) = 0, \tag{56}$$

$$\mathbb{E}(\mathbf{u}_i | y_i, \mathbf{x}_i) = \mathbf{0}, \tag{57}$$

and

$$\mathbb{V}ar(q_i|y_i, \mathbf{x}_i) = \sigma^2_{qi}, \tag{58}$$

$$\mathbb{C}ov(\mathbf{u}_i|y_i, \mathbf{x}_i) = \mathbf{\Sigma}_{ui}, \tag{59}$$

$$\mathbb{C}ov(\mathbf{u}_i, q_i|y_i, \mathbf{x}_i) = \mathbf{\Sigma}_{uqi}. \tag{60}$$

In this context, $\mathbf{u}_i \in \mathbb{R}^{(p+1)\times 1}$ is the vector of measurement errors that is added to the true covariate values for observation i. Note that the first entry of $\mathbf{u}_i$ corresponds to a zero which is added to the intercept constant in $\mathbf{x}_i$. D_i and $\mathbf{W}_i$ contain the inaccurate covariate values that the practitioner observes. By assuming $\mathbb{E}(q_i|y_i, \mathbf{x}_i) = 0$ and $\mathbb{E}(\mathbf{u}_i|y_i, \mathbf{x}_i) = \mathbf{0}$, we generally imply that undesired errors arising due to random variations in the measurement process add up to zero if a measuring system is well-calibrated (Fuller, 1987). In situations without error in the response, it holds that $\sigma^2_{qi} = 0$ and $\mathbf{\Sigma}_{uqi} = 0$ which then results in a simpler model. Likewise, if parts of the covariate vector $\mathbf{x}_i$ are measured without error, all components of $\mathbf{u}_i$, $\mathbf{\Sigma}_{ui}$ and $\mathbf{\Sigma}_{uqi}$ that belong to those perfect covariate measurements equal zero. Note that if one allows the above measurement error variances and covariances to change with i, their respective average values across all n observations can be calculated by means of the following equations:

$$\mathbf{\Sigma}_{uu} = \sum_{i=1}^{n} \mathbf{\Sigma}_{ui}/n, \tag{61}$$

$$\mathbf{\Sigma}_{uq} = \sum_{i=1}^{n} \mathbf{\Sigma}_{uqi}/n, \tag{62}$$

$$\sigma^2_q = \sum_{i=1}^{n} \sigma^2_{qi}/n. \tag{63}$$

Compared to the Berkson model, the classical measurement error model describes situations where instead of the true covariate occurrences $\mathbf{x}_i$, we observe a random vector $\mathbf{W}_i$ containing the true covariate values $\mathbf{x}_i$ plus a certain amount of random measurement error $\mathbf{u}_i$. Thus, with classical measurement error we model the distribution of the observed

values $\mathbf{W}_i$ given the true covariate values $\mathbf{x}_i$. Note that the random measurement errors $\mathbf{u}_i$ and the true covariate values $\mathbf{x}_i$ are assumed to be independent, while it follows from Equation 55 that $\mathbf{W}_i$ and $\mathbf{u}_i$ are correlated. As detailed above, in the case of Berkson error the naive OLS estimator provides unbiased coefficient estimates. However, with $\mathbf{W}_i$ and $\mathbf{u}_i$ being correlated random variables, "naively" applying the OLS estimator to data with classical measurement error results in biased coefficient estimates. The naive OLS estimator for error-prone design is defined as

$$\hat{\boldsymbol{\beta}}_{naive} = (\mathbf{W}^{\mathbf{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathbf{T}}\mathbf{D}, \tag{64}$$

where by analogy to Equation 2, one has $\mathbf{W} \in \mathbb{R}^{n\times(p+1)}$ and $\mathbf{D} \in \mathbb{R}^{n\times 1}$. In line with Equation 47 and Equation 48, $\hat{\boldsymbol{\beta}}_{naive}$ can alternatively be obtained by means of Equation 65 and Equation 66, where one has $\mathbf{C}_{ww} \in \mathbb{R}^{p\times p}$ and $\mathbf{C}_{wd} \in \mathbb{R}^{p\times 1}$. Note that $\mathbf{C}_{ww}$ and $\mathbf{C}_{wd}$ can be calculated by applying Equation 45 and Equation 46, but replacing $\mathbf{X}$ by $\mathbf{W}$ and $\mathbf{Y}$ by $\mathbf{D}$.

$$\hat{\boldsymbol{\beta}}_{1naive} = \mathbf{C}_{ww}^{-1}\mathbf{C}_{wd} \tag{65}$$

$$\hat{\beta}_{0naive} = \bar{D} - \hat{\boldsymbol{\beta}}_{1naive}^{\mathbf{T}}\bar{\mathbf{W}}, \tag{66}$$

$$\mathbb{E}(\mathbf{C}_{ww}) = \mathbb{E}(\mathbf{C}_{xx}) + \boldsymbol{\Sigma}_{uu} \approx \boldsymbol{\Sigma}_{xx} + \boldsymbol{\Sigma}_{uu} \tag{67}$$

$$\mathbb{E}(\mathbf{C}_{wd}) = \mathbb{E}(\mathbf{C}_{xy}) + \boldsymbol{\Sigma}_{uq} \approx \boldsymbol{\Sigma}_{xy} + \boldsymbol{\Sigma}_{uq} \tag{68}$$

Consider the expectations for $\mathbf{C}_{ww}$ and $\mathbf{C}_{wd}$ given in Equation 67 and Equation 68.[12] Given the latter, it follows from Equation 65 and Equation 66 that the naive OLS estimator clearly produces biased estimates for $\boldsymbol{\beta}$ if it is applied to data with classical measurement error.

12 Note that these expectations are exact under the normal structural model with normal measurement error and constant measurement error covariance matrix. For all other cases, they are only approximate (Buenaccorsi, 2010, p. 109).

More specifically, with measurement error in the covariates and in the response it follows that

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{1naive}) = \mathbb{E}(\mathbf{C}_{ww}^{-1}\mathbf{C}_{wd}) \approx \boldsymbol{\gamma}_1 = (\boldsymbol{\Sigma}_{xx} + \boldsymbol{\Sigma}_{uu})^{-1}\boldsymbol{\Sigma}_{xx}\,\boldsymbol{\beta}_1 + (\boldsymbol{\Sigma}_{xx} + \boldsymbol{\Sigma}_{uu})^{-1}\boldsymbol{\Sigma}_{uq} \tag{69}$$

and

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{0naive}) \approx \gamma_0 = \beta_0 + (\boldsymbol{\beta}_1 - \boldsymbol{\gamma}_1)^{\mathbf{T}}\boldsymbol{\mu}_x. \tag{70}$$

In our simulation study presented in section 6, we focus on the LASSO in the presence of classical additive measurement error in the covariates. In doing so, we assume error-free responses y_i $(i = 1, ..., n)$, i.e., we imply that $q_i = 0$. As a consequence, one has $\boldsymbol{\Sigma}_{uqi} = 0$ which means that there is no correlation between the error in the response and any errors in the covariates. Given the above equations and assuming that $q_i = 0$ leads to the following simplified equation which shows the bias for the OLS estimator in the presence of classical measurement error in the covariates:

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{1naive}) \approx (\boldsymbol{\Sigma}_{xx} + \boldsymbol{\Sigma}_{uu})^{-1}\boldsymbol{\Sigma}_{xx}\,\boldsymbol{\beta}_1 = \boldsymbol{\kappa}\boldsymbol{\beta}_1, \tag{71}$$

where

$$\boldsymbol{\kappa} = (\boldsymbol{\Sigma}_{xx} + \boldsymbol{\Sigma}_{uu})^{-1}\boldsymbol{\Sigma}_{xx}. \tag{72}$$

$\boldsymbol{\kappa}$ is referred to as *reliability ratio* (Fuller, 1987; Carroll et al., 2006) or *reliability matrix* (Buonaccorsi, 2010). As can be seen from Equation 71 and Equation 72, the fact that $\boldsymbol{\beta}_1$ is multiplied by the reliability matrix leads to a downward bias of the resulting naive OLS estimators for $\boldsymbol{\beta}_1$. In the prevalent literature, this effect is referred to as *attenuation bias* (Fuller, 1987; Carroll et al., 2006; Buonaccorsi, 2010). For the models presented in this section and also in section 6, we implicitly assume that each of the p considered covariates is measured with error. Note, however, that measurement error in only one of the covariates often leads to biased estimates for all covariate coefficients including such coefficients that are actually measured without error (see Buonaccorsi, 2010, Illustration 1-3 in Chapter 5.3).

4.2 Methods to correct for additive measurement error in the covariates

We illustrated in subsection 4.1 that the naive OLS estimator produces biased estimates for the coefficients $\beta_j (j = 0, ..., p)$ if the design matrix contains covariate values that

are measured with classical additive measurement error. Hence, there is a need for methods that allow the user to correct for measurement error in cases where the latter is inherent to the data of interest. In fact, there are numerous settings and assumptions under which corresponding estimators can be developed. This has lead to quite a few alternative estimators for measurement error data (see, e.g., Fuller, 1987; Cheng and Van Ness, 1999; Buonaccorsi, 2010). However, in the following we confine ourselves to briefly sketching some basic approaches that facilitate the estimation of reliable (in the sense of unbiased) coefficients β_j, albeit the presence of error-prone covariate values as described in Equation 55. We present the so-called *method of moments estimator* which uses knowledge about the true measurement error variance to correct for potential error in the covariates. For the more likely case that the practitioner has no knowledge of the true measurement error variance, we show how to use replicate data to estimate $\boldsymbol{\Sigma}_{uu}$ before applying the method of moments (see subsubsection 4.2.2).

4.2.1 Error correction using the method of moments

A common and also very comprehensive method to correct for classical additive measurement error and thus to avoid biased coefficient estimates when using the OLS estimator is the *method of moments* (see, e.g., Carroll et al., 2006, Capter 3). As shown in subsection 4.1, the expectation for the naive coefficient estimate $\hat{\boldsymbol{\beta}}_{1naive}$ can be expressed as the corresponding true coefficient $\boldsymbol{\beta}_1$ times a factor. This factor is constituted by the reliability ratio $\boldsymbol{\kappa}$, which is defined as $\boldsymbol{\kappa} = (\boldsymbol{\Sigma}_{xx} + \boldsymbol{\Sigma}_{uu})^{-1}\boldsymbol{\Sigma}_{xx}$. If one had notice of the reliability ratio, one could simply calculate the naive OLS estimator $\hat{\boldsymbol{\beta}}_{1naive}$ and subsequently correct for its bias by dividing $\hat{\boldsymbol{\beta}}_{1naive}$ by the reliability ratio $\boldsymbol{\kappa}$. However, since $\boldsymbol{\kappa}$ is unlikely to be known in practice, it typically needs to be estimated.

To derive an appropriate estimator $\hat{\boldsymbol{\kappa}}$ for the reliability ratio, we first assume that the individual or common estimates of the measurement error variances and covariances are available. Let the estimator for the measurement error variance of the ith unit be denoted by $\hat{\boldsymbol{\Sigma}}_{ui}$. If the measurement error variances are heteroscedastic, i.e., if $\hat{\boldsymbol{\Sigma}}_{ui}$ changes with i, the estimated average or inter-individual measurement error variance is $\hat{\boldsymbol{\Sigma}}_{uu} = \sum_{i=1}^{n} \hat{\boldsymbol{\Sigma}}_{ui}/n$. Alternatively, it might hold that $\hat{\boldsymbol{\Sigma}}_{ui} = \hat{\boldsymbol{\Sigma}}_{uu}$ which means that the measurement error variances are constant.

Given the fact that the OLS estimator and the reliability ratio only depend on moments of the observed data, the method of moments provides an appropriate estimator for $\boldsymbol{\beta}$

in the presence of measurement error (Carroll et al., 2006). Using the above introduced relations $\hat{\mathbf{\Sigma}}_{xx} = \mathbf{C}_{ww} - \hat{\mathbf{\Sigma}}_u$ and $\hat{\mathbf{\Sigma}}_{xy} = \mathbf{C}_{wy}$ and assuming that $\hat{\mathbf{\Sigma}}_{uq} = 0$, the estimator $\hat{\boldsymbol{\kappa}}$ can be written as

$$\hat{\boldsymbol{\kappa}} = (\hat{\mathbf{\Sigma}}_{xx} + \hat{\mathbf{\Sigma}}_{uu})^{-1}\hat{\mathbf{\Sigma}}_{xx} = (\mathbf{C}_{ww} - \hat{\mathbf{\Sigma}}_u + \hat{\mathbf{\Sigma}}_u)^{-1}(\mathbf{C}_{ww} - \hat{\mathbf{\Sigma}}_u) = \mathbf{C}_{ww}^{-1}(\mathbf{C}_{ww} - \hat{\mathbf{\Sigma}}_u). \tag{73}$$

Provided that there is an estimator for the measurement error variance $\mathbf{\Sigma}_{uu}$, one can calculate $\hat{\boldsymbol{\kappa}}$ and subsequently use it to correct the naive estimate by dividing it by $\hat{\boldsymbol{\kappa}}$:

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{C}_{ww} - \hat{\mathbf{\Sigma}}_u)^{-1}\mathbf{C}_{ww}\hat{\boldsymbol{\beta}}_{1naive} = \hat{\boldsymbol{\kappa}}^{-1}\hat{\boldsymbol{\beta}}_{1naive}. \tag{74}$$

Directly formulating the *unweighted moment corrected estimators* (Buonaccorsi, 2010) leads to

$$\hat{\boldsymbol{\beta}}_{1(moment)} = \hat{\mathbf{\Sigma}}_{xx}^{-1}\hat{\mathbf{\Sigma}}_{xy} \tag{75}$$

and

$$\hat{\beta}_{0moment} = \bar{D} - \hat{\boldsymbol{\beta}}_{1moment}^{\mathbf{T}}\bar{\mathbf{W}}, \tag{76}$$

where $\hat{\mathbf{\Sigma}}_{xx} = \mathbf{C}_{ww} - \hat{\mathbf{\Sigma}}_u$ and $\hat{\mathbf{\Sigma}}_{xy} = \mathbf{C}_{wd} - \hat{\mathbf{\Sigma}}_{uq}$.[13] For cases where there is no error in the response, $\hat{\mathbf{\Sigma}}_{xy}$ equals $\mathbf{C}_{wd}$.

4.2.2 Error correction using replicate data

In this subsubsection, we describe how to obtain an estimate of the measurement error variances and covariances if there is replicate data available or if there is - at least - the possibility to perform replicate measurements to create such data. To this end, we ignore potential measurement error in the response and solely focus on the procedure for estimating covariate measurement error variances. We assume that all considered covariates are measured with additive error and that they are measured together in each replicate so that one has balanced replicate data. However, the prevalent literature also proposes estimation procedures for the case of unbalanced replicates. For an extended discussion of the usage of replicate data to estimate measurement error variances, we refer the interested reader to Buonacorrsi (2010, Chapter 5.4.3 and 6.5.1).

13 Note that the matrix $\hat{\mathbf{\Sigma}}_{xx} = \mathbf{C}_{ww} - \hat{\mathbf{\Sigma}}_u$ can generally become negative which then requires modifications of the estimators. This topic is discussed in Buonaccorsi (2010, Section 5.4.4.).

In the following, we assume that for each observation i ($i = 1, ..., n$) there are at least $m_i > 1$ corrupted replicate values W_{ija} available for each of the p covariates. The vector $W_{ia} \in \mathbb{R}^{(p+1)}$ contains the corrupted covariate values together with the intercept for the ith observation measured in the ath replication. Altogether, there are m_i such data vectors $\mathbf{W}_{i1}, \mathbf{W}_{i2}, ..., \mathbf{W}_{im_i}$ for the ith observation. Moreover, we assume that given y_i and $\mathbf{x}_i$ it holds that

$$\mathbf{W}_{ij} = \mathbf{x}_i + \mathbf{u}_{ij}, \tag{77}$$

with

$$\mathbb{E}(\mathbf{u}_{ij}) = \mathbf{0}, \tag{78}$$

$$\mathbb{V}ar(\mathbf{u}_{ij}) = \sigma^2_{ui(1)}, \tag{79}$$

and

$$\mathbb{C}ov(\mathbf{u}_{ij}) = \mathbf{\Sigma}_{ui(1)}, \tag{80}$$

where $\mathbf{x}_i \in \mathbb{R}^{(p+1)}$ is the vector with error-free covariate values belonging to the ith observation and $\mathbf{\Sigma}_{ui(1)}$ is the per-replicate measurement error covariance matrix for the ith observation which has the corresponding per-replicate variance $\sigma^2_{ui(1)}$ on its diagonal. Note that by assuming $\sigma^2_{uia(1)} = \sigma^2_{ui(1)}$ for each covariate a, we imply that the per-replicate error variance is the same for each covariate. Since we have $m_i > 1$ replicate vectors $\mathbf{W}_{ij}$ which contain corrupted covariate values, we can calculate the corresponding inter-replicate mean

$$\mathbf{W}_i = \sum_{j=1}^{m_i} \mathbf{W}_{ij}/m_i. \tag{81}$$

With $\mathbf{W}_i = \mathbf{x}_i + \mathbf{u}_i$, one then has

$$\mathbb{V}ar(\mathbf{u}_i) = \sigma^2_{ui} = \sigma^2_{ui(1)}/m_i, \tag{82}$$

and

$$\mathbb{C}ov(\mathbf{u}_i) = \mathbf{\Sigma}_{ui} = \mathbf{\Sigma}_{ui(1)}/m_i. \tag{83}$$

From Equation 83 we can conclude that one can estimate the measurement error variance by simply dividing the per-replicate measurement error variance $\mathbf{\Sigma}_{ui(1)}$ by the number of replicates m_i. Thus, we first need to estimate $\mathbf{\Sigma}_{ui(1)}$. Given $m_i > 1$, the per-replicate

sample variance of $\mathbf{W}_i$ denoted by $\mathbf{S}_{Wi}$ can be obtained by using Equation 84. Note that if we had covariates measured with error and such measured without error, any entries of $\mathbf{S}_{Wi}$ involving uncorrupted covariates would be zero.

$$\hat{\mathbf{\Sigma}}_{ui(1)} = \mathbf{S}_{Wi} = \frac{\sum_{j=1}^{m_i}(\mathbf{W}_{ij} - \mathbf{W}_i)(\mathbf{W}_{ij} - \mathbf{W}_i)^{\mathbf{T}}}{(m_i - 1)} \tag{84}$$

In situations where the measurement errors of two variables are correlated, one needs at least some paired replicate values for those two variables, say $m_{iab} > 1$ for variables a and b. One can then use

$$\hat{\sigma}_{uiab(1)} = S_{iab} = \frac{\sum_{j=1}^{m_{iab}}(W_{ija} - W_{ia})(W_{ijb} - W_{ib})}{(m_{iab} - 1)}, \tag{85}$$

where the sum is over those replications j for which W_{ija} and W_{ijb} are both available. An estimate of the abth element of $\mathbf{\Sigma}_{ui}$ is given by

$$\hat{\sigma}_{uiab} = S_{iab}/m_{iab}. \tag{86}$$

Combining Equation 83 and Equation 84 leads to $\hat{\mathbf{\Sigma}}_{ui} = \frac{\mathbf{S}_{W_i}}{m_i}$ as estimator for the ith measurement error covariance matrix. If we assume that the error variance is the same for all observations, one can set $\hat{\mathbf{\Sigma}}_{ui} = \hat{\mathbf{\Sigma}}_{uu}$. Otherwise one can simply estimate $\hat{\mathbf{\Sigma}}_{uu}$ by using Equation 61.

5 Analytical results for the linear LASSO with additive measurement error in the covariates

Introducing the linear LASSO and related theoretical aspects in section 2 and section 3, we assumed that the design matrix $\mathbf{X}$ passed to the LASSO contained error-free covariate measurements. However, in practice, where data corrupted by measurement error are rather the norm than the exception, this assumption mostly does not hold. Against this background, we now study the linear LASSO given in Equation 7 in the presence of measurement error in the design matrix. The analytical results with respect to the LASSO with error-prone design presented in this section involve classical additive measurement

error only.

In subsection 5.1, we outline the precise model setup for the subsequent analysis by combining the theoretical components introduced in section 2, section 3 and section 4. We then examine the impact of ignoring classical additive measurement error in the design on the standard LASSO's performance with regard to estimation and prediction (subsection 5.2) and concerning variable selection (subsection 5.3). In doing so, we use the central results for error-free design matrices presented in subsubsection 3.3.1 and subsubsection 3.3.2 as a reference. We then investigate how these results are influenced by additive measurement error in $\mathbf{X}$. The theoretical results given in this section are from Sørensen et al. (2014). We made, however, some notational modifications to achieve consistent notation throughout our work.

5.1 Model Setup

In this section, we apply the above presented classical measurement error framework to the linear regression model defined in Equation 1. Note that in line with subsection 3.1, we assume that the LASSO is applied to standardized data that satisfies $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i = 0$ and $\hat{\sigma}_j^2 = n^{-1}\sum_{i=1}^{n}(W_i^{(j)} - \bar{W}^{(j)})^2 = 1$ for all j. This assumption implies that the intercept of our linear model is zero and that all covariates are centered and measured on the same scale so that all diagonal elements of the error-prone Gram matrix $\mathbf{C}_{ww} = \frac{1}{n}\mathbf{W}^T\mathbf{W}$ equal 1. Note that applying the empirical mean $\bar{\boldsymbol{W}}^{(j)}$ for centring the error-prone covariate vectors $\mathbf{W}^{(j)}$ $(j = 1, ..., p)$ in $\mathbf{W}$, one has to be aware of the fact that the measurement error is inherent to the centring process. Therefore, we would like to point out that assessing the impact of empirical mean centring using the error-prone mean on the resulting LASSO estimates is surely a worthwhile focus for future research.

Altogether, we assume the following model setup:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{87}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the perfectly measured, continuous response observed on n individuals, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the true, but unknown coefficient vector for the p covariates and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the vector of model errors. We assume that the model errors $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$ are i.i.d. normally distributed with mean zero and variance σ^2, i.e., $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. However, unlike

in Equation 1, we observe $\mathbf{W} \in \mathbb{R}^{n \times p}$ containing noisy covariate measurements instead of the true design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. More in particular, we have

$$\mathbf{W} = \mathbf{X} + \mathbf{U}, \tag{88}$$

where the matrix $\mathbf{U} \in \mathbb{R}^{n \times p}$ contains the covariate measurement errors. $\mathbf{U}$ is assumed to have normally distributed rows with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}_{uu}$. Hence, $\mathbf{U}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{uu})$ which means that the measurement error variances and covariances do not change with i. Furthermore, we assume $\mathbf{U}$ and $\mathbf{X}$ to be independent (classical measurement error). Directly applying the LASSO given by Equation 7 to the above error-prone linear model, yields the *naive LASSO* (Sørensen et al., 2014):

$$\hat{\boldsymbol{\beta}}_{LASSO,naive}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left(1/n \, \|\mathbf{y} - \mathbf{W}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right). \tag{89}$$

Note that the only difference between Equation 7 and Equation 89 is that the latter contains the noisy design matrix $\mathbf{W}$ instead of the true design matrix $\mathbf{X}$. For the rest, we adopt the assumptions and notation outlined in subsection 3.1. Additionally, we introduce the partitioning of the error-prone design matrix $\mathbf{W} = (\mathbf{W}_{S_0}, \mathbf{W}_{S_0^c})$, where $\mathbf{W}_{S_0} \in \mathbb{R}^{n \times q}$ and $\mathbf{W}_{S_0^c} \in \mathbb{R}^{n \times (p-q)}$ are the first q and the last $(p-q)$ columns of $\mathbf{W} \in \mathbb{R}^{n \times p}$, respectively. Thus, $\mathbf{W}_{S_0}$ contains the n corrupted measurements of the q influential covariates, while $\mathbf{W}_{S_0^c}$ contains the n corrupted measurements of the $(p-q)$ unimportant covariates. Correspondingly, let $\mathbf{C}_{ww} = \frac{1}{n}\mathbf{W}^T\mathbf{W}$ be the scaled Gram matrix of the corrupted covariate measurements. In line with $\mathbf{C}_{xx}$, we can write $\mathbf{C}_{ww}$ in a block-wise form as

$$\mathbf{C}_{ww} = \begin{pmatrix} \mathbf{C}_{ww}(S_0, S_0) & \mathbf{C}_{ww}(S_0, S_0^c) \\ \mathbf{C}_{ww}(S_0^c, S_0) & \mathbf{C}_{ww}(S_0^c, S_0^c) \end{pmatrix}, \tag{90}$$

where $\mathbf{C}_{ww}(S_0, S_0) \in \mathbf{R}^{q \times q}$, $\mathbf{C}_{ww}(S_0^c, S_0^c) \in \mathbf{R}^{(p-q) \times (p-q)}$ and $\mathbf{C}_{ww}(S_0, S_0^c) = \mathbf{C}_{ww}(S_0^c, S_0)^T \in \mathbf{R}^{q \times (p-q)}$. Accordingly, the population covariance block matrix of $\mathbf{W}$ is given by

$$\boldsymbol{\Sigma}_{ww} = \begin{pmatrix} \boldsymbol{\Sigma}_{ww}(S_0, S_0) & \boldsymbol{\Sigma}_{ww}(S_0, S_0^c) \\ \boldsymbol{\Sigma}_{ww}(S_0^c, S_0) & \boldsymbol{\Sigma}_{ww}(S_0^c, S_0^c) \end{pmatrix}, \tag{91}$$

where the dimensions of the population block matrices equal those of the sample block matrices above.

The LASSO estimates are partitioned according to the same pattern as $\boldsymbol{\beta}$. This means that one has $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{Lasso,naive}(\lambda) = (\hat{\boldsymbol{\beta}}_{S_0}^T, \hat{\boldsymbol{\beta}}_{S_0^c}^T)^T$, where $\hat{\boldsymbol{\beta}}_{S_0} \in \mathbb{R}^q$ and $\hat{\boldsymbol{\beta}}_{S_0^c} \in \mathbb{R}^{(p-q)}$. Note that according to the above definition, the dependence of $\hat{\boldsymbol{\beta}}$ on λ is implicit. We would like to point out that since the LASSO does not necessarily provide correct estimations of the true $\boldsymbol{\beta}$, the elements of $\hat{\boldsymbol{\beta}}_{S_0}$ are not necessarily non-zero neither are the elements of $\hat{\boldsymbol{\beta}}_{S_0^c}$ necessarily zero. To differentiate the true active set S_0 from the active set of the LASSO, we define for any $\lambda \geq 0$ the active set of the LASSO as $\hat{S}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0\}$. Thus, $\hat{S}(\lambda)$ contains all LASSO non-zero coefficient estimates.

Note that unless otherwise specified, we assume fixed true covariates that satisfy the regularity conditions introduced in subsection 3.1 (see Equation 30 and Equation 31). In addition, we adopt the following regularity conditions for the random measurement errors in $\mathbf{U}$ defined by Sørensen et al. (2014):

$$\mathbf{C}_{uu} \to \boldsymbol{\Sigma}_{uu} \text{ , as } n \to \infty, \tag{92}$$

$$(1/n) \max_{1 \leq i \leq n} (\mathbf{u}_i^T \mathbf{u}_i) \to 0 \text{ , as } n \to \infty. \tag{93}$$

Since $\mathbf{U}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{uu})$, the convergences Equation 92 and Equation 93 can be shown to hold with probability 1 (see Anderson (2003, Th. 3.4.4) cited in Sørensen et al. (2014)). Given Equation 30, Equation 31, Equation 92 and Equation 93, it follows that $\mathbf{W} = \mathbf{X} + \mathbf{U}$ satisfies

$$\mathbf{C}_{ww} \to \boldsymbol{\Sigma}_{ww} \text{ , as } n \to \infty \tag{94}$$

and

$$(1/n) \max_{1 \leq i \leq n} (\mathbf{w}_i^T \mathbf{w}_i) \to 0 \text{ , as } n \to \infty \tag{95}$$

with probability 1.

5.2 Prediction and estimation accuracy

In subsubsection 3.3.1, we introduced two basic conditions that have to be satisfied for the LASSO to be consistent in estimation and prediction. More precisely, we allowed for a condition on the noise $\boldsymbol{\epsilon}$ (Equation 16) and the CC (Equation 17). Both of these conditions are crucial for establishing oracle results for the LASSO with regard to estimation and prediction. The oracle results for the LASSO with perfect design are given in Equation 18 to Equation 20. In summary, Equation 20 constitutes a probability inequality for the statistical error of the LASSO with fixed design and Gaussian errors. Given an appropriate choice for the tuning parameter λ, Equation 20 holds with high probability. In particular, the oracle inequalities for the LASSO provides bounds for the prediction and estimation error of the LASSO. Against the background of Equation 18 to Equation 20, we concluded in subsubsection 3.3.1 for the LASSO with true design matrix $\mathbf{X}$ that as long as the number of observations n tends to infinity with a rate faster than $q \ \log(p)$, the LASSO will be consistent for prediction (see Equation 21). If n tends to infinity with a rate faster than $q^2 \ \log(p)$, the LASSO will also be consistent for estimation in the l_1-norm (see Equation 22) (Sørensen et al., 2014). To derive corresponding results for the naive LASSO with corrupted covariate measurements $\mathbf{W}$, we need to consider two sources of random error. More precisely, we need to bound the observational noise $\boldsymbol{\epsilon}$ and the covariate measurement error $\mathbf{U}$. To establish oracle results for the LASSO with corrupted design, Sørensen et al. (2014) propose the following CC which involves the observed covariate measurements $\mathbf{W}$:

Compatibility condition for the naive LASSO with error-prone design

The CC is met for the index set S_0, for some $\phi_0 > 0$ and for all $\boldsymbol{\gamma} \in \mathbb{R}^p$ satisfying $\|\boldsymbol{\gamma}_{S_0^c}\|_1 \leq 3\|\boldsymbol{\gamma}_{S_0}\|_1$, if it holds that

$$\|\boldsymbol{\gamma}_{S_0}\|_1^2 \leq \frac{q \ \|\mathbf{W}\boldsymbol{\gamma}\|_2^2}{n \ \phi_0^2}. \tag{96}$$

As in the CC for uncorrupted design (Equation 17), the compatibility constant ϕ_0 in Equation 96 corresponds to a lower bound for the restricted eigenvalue of the Gram matrix $\mathbf{C}_{ww}$. By analogy to Equation 13 and Equation 14 for uncorrupted design, a basic inequality serves as technical base for all derivations concerning the estimation and prediction error of the naive LASSO with measurement error in the design. For the the latter, the *basic inequality* is given by Equation 97.

$$(1/n)\ \|\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq (1/n)\ \|\mathbf{y} - \mathbf{W}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{97}$$

Recall that we defined $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{Lasso,naive}(\lambda)$. Building on Equation 97, one can use Equation 96 combined with the bound for the observational noise given in Equation 98 to establish a bound for the estimation and prediction error of the naive LASSO with additive measurement error. The resulting bound is defined in Equation 99. Note that the derivation of Equation 98 is not trivial and requires several mathematical rearrangements and manipulations. However, for being rather technical the latter are not discussed here. The interested reader is refered to the appendix of Sørensen et al. (2014).

Prediction and estimation bound for the naive LASSO with error-prone design
Assume that the compatibility condition with constant ϕ_0 holds for S_0 and that there exists a constant λ_0 such that

$$(2/n)\ \|(\boldsymbol{\epsilon} - \mathbf{U}\boldsymbol{\beta})^T\mathbf{W}\|_\infty \leq \lambda_0 \tag{98}$$

Then, with a regularization parameter $\lambda \geq 2\lambda_0$, the following bound holds for the naive LASSO:

$$(1/n)\ \|\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq \frac{4\lambda^2\, q}{\phi_0^2}. \tag{99}$$

We can conclude from Equation 99 that - as for the LASSO with perfect design $\mathbf{X}$ - one can generally bound the estimation and prediction error for the LASSO if additive measurement error in the design is present. A problem lies, however, in the noise bound given in Equation 98. Sørensen et al. (2014) rearrange Equation 98 by means of the triangle inequality and obtain Equation 100 which implies the bound defined in Equation 98.

$$(2/n)\ \|\boldsymbol{\epsilon}^T\mathbf{W}\|_\infty + (2/n)\|\boldsymbol{\beta}^T\mathbf{U}^T\mathbf{X}\|_\infty + (2/n)\ \|\boldsymbol{\beta}\|_1\|\mathbf{U}^T\mathbf{U}\|_\infty \leq \lambda_0 \tag{100}$$

As stated by Sørensen et al. (2014), the naive LASSO with corrupted design would be consistent if all three terms in Equation 100 converged to zero. It is clear that the first two terms in Equation 100 converge to zero as we raise the sample size to infinity, since then $(2/n) \to 0$. By contrast, the term $\|\mathbf{U}^T\mathbf{U}\|_\infty$ is quadratic in the measurement error so that the third term on the left hand side of Equation 100 does not converge to zero. This is due to the fact that $\mathbf{U}^T\mathbf{U}$ converges to $n\boldsymbol{\Sigma}_{uu}$ as $n \to \infty$ so that the n factors cancel out while at

the same time it holds that $\|\boldsymbol{\Sigma}_{uu}\|_\infty \neq 0$. Hence, one can conclude that the naive LASSO with corrupted design is not consistent in estimation and prediction (see Sørensen et al., 2014). Against this background, Sørensen et al. (2014) provide the following asymptotic result for the naive LASSO estimates:

Assume $\lambda \to 0$ *as* $n \to \infty$*. Then, as* $n \to \infty$ *with fixed* p, one has

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\Sigma}_{ww}^{-1}\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}. \tag{101}$$

We can derive from Equation 101 that the resulting estimates are clearly biased if the naive LASSO is applied to covariate data corrupted by additive measurement error. On the other side, as pointed out by Sørensen et al. (2014), it has been shown by Knight and Fu (2000) that given a perfect design matrix, the estimates of the LASSO converge in probability to $\boldsymbol{\beta}$ under the same conditions ($\lambda \to 0$ as $n \to \infty$ and $n \to \infty$ with fixed p) as assumed for Equation 101. We can use the asymptotic result given by Equation 101 to correct the naive LASSO for the bias induced by additive measurement error in the following way:

$$\hat{\boldsymbol{\beta}}_{Lasso,corr} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{ww}\hat{\boldsymbol{\beta}}_{Lasso,naive} \approx (\mathbf{C}_{ww} - \boldsymbol{\Sigma}_{uu})^{-1}\mathbf{C}_{ww}\,\hat{\boldsymbol{\beta}}_{Lasso,naive} \tag{102}$$

It is not difficult to recognize that the above equation for measurement error correction parallels the correction mode for additive measurement error in the common linear model (see Equation 74). This implies that with λ being scaled properly, the bias induced by additive measurement error is the same for the LASSO as for a multivariate linear model (Sørensen et al., 2014; Carroll et al., 2006).

5.3 Variable selection accuracy

In subsubsection 3.3.2, we introduced the IC (Zhao and Yu, 2006) for the LASSO with perfect design matrix $\mathbf{X}$. In doing so, we showed that under certain regularity conditions the latter is sufficient and *essentially* necessary for the LASSO to select the true model both in the classical fixed p setting and in the large p setting where the number of covariates increases with n. We now turn to the topic of variable selection consistency for the naive

LASSO with measurement error in the design.

Equation 103 shows the SIC for the LASSO with corrupted design $\mathbf{W}$. In line with Sørensen et al. (2014), we refer to Equation 103 as *IC with measurement error (IC-ME)*. Note that the only difference between the IC-ME in Equation 103 and the SIC in Equation 27 is that the partitions of the perfect Gram matrix $\mathbf{C}_{xx}$ are replaced by the error-prone partitions of $\mathbf{C}_{ww}$ in the IC-ME. Against this background, the interpretation of the IC-ME parallels the one proposed for the SIC in subsubsection 3.3.2.

Irrepresentable condition with covariate measurement error (IC-ME): *The IC-ME holds if there exists a constant* $\theta \in [0, 1)$*, such that*

$$\|\boldsymbol{C}_{ww}(S_0^c, S_0)\boldsymbol{C}_{ww}(S_0, S_0)^{-1} sign(\boldsymbol{\beta}_{S_0})\|_\infty \leq \theta. \tag{103}$$

However, Sørensen et al. (2014) show that besides the IC-ME an additional condition is needed for the naive LASSO with corrupted design to sign consistently select covariates with high probability in the large sample limit. As a consequence, they introduce the *Measurement Error Condition (MEC)* which goes as follows:

Measurement Error Condition: *The MEC is satisfied if*

$$\boldsymbol{\Sigma}_{ww}(S_0^c, S_0)\boldsymbol{\Sigma}_{ww}(S_0, S_0)^{-1}\boldsymbol{\Sigma}_{uu}(S_0, S_0) - \boldsymbol{\Sigma}_{uu}(S_0^c, S_0) = \mathbf{0}. \tag{104}$$

First of all, it is noticeable that compared to the IC-ME which involves sample covariance matrices, the MEC involves population covariance matrices. From a technical point of view, Equation 104 is satisfied if $\boldsymbol{\Sigma}_{xx}(S_0^c, S_0) = \boldsymbol{\Sigma}_{uu}(S_0^c, S_0) = \mathbf{0}$. This can be explained by the fact that with $\mathbf{W} = \mathbf{X} + \mathbf{U}$ and $\mathbf{U}$ and $\mathbf{X}$ being independent, it holds that

$$\boldsymbol{\Sigma}_{ww}(S_0^c, S_0) = \boldsymbol{\Sigma}_{xx}(S_0^c, S_0) + \boldsymbol{\Sigma}_{uu}(S_0^c, S_0). \tag{105}$$

Thus, if $\boldsymbol{\Sigma}_{xx}(S_0^c, S_0) = \boldsymbol{\Sigma}_{uu}(S_0^c, S_0) = \mathbf{0}$, it follows that $\boldsymbol{\Sigma}_{ww}(S_0^c, S_0) = \mathbf{0}$ which then leads Equation 104 to hold. Note that the IC-ME requires the sample covariance matrix $\mathbf{C}_{ww}(S_0^c, S_0) = \mathbf{C}_{xx}(S_0^c, S_0) + \mathbf{C}_{uu}(S_0^c, S_0)$ to show a moderate degree of correlation. This implies that the empirical correlation between the unimportant error-prone covariates

in S_0^c and the important error-prone covariates in S_0 must be reasonably small for Equation 103 to hold. On the other hand, the MEC requires the population matrices $\mathbf{\Sigma}_{xx}(S_0^c, S_0)$ and $\mathbf{\Sigma}_{uu}(S_0^c, S_0)$ to be exactly zero. This requirement is satisfied in situations where there is absolutely no correlation between the unimportant covariates and the covariates being part of the true model. However, especially in high-dimensional data settings where p grows with n, such entire absence of correlation is rather unlikely to occur. Alternatively, the MEC holds if $\mathbf{\Sigma}_{uu} = c\mathbf{\Sigma}_{xx}$ for some constant c. This corresponds to situations where the population covariance matrix of the measurement errors has the same shape as the population covariance matrix of the true covariates (Sørensen et al., 2014).

For the finite sample case, Sørensen et al. (2014) show that the IC-ME serves to establish a positive lower bound on the probability of sign consistent covariate selection for the LASSO with additive error in the covariates. For the large sample case where p is fixed, they show that the MEC together with a couple of other conditions is sufficient for the LASSO to sign consistently select the true model with probability approaching one (see Theorem 1 in Sørensen et al. (2014)). While the IC forms a sufficient and essentially necessary condition for consistent covariate selection in the absence of measurement error, this is not the case for the LASSO with error-prone design matrix. For the latter, the IC-ME which is equivalent to the IC in the error-free case can be applied to establish a lower bound on the probability of consistent covariate selection. In addition, the MEC which is a much stronger condition than the IC-ME must also be satisfied for the LASSO to perform consistent covariate selection in the presence of measurement error. Note that the MEC is, however, not necessary for asymptotically sign consistent covariate selection. This means that the LASSO can still perform sign consistent covariate selection, even though the MEC might not be satisfied (Sørensen et al., 2014). In summary, it can be stated that the LASSO with measurement error does require a much stronger condition for sign consistent covariate selection than the LASSO with perfect design does. It should be mentioned that, while we already found it cumbersome to transfer the SIC into conditions for practical application, translating the population variances-based MEC into practically applicable terms appears almost impossible.

5.4 Correction methods for the LASSO with additive measurement error in the covariates

As for the classical linear model, the correction factor involved in Equation 102 represents a very straightforward solution to adjust the naive LASSO for the bias induced by additive measurement error in the design. In general, the literature on the LASSO with measurement error is very scarce and there are only very few studies that propose potential correction methods for the naive LASSO. A small strand of literature focusing on this topic and related questions has just emerged. Nevertheless, there are some interesting studies providing insightful results with respect to the performance of the LASSO in the presence of measurement error and potential approaches to correct for the bias induced by the latter.

Rosenbaum and Tsybakov (2010) studied a sparse linear regression model with $p \gg n$ whose design matrix was observed with additive measurement error. In doing so, they introduced the notion of *matrix uncertainty* refering to situations where $\mathbf{X}$ is observed with some kind of measurement error. They proposed a new class of *matrix uncertainty selectors (MU selectors)* and performed a MC simulation study to compare the latter with the naive LASSO and several other suitable estimators under matrix uncertainty. Examining the estimation accuracy and the number of estimated non-zero coefficients belonging to the true sparsity pattern, they showed the naive LASSO to be unstable in the selection of the true sparsity pattern and also to exhibit less efficiency in the estimation of $\boldsymbol{\beta}$ than the MU selectors. In particular, the naive LASSO lead to coefficients that were mostly too large and it also included many irrelevant covariates. Furthermore, the it turned out to be sensitive to the explicit values of the underlying $\boldsymbol{\beta}$. Nevertheless, Rosenbaum and Tsybakov (2010) also showed that the naive LASSO's sets of non-zero coefficients consistently included the true set of coefficients.

In their study of high-dimensional sparse linear regression, Loh and Wainwright (2012) discussed the issue of non-convex optimization problems due to noisy and missing covariate data. For the case of covariate data with additive measurement error, they presented the following reasoning to motivate why directly correcting the naive LASSO by using Equation 106 as estimator for $\boldsymbol{\Sigma}_{xx}$ cannot lead to meaningful results:
As mentioned before, the Gram matrix $\mathbf{C}_{xx} = 1/n\ \mathbf{X}^T\mathbf{X}$ is required to be positive semi-definite for the LASSO program in Equation 7 to be convex (Loh and Wainwright, 2012).

The convexity of optimization problems implies that any locally optimal point is also (globally) optimal (Boyd and Vandenberghe, 2004). On the other hand, non-convexity implies the presence of local minima. Note that $\mathbf{C}_{xx}$ is positive semi-definite if and only if $\lambda_i \geq 0$ for $i = 1, ..., n$ and $\lambda_i = 0$ for at least one eigenvalue (see subsection A.1). With measurement error as defined in Equation 55, one has $\mathbf{X} = \mathbf{W} - \mathbf{U}$, where $\mathbf{U} \in \mathbb{R}^{n \times p}$ contains the random covariate measurement errors.[14] $\mathbf{U}$ is independent of the true covariate measurements $\mathbf{X}$. In subsection 4.1, we introduced the relation $\mathbb{E}(\mathbf{C}_{ww}) = \mathbb{E}(\mathbf{C}_{xx}) + \boldsymbol{\Sigma}_{uu} \approx \boldsymbol{\Sigma}_{xx} + \boldsymbol{\Sigma}_{uu}$. Theoretically, we could apply this relation in order to estimate $\boldsymbol{\Sigma}_{xx}$ by means of Equation 106.

$$\hat{\boldsymbol{\Sigma}}_{xx,addME} = \frac{1}{n}\mathbf{W}^T\mathbf{W} - \boldsymbol{\Sigma}_{uu}. \tag{106}$$

$$\mathbb{E}(\|\mathbf{y} - \mathbf{W}\boldsymbol{\beta}\|_2^2 | \mathbf{X}, \mathbf{y}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + n\, \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{uu} \boldsymbol{\beta} \tag{107}$$

It can easily be seen from Equation 106 that given the absence of measurement error in the covariates, i.e., if $\boldsymbol{\Sigma}_{uu} = \mathbf{0}$, the estimator $\hat{\boldsymbol{\Sigma}}_{xx,addME}$ reduces to the error-prone Gram matrix comprised by the naive LASSO. On the other hand, $\hat{\boldsymbol{\Sigma}}_{xx,addME}$ is no longer positive semi-definite for $p \gg n$ if there is measurement error in the covariates which implies that $\boldsymbol{\Sigma}_{uu} > \mathbf{0}$. This can be explained by the fact that in high-dimensional data settings, $\mathbf{W}$ has at most rank n which means that $\mathbf{W}$ exhibits at least $(p - n)$ eigenvalues equal to zero. In this context, subtracting $\boldsymbol{\Sigma}_{uu} > \mathbf{0}$ ($\boldsymbol{\Sigma}_{uu} \in \mathbb{R}^{n \times p}$) from $\mathbf{C}_{ww} = \frac{1}{n}\mathbf{W}^T\mathbf{W}$ may result in $\hat{\boldsymbol{\Sigma}}_{xx,addME}$ showing a large number of negative eigenvalues. As an example, we state that if $\boldsymbol{\Sigma}_{uu} = \sigma_u^2\mathbf{I}$ with $\sigma_u^2 > 0$, $\hat{\boldsymbol{\Sigma}}_{xx,addME}$ has $p - n$ eigenvalues equal to $-\sigma_u^2$ and thus $p - n$ negative eigenvalues. This situation clearly leads to the non-convexity of the LASSO program. Hence, we conclude that inserting the above relation in Equation 89 to correct the naive LASSO for measurement error would turn the quadratic losses involved in Equation 89 into a non-convex optimization problem and would thus not be sensible (Loh and Wainwright, 2012). Note that if $\mathbf{C}_{xx}$ has at least one negative eigenvalue and is therefore not positive semi-definite, the objective equation in Equation 89 is unbounded from below and the RE condition proposed by Bickel et al. (2009) is violated. Given this context and the fact that the loss function of the naive LASSO is biased (see Equation 107),

14 Recall that $\mathbf{U}$ has its rows $\mathbf{u}_i$ drawn i.i.d. from $N(\mathbf{0}, \boldsymbol{\Sigma}_{uu})$.

Loh and Wainwright (2012) introduced the *regularized corrected LASSO (RCL)* which is given by Equation 108.

$$\hat{\boldsymbol{\beta}}_{LASSO,RC}(\lambda) = \underset{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_1 \leq R}{\arg\min}\{(1/n)\|\mathbf{y} - \mathbf{W}\boldsymbol{\beta}\|_2^2 - \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{uu}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1\} \tag{108}$$

They pointed out that the RCL facilitates sign consistent covariate selection under an IC-type condition and that the loss function of the RCL is always non-convex when $p > n$. Moreover, they showed that the RCL shows an improved estimation accuracy compared to the naive LASSO when additive measurement error is present. With respect to covariate selection consistency, Sørensen et al. (2014) showed that the RCL performs consistently under a condition which is very similar to the IC for the LASSO without measurement error. This means that compared to the MEC which is sufficient for the naive LASSO to consistently select covariates in the presence of measurement error, the RCL is a good alternative, since it requires conditions that are less strict and because it also shows a better performance than the naive LASSO.

6 Simulation study for the LASSO with additive measurement error in the covariates

This section aims at illustrating the finite sample performance of the LASSO in the presence of measurement error in the design. In the following, we focus on the LASSO for a linear model whose design matrix is corrupted by classical additive measurement error. In doing so, we draw comparisons between the naive OLS estimator, the naive LASSO as well as their corrected estimates, respectively. Furthermore, we allow for two different prevalent implementations of the linear LASSO in the `R` software package (R Development Core Team, 2013) to assess the differences between the resulting LASSO estimators. Another focus is on the distribution of the tuning parameter. In this context, we consider the impact of the general noise level ϵ on tuning parameter selection with cross-validation. The general structure of our simulation is presented in subsection 6.1. The corresponding results are presented and discussed in subsection 6.3.

6.1 Simulation model setup and methods

In our simulation, we generated data from the multiple linear regression model

$$\mathbf{Y}_k = \mathbf{X}\boldsymbol{\beta}_k + \boldsymbol{\epsilon}, \tag{109}$$

where $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ contains the error-free measurements of the p covariates included in the model together with a constant for the intercept and where $\boldsymbol{\beta}_k \in \mathbb{R}^{(p+1)}$ is the corresponding coefficient vector to estimate. The linear model presented in Equation 109 is sparse with active set $S_0 = \{1, ..., 5\}$ and $S_0^c = \{6, ..., p\}$, provided that one ignores the intercept constant. The corresponding sparsity index is $q = |S_0| = 5$ for either choice of $\boldsymbol{\beta}_k$ ($k = 1, 2, 3$). This means that irrespective of the intercept and for either choice of $\boldsymbol{\beta}_k$, the model in Equation 109 contains $q = 5$ influencing covariates with non-zero coefficients $\beta_{k,j}$ and $p - q$ irrelevant covariates with corresponding zero coefficients. As can be seen from Table 1, the parameter vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ only differ in terms of the magnitude of their non-zero coefficients. The model with $\boldsymbol{\beta}_1$ accounts for a situation where $q = 5$ out of p covariates have a reasonably strong impact on the response. The non-zero coefficients of $\boldsymbol{\beta}_2$ are half the size of the non-zero coefficients of β_1 which reflects a situation where the important covariates only have a moderate impact on the response. Finally, the coefficients of $\boldsymbol{\beta}_3$ are exponentially decreasing from coefficient $\beta_{3,1}$ to $\beta_{3,5}$. This describes a situation where the pertinent covariates are very heterogeneous with respect to their relative importance. Note that augmenting the number of covariates p according to Table 2 leads to smaller proportions q/p and hence to more sparsity in the above model.

	Parameter vector $\boldsymbol{\beta}_k$ ($k = 1, 2, 3$) belonging to non-zero and zero covariates	
	$\boldsymbol{\beta}_{k,S_0} \in \mathbb{R}^5$	$\boldsymbol{\beta}_{k,S_0^c} \in \mathbb{R}^{(p-5)}$
$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_{1,S_0} = (1, 1, 1, 1, 1)^T$	$\boldsymbol{\beta}_{1,S_0^c} = (0, ..., 0)^T$
$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_{2,S_0} = (0.5, 0.5, 0.5, 0.5, 0.5)^T$	$\boldsymbol{\beta}_{2,S_0^c} = (0, ..., 0)^T$
$\boldsymbol{\beta}_3$	$\boldsymbol{\beta}_{3,S_0} = (1, 0.5, 0.25, 0.125, 0.0625)^T$	$\boldsymbol{\beta}_{3,S_0^c} = (0, ..., 0)^T$

Table 1: Composition of the parameter vectors $\boldsymbol{\beta}_k$ ($k = 1, 2, 3$)

We performed simulations which involved two different sample sizes ($n = 100, 500$). Table 2 provides an overview of the pertinent parameters and their respective levels. To compare the naive OLS and LASSO estimators (Equation 2 and Equation 89) with their corrected analogues (Equation 75 and Equation 102) we implemented the following simulation procedure:

- The parameter vectors $\boldsymbol{\beta}_k$ ($k = 1, 2, 3$) accounting for three different data situations were defined according to Table 1. Ignoring the intercept term in $\boldsymbol{\beta}_k$, every $\boldsymbol{\beta}_k$ has the active set $S_0 = \{1, 2, 3, 4, 5\}$ with corresponding sparsity index $q = 5$.
- The variance-covariance (or rather correlation) matrix for the p covariates was predefined as $\boldsymbol{\Sigma}_{xx} \in \mathbb{R}^{p \times p}$, with $diag(\boldsymbol{\Sigma}_{xx}) = (1, ..., 1)$ and the off-diagonal entries corresponding to ρ. Note that ρ denotes the constant correlation between the p covariates.
- The design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ was generated with rows $\mathbf{x}_i$ i.i.d. distributed according to $\mathcal{N}_p(1, \boldsymbol{\Sigma}_{xx})$. The constant for the intercept was then added to $\mathbf{X}$ as first column.
- The responses $\mathbf{y}_k$ for $k = 1, 2, 3$ were generated applying Equation 109. $\boldsymbol{\epsilon} \in \mathbb{R}^n$ was i.i.d. drawn from $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\mathbf{I}_n$ denotes the identity matrix. Note that σ^2 was chosen according to Table 2, respectively.
- The error-prone design matrix $\mathbf{W} = \mathbf{X} + \mathbf{U}$ was created by adding classical measurement error $\mathbf{U}$ to the true covariate values in $\mathbf{X}$. The rows of $\mathbf{U}$ were i.i.d. drawn from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{uu})$ with $\boldsymbol{\Sigma}_{uu} = \sigma_u^2 \mathbf{I}_p$. Hence, we assumed that all p covariates were measured with constant measurement error which implies that the measurement error variance σ_u^2 does change neither with j (index for covariates) nor with i (index for observations). Note further that for one individual i, the measurement errors of different covariates are independent from each other. This means that they are uncorrelated and that the rows $\mathbf{w}_i$ and $\mathbf{x}_i$ in $\mathbf{W} = \mathbf{X} + \mathbf{U}$ are independent. Finally, the model noise $\boldsymbol{\epsilon}$ is independent from the measurement error in the covariates σ_u^2.
- In subsection 3.1 and subsection 5.1, we proposed to apply the LASSO to linear models with zero intercept and standardized covariates. For the estimation of the LASSO in our simulation, we ignored the intercept term in Equation 109 by passing $\mathbf{W}[, -1]$ to the respective `R` functions. The standardization of the covariates was then automatically performed by the latter.
- The naive OLS and the naive LASSO estimates for $\boldsymbol{\beta}_k$ ($k = 1, 2, 3$) were computed. Thereby, two alternative estimation methods for the LASSO were considered. The first estimate was computed by means of the `R` package `GLMNET` (Friedman et al., 2010). The tuning parameter λ was chosen according to the minimal value of the 10-fold cross-validation curve calculated by means of the `cv.glmnet` function with default parameters. The second estimate was obtained by means of the `R` package `penalized`

(Goeman et al., 2010) and the function `OptL1` therein. The tuning parameter λ was chosen according to the minimal value of the 10-fold cross-validation curve calculated by means of `OptL1` with default parameters. Note that using the default parameters of the above functions implies that all covariates are standardized properly.

- The corrected versions of the OLS and the LASSO estimates were subsequently calculated applying Equation 75 and Equation 102. In doing so, we assumed two different scenarios: in the first scenario, we assumed that the measurement error variance σ_u^2 was known and could directly be used to correct the naive estimates; in the second scenario, we assumed that σ_u^2 was unknown and hence had to be estimated. $\hat{\sigma}_u^2$ was calculated as described in subsubsection 4.2.2 by using two error-prone replicates of each covariate measurement which we assumed to be available in the absence of a notion of the true σ_u^2.
- Overall, the above procedure was performed for the different value combinations of p, ρ, σ_u^2 and ϵ displayed in Table 2. For each unique parameter combination the above procedure was repeated 500 times (MC simulation with 500 runs).

For each estimator, the estimation accuracy was measured by the global bias which equals the bias induced by the over- or underestimation of the $p - 5$ zero coefficients as well as by the MSE. Note that if we say that an estimator is *unbiased*, we only refer to the fact that its *global bias* (as defined above) is zero. However, this does not imply that its overall bias is also zero. The variable selection performance of the LASSO estimators was evaluated by the number of non-zero coefficients that were correctly estimated as non-zero (true positives) and the number of zero coefficients that were incorrectly estimated as non-zero (false positives). The true positives are expressed in relation to the number of important covariates $q = 5$ (true positive fraction, TPF) whereas the false positives are expressed in relation to the number of irrelevant covariates $p - 5$ (false positive fraction, FPF). The simulation results are presented and discussed in subsection 6.3. All results presented are averages over 500 MC simulations. The numbers in parentheses are the corresponding standard errors for TPF and FPF. We also considered the distributions of the 500 cross-validated tuning parameters for different noise levels σ^2.

6.2 Estimating LASSO in R with the packages `GLMNET` and `penalized`

As indicated above, we used two functions implemented in the `R` software package (R Core Development Team, 2013) to estimate the LASSO in the above simulation procedure.

	β_1		β_2		β_3	
	$n = 100$	$n = 500$	$n = 100$	$n = 500$	$n = 100$	$n = 500$
Number of covariates p	10	50	10	50	10	50
	50	250	50	250	50	250
	150		150		150	
Correlation btw. covariates ρ	0.1	0.1	0.1	0.1	0.1	0.1
	0.5	0.5	0.5	0.5	0.5	0.5
	0.9	0.9	0.9	0.9	0.9	0.9
Measurement error variance σ_u^2	0.01	0.01	0.01	0.01	0.01	0.01
	0.10	0.10	0.10	0.10	0.10	0.10
	1.00	1.00	1.00	1.00	1.00	1.00
Noise variance σ^2	1.0	1.0	1.0	1.0	1.0	1.0
	2.5	2.5	2.5	2.5	2.5	2.5
	5.0	5.0	5.0	5.0	5.0	5.0

Table 2: Parameter combinations involved in the MC simulation study

More in particular, we used the packages `GLMNET` (Friedman et al., 2010) and `penalized` (Goeman, 2010, 2014) and the corresponding functions for LASSO estimation therein. The resulting estimates are compared in subsection 6.3.

The `penalized` package uses likelihood cross-validation for all models. Cross-validation is a popular technique used to assess the predictive quality of the penalized prediction model or to compare the predictive ability of models with different values of the tuning parameter λ (Goeman, 2014). In our simulation, we performed 10-fold cross-validation using the `optL1` function provided by the `penalized` package. A 10-fold cross-validation randomly divides the observations contained in the dataset into 10 disjunct groups of equal size. The first fold is then used as validation set, while the model is fitted on the remaining 9 folds (see, e.g., Hastie et al., 2009, Chapter 7). We used the default `standardize=TRUE` to standardize the covariates to unit second central moment before applying penalization. In general, `OptL1` optimizes the cross-validated log-likelihood with respect to the tuning parameter λ and thus allows to identify the optimal amount of regularization. Note that the allocation of the subjects to the folds is random which implies that λ can be considered to be a random variable. This consideration forms the motivation for examining the distributions of the optimal $\lambda's$ resulting from 500 MC simulations, respectively.

`OptL1` provides an efficient algorithm for l_1-penalized estimation. The latter follows the gradient to the maximum of the penalized likelihood, using a series of directional Taylor approximations. In doing so, it uses *Brent's algorithm* for minimization without derivatives (Brent, 1973) to identify the optimal value of the tuning parameter λ. While leaving the above approximation as an option for other model classes, `penalized` automatically applies the approximation method for linear models. The approximation method is much faster than common cross-validation and for linear models it actually provides exact answers. However, applying `OptL1` one should be aware of the fact that Brent's algorithm is guaranteed to work only for uni-modal functions. This implies that `OptL1` has the inherent risk of converging to a local instead of to a global optimum (Goeman, 2010). Since the cross-validated likelihood as a function of λ has very often several local optima, this consideration especially refers to l_1-optimization. Moreover, it is noticeable that although the results of the approximation are generally quite accurate, the approximation method has nevertheless the tendency to be overly optimistic in the sense that it leads to optimal values of λ that are a slightly smaller than the ones resulting from usual cross-validation (Goeman, 2014).
Besides the LASSO estimates resulting from `OptL1` with the function arguments chosen as indicated above, we also consider the LASSO estimates resulting from choosing λ according to the minimum of the 10-fold cross-validation curve computed by the `cv.glmnet` function (with default parameters) implemented in the `GLMNET` package (Friedman et al., 2011). The latter facilitates extremely efficient procedures for fitting the entire LASSO path for linear regression, logistic and multinomial regression models, Poisson regression and the Cox model. The underlying algorithm uses cyclical coordinate descent in a pathwise fashion (Friedman et al., 2011). Note that as the $\lambda's$ resulting from `OptL1`, also the results of `cv.glmnet` are random, since the folds are selected at random. Friedman et al. (2011) point out that this randomness can be reduced by running `cv.glmnet` many times and averaging the error curves.

6.3 Simulation results

We now present the results obtained by conducting the above described MC study regarding the performance of the LASSO in the presence of additive measurement error in the design. As far as the performance of the LASSO is concerned, we adopt the two main perspectives presented in subsection 3.3. More precisely, we consider the performance of the naive LASSO and its corrected versions with respect to estimation in subsubsec-

tion 6.3.1. Their variable selection performance is examined in subsubsection 6.3.2. To draw comparisons with the LASSO, in subsubsection 6.3.1 we also consider the estimation performance of the naive OLS estimator and its corrected estimates for some parameter combinations. In subsubsection 6.3.3, we briefly outline the topic of tuning parameter selection. In doing so, we examine the distributions of the tuning parameter λ constituted by 500 runs of the MC procedure. We do this for different combinations of the considered parameters. In the appendix (see subsubsection A.2.5), we provide some additional results. The latter regard the differences between the estimates resulting from the `cv.glmnet` function provided by the `GLMNET` package and the `OptL1` function provided by the `penalized` package in `R`.

6.3.1 Results with respect to parameter estimation consistency

In the following, we present the simulation results regarding the estimation accuracy of the naive OLS (OLS_{naive}) and LASSO ($Lasso_{naive}$) estimators in the presence of additive measurement error in the design. Recall that by *naive* we simply mean ignoring matrix uncertainty. We also allow for the corrected estimates of OLS and LASSO, respectively (see subsection 4.2 and subsection 5.3). The corrected estimates were computed by using Equation 75 for the OLS estimates and Equation 102 for the LASSO estimates. The indices $TMEV$ (*True Measurement Error Variance*) and $EMEV$ (*Estimated Measurement Error Variance*) are used to denote the LASSO and OLS estimates which were corrected by the true and the estimated measurement error variance, respectively. In this section, we only consider the LASSO estimates that were obtained by using `cv.glmnet` (package `GLMNET` by Friedman et al., 2011). The corresponding `OptL1`-LASSO estimates (package `penalized` by Goeman, 2010, 2014) are shown in the appendix of this work.[15] Moreover, in this section we only allow for the results implied by Gaussian noise with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. Table 3 to Table 13 summarize the empirical averages of the global bias and the MSE values for the naive OLS and LASSO estimates as well as for their corrected estimates, respectively. They account for several combinations of the parameters n, p, ρ and σ_u^2. The global bias and MSE values for further parameter combinations can be found in the appendix of this work (see Appendix A).

15 Note that the `cv.glmnet` LASSO estimates are denoted by $Lasso1, naive$, $Lasso1, TMEV$ and $Lasso1, EMEV$, respectively, while the `penalized` LASSO estimates are denoted by $Lasso2, naive$, $Lasso2, TMEV$ and $Lasso2, EMEV$.

In general, our results indicate that the naive LASSO - and therefore also the corrected LASSO estimates - are somehow susceptible to the underlying coefficient values in $\boldsymbol{\beta}$. In particular, the naive LASSO clearly shows the worst estimation accuracy (in terms of the highest global bias and MSE values) for $\boldsymbol{\beta}_1$. There is only a small difference in the global bias and MSE values computed for the estimates of $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$. Nonetheless, $\boldsymbol{\beta}_3$ exhibits the lowest global bias and MSE values for all considered parameter combinations. This is in line with prevalent theory on the LASSO which points towards the fact that the LASSO performs best for a small upto a moderate number of moderate-sized effects (see Tibshirani, 1996). In the following, we consider results for the estimates of $\boldsymbol{\beta}_3$ only. For corresponding results with respect to the estimates for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, we refer to subsection A.2 in the appendix.

Considering the estimation accuracy of the naive LASSO and OLS estimates (see Table 3, Table 4, Table 5 and Table 6), our first observation is that the global bias of the estimates resulting from either method are reasonably small and always positive. Moreover, for both naive estimators the global bias values are the largest if $p = 10$ and $n = 100$, while they gradually decrease with a growing number of parameters p and growing sample size n. This implies a fixed combination of the remaining parameters (ρ and σ_u^2). If $p = 250$ and $n = 500$, both estimators are almost unbiased (in terms of global bias), even if there is a high degree of covariate correlation ρ and a high level of measurement error variance σ_u^2. If $p = 10$, the naive OLS estimates are generally unbiased if the measurement error variance is small ($\sigma_u^2 = 0.01$). Nevertheless, a small global bias is induced by moderate or high measurement error variance ($\sigma_u^2 = 0.1$ and $\sigma_u^2 = 1$). The aforementioned effect is further enhanced with an increasing covariate correlation ρ. Interestingly, if $p = 50$ the tendency of global biases growing with σ_u^2 is at least partly still visible, but alleviated. Under the same conditions, the naive OLS is still unbiased also for $\rho = 0.5$ in combination with $\sigma_u^2 = 0.1$. However, if $\sigma_u^2 = 1$, there is a small bias which grows if ρ is raised. If for $p = 50$ the sample size is increased from $n = 100$ to $n = 500$, the naive OLS still shows a behaviour which is similar to the above described one. If $n = 500$ and $p = 250$, the OLS is unbiased except for moderately and highly correlated covariates ($\rho = 0.5$ or $\rho = 0.9$) combined with high measurement error variance ($\sigma_u^2 = 1$) where it, nevertheless, exhibits only a tiny global bias.
Basically, the naive LASSO behaves in a way that is very similar to the one described for the naive OLS. In general, the naive LASSO's global bias values do not strongly diverge

from the ones of the naive OLS. If $n = 500$, there is almost no difference between the bias values of the LASSO and the OLS. Nevertheless, if $p = 10$ and $\sigma_u^2 = 0.01$ or $\sigma_u^2 = 0.1$, the LASSO partly shows a larger bias than the OLS, while if $p = 10$ and $\sigma_u^2 = 1$, the LASSO and the OLS are almost identical in terms of bias. For growing p, the difference in bias between the LASSO and the OLS almost completely vanishes for all possible values of ρ and σ_u^2.
The LASSO estimates by themselves show increasing bias values with growing ρ and σ_u^2. However, with growing p and n the biases approach a value of zero. This may be interpreted as indication that the bias induced by the over- or underestimation of unimportant coefficients becomes less important with growing p.

With respect to the MSE values of the naive estimates, we observe that the naive LASSO exhibits smaller MSE values than the naive OLS for almost every value of p. Overall, the LASSO shows a reasonably low MSE level also for an elevated number of model parameters. In line with theory (see section 2), this reflects that despite the fact that the LASSO exhibits a larger global bias than the naive OLS, it is clearly favourable in terms of comparatively low MSE values. This advantage over the OLS appears particularly strong if the number of covariates p is raised. Also for small p ($p = 10$), the naive LASSO generally outperforms the naive OLS in terms of showing considerably smaller MSE values. However, note that if $\sigma_u^2 = 1$, both estimate types have almost identical MSE and global bias values regardless the actual value for ρ. This suggests that the presence of a reasonably high amount of error in the covariates for quite small p makes the LASSO lose its advantage in terms of low variance, while it makes the OLS lose its advantage in terms of low bias. As a result, both methods approach in terms of estimation accuracy measured by the MSE of their estimates. Interestingly, the degree of correlation among covariates represented by ρ does not seem to play a role in the above situation.

Overall, the MSE values of both the naive LASSO and the naive OLS increase with growing σ_u^2 and ρ. It is known that the problems of measurement error in the design and correlation among the model covariates by themselves already impede both estimation methods to correctly identify the underlying true model. Thus, it is not surprising that the combination of both adversities makes it especially hard for the methods to consistently estimate the coefficients of $\boldsymbol{\beta}_3$. Focusing on the naive OLS estimates, we can observe an interesting phenomenon. More specifically, for high ($\rho = 0.9$) and partly also for intermediate ($\rho = 0.5$)

correlation the MSE values of the naive OLS diminish with growing σ_u^2. For instance, if $n = 100$, $p = 50$ and $\rho = 0.9$, the naive OLS estimates have a MSE of 9.68 if $\sigma_u^2 = 0.01$, a MSE of 5.88 if $\sigma_u^2 = 0.1$ and a MSE of 2.16 if $\sigma_u^2 = 1$. Likewise, there results a MSE of 9.21 for $\sigma_u^2 = 0.01$, a MSE of 5.66 for $\sigma_u^2 = 0.01$ and a MSE of 2.12 for $\sigma_u^2 = 1$, if $n = 500$ and $p = 250$. Raising the sample size n has generally a mitigating effect on the MSE values for both estimate types, while their estimation accuracy becomes worse (global bias and MSE values increase) if ρ becomes larger. In summary, the MSE values of the naive LASSO increase with σ_u^2 for all values of p and ρ.

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.00	0.01	0.03	0.00	0.00	0.01				0.00	0.00	0.02	0.00	0.00	0.00
$\rho = 0.5$	0.00	0.02	0.10	0.00	0.00	0.02				0.00	0.01	0.02	0.00	0.00	0.01
$\rho = 0.9$	0.03	0.09	0.15	0.00	0.02	0.04				0.00	0.02	0.03	0.00	0.00	0.01

Table 3: Bias of the naive OLS estimator for β_3 (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.01	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
$\rho = 0.5$	0.03	0.04	0.10	0.01	0.01	0.02	0.00	0.01	0.01	0.01	0.01	0.02	0.00	0.00	0.01
$\rho = 0.9$	0.07	0.11	0.15	0.02	0.03	0.03	0.01	0.01	0.01	0.01	0.02	0.03	0.00	0.01	0.01

Table 4: Bias of the naive $LASSO_{\texttt{cv.glmnet}}$ estimator for β_3 (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.12	0.15	0.42	1.15	1.18	1.19				0.12	0.14	0.46	1.13	1.14	1.24
$\rho = 0.5$	0.20	0.22	0.56	2.04	1.89	1.52				0.21	0.23	0.66	1.95	1.89	1.55
$\rho = 0.9$	1.01	0.84	0.93	9.68	5.88	2.16				1.02	0.90	1.15	9.21	5.66	2.12

Table 5: MSE of the naive OLS estimator for β_3 (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.09	0.13	0.49	0.16	0.20	0.63	0.21	0.38	0.85	0.03	0.06	0.43	0.06	0.10	0.56
$\rho = 0.5$	0.12	0.16	0.56	0.23	0.30	0.75	0.34	0.39	1.04	0.05	0.11	0.63	0.09	0.16	0.73
$\rho = 0.9$	0.52	0.67	0.92	0.82	1.10	1.32	1.17	1.32	1.44	0.25	0.58	1.12	0.37	0.71	1.27

Table 6: MSE of the naive $LASSO_{\texttt{cv.glmnet}}$ estimator for β_3 (averages of 500 MC simulations, $\epsilon = 1$)

Table 7 to Table 13 report the simulation results for the corrected OLS and LASSO estimates. Recall that we used the true measurement error variance to calculate OLS_{TMEV} and $Lasso_{TMEV}$, while we assumed it to be unknown and therefore had to estimate it by means of two error-prone replicate measurements when calculating OLS_{EMEV} and $Lasso_{EMEV}$. As far as the performance of OLS_{TMEV}, $Lasso_{TMEV}$ and OLS_{EMEV}, $Lasso_{EMEV}$ is concerned, we find that the results provide no overall evidence of the efficacy of the applied measurement error correction factors. There are cases where at least the $TMEV$-corrected estimates exhibit a marginally smaller global bias than the naive estimates. For instance, if ρ=0.9, $p = 10$ and $\sigma_u^2 = 0.01$, one has $bias_{OLS,naive} = 0.03$ and $MSE_{OLS,naive} = 1.01$, $bias_{OLS,TMEV} = 0.01$ and $MSE_{OLS,TMEV} = 1.29$, and finally $bias_{OLS,EMEV} = 0.21$ and $MSE_{OLS,EMEV} = 424.72$. Under the same conditions, it results for the LASSO that $bias_{Lasso,naive} = 0.07$ and $MSE_{Lasso,naive} = 0.52$, $bias_{Lasso,TMEV} = 0.06$, $MSE_{Lasso,TMEV} = 0.61$ and $bias_{Lasso,EMEV} = 0.26$ and $MSE_{Lasso,EMEV} = 93.98$. However, the above example also shows that the MSE values of the corrected estimates are consistently larger than the for the naive estimators regardless of the estimation method used. For the majority of parameter combinations the increase in terms of MSE values exhibited by the corrected estimates as compared to their naive analogues is substantial. Here, we give the example of $\rho = 0.9$, $p = 10$ and $\sigma_u^2 = 1$, where one has $bias_{OLS,naive} = 0.15$, $MSE_{OLS,naive} = 0.93$ and $bias_{OLS,TMEV} = -0.13$, $MSE_{OLS,TMEV} = 2409.62$ and $bias_{OLS,EMEV} = -9.02$, $MSE_{OLS,EMEV} = 470450.00$. For the LASSO estimates, one observes under the same conditions that $bias_{Lasso,naive} = 0.15$, $MSE_{Lasso,naive} = 0.92$ and $bias_{Lasso,TMEV} = -0.12$, $MSE_{Lasso,TMEV} = 2269.63$ and $bias_{Lasso,EMEV} = -8.68$, $MSE_{Lasso,EMEV} = 435774.00$. A similar situation can be observed if $p = 50$, $n = 500$, $\rho = 0.5$ and $\sigma_u^2 = 0.01$. Here, the general MSE level of both $TMEV$-estimates and $EMEV$-estimates is, however, of smaller order of magnitude. In the last-mentioned example, the MSE values of the $TMEV$-estimates obtained by either method are considerably larger than those obtained for the naive estimates. However, they are by far smaller than the MSE values for the $EMEV$-estimates which

show enormous MSE values. The MSE values of the $TMEV$- and $EMEV$-estimates can even be said to literally explode compared to the MSEs of the naive estimates, at least in the majority of simulated scenarios. While, in general, we find the above described behaviour of the corrected OLS and LASSO estimates absolutely surprising, it seems intuitive that $TMEV$-estimates show smaller MSE values than $EMEV$-estimates. The reason therefore simply is that for the latter, we estimated the measurement error variance by means of Equation 83 and Equation 84. It is common sense that the use of estimated parameters instead of true parameters induces additional uncertainty. In the above case, this results in larger MSE values for the $EMEV$-estimates. In our simulation results, we also observe situations where - by contrast - $EMEV$-estimates show considerably smaller MSE values than $TMEV$-estimates. For instance, if $p = 50$, $n = 100$, $\rho = 0.9$, and $\sigma_u^2 = 1$, it results that $bias_{OLS,naive} = 0.04$, $MSE_{OLS,naive} = 2.16$, $bias_{OLS,TMEV} = 0.04$, $MSE_{OLS,TMEV} = 27793.11$ and $bias_{OLS,EMEV} = 0.00$, $MSE_{OLS,EMEV} = 5.48$. For the LASSO, it holds under the same conditions that $bias_{Lasso,naive} = 0.03$, $MSE_{Lasso,naive} = 1.32$ and $bias_{Lasso,TMEV} = 0.06$, $MSE_{Lasso,TMEV} = 3085.89$ and $bias_{Lasso,EMEV} = 0.00$, $MSE_{Lasso,EMEV} = 4.95$. However, based on the prevalent statistical theory, we are not able to explain the above described relation between the $TMEV$- and $EMEV$-estimates.

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.00	0.00	-0.21	0.00	-0.01	-0.02				0.00	0.00	-0.03	0.00	-0.01	0.00
$\rho = 0.5$	0.00	-0.01	-22.19	0.00	-0.07	-0.10				0.00	0.00	-0.15	0.00	0.00	-0.06
$\rho = 0.9$	0.01	0.36	-0.13	-0.04	0.09	0.04				0.00	0.08	0.06	0.00	-0.01	0.00

Table 7: Bias of the OLS estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.01	0.01	-0.12	0.00	0.01	-0.01	0.00	-0.01	-0.07	0.00	0.00	-0.04	0.00	0.00	0.00
$\rho = 0.5$	0.03	0.01	-21.80	0.01	-0.06	-0.06	0.01	0.01	0.00	0.00	0.00	-0.14	0.00	0.00	-0.01
$\rho = 0.9$	0.06	0.23	-0.12	0.02	0.05	0.06	0.01	0.01	0.01	0.01	0.04	0.05	0.00	-0.01	-0.01

Table 8: Bias of the $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

Overall, the above results propose that the phenomenon of exploding MSE values of the corrected LASSO and OLS estimates appears to be independent of the precise parameter

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.12	0.18	725.67	1.28	5718.10	12193.50				0.13	0.17	231422.00	1.25	7440.42	38465.90
$\rho = 0.5$	0.21	0.33	70245900.00	2.48	76129.60	23230.00				0.22	0.34	222620.00	2.32	155438.00	6495340.00
$\rho = 0.9$	1.29	76754.30	2409.62	8066.53	130568.00	27793.11				1.29	114371.00	5365.51	19331.20	113155.00	1616120.00

Table 9: MSE of OLS estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.09	0.12	281.20	0.15	381.95	4162.58	0.57	1532.49	32888.60	0.03	0.04	37340.20	0.05	2138.64	3280.45
$\rho = 0.5$	0.12	0.19	67750200.00	0.24	35195.40	7469.23	0.71	1466.48	6933.58	0.05	0.08	275087.00	0.08	15638.00	1098690.00
$\rho = 0.9$	0.61	22890.90	2269.63	152.09	6618.64	3085.89	253.21	13554.06	2506.51	0.24	23566.90	4457.17	787.35	20747.50	417329.00

Table 10: MSE of the $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.33	0.00	0.00	0.10	0.01	0.00				0.10	-0.01	0.00	0.01	-0.02	0.00
$\rho = 0.5$	0.50	-0.10	-0.02	0.05	-0.01	0.11				0.05	-1.26	0.00	0.01	0.01	0.00
$\rho = 0.9$	0.21	0.26	-9.02	0.05	-0.04	0.00				0.05	-0.06	-0.01	0.01	0.01	0.00

Table 11: Bias of the OLS estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	0.30	0.00	0.00	0.08	0.01	0.00	0.01	0.01	0.00	0.08	-0.01	0.00	0.00	-0.02	0.00
$\rho = 0.5$	0.48	-0.10	-0.02	0.06	-0.01	0.10	0.03	0.03	-0.04	0.05	-1.21	-0.01	0.01	0.01	0.00
$\rho = 0.9$	0.26	0.25	-8.68	0.04	-0.04	0.00	0.01	0.02	-0.04	0.04	-0.07	-0.01	0.01	0.01	0.00

Table 12: Bias of the $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	5461.30	7.63	2.93	4331.30	24.82	7.82				2161.83	25.71	2.67	18051.90	443.34	7.83
$\rho = 0.5$	546.77	87.50	2.01	9906.70	71.77	435.45				1714.72	38488.00	2.92	2935.00	1143.08	21.83
$\rho = 0.9$	424.72	427.14	470450.00	158.60	43537.80	5.48				134.41	23287.90	30.64	648.46	90.02	18.82

Table 13: MSE of the OLS estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	4546.07	6.04	3.02	3658.34	65.41	12.48	2194.10	256	30.43	2138.80	33.11	3.38	12344.30	110.49	3.41
$\rho = 0.5$	484.68	81.75	1.93	10332.30	56.57	570.64	19998.80	29426.40	72.36	1463.75	35428.10	4.13	1327.20	75.73	19.13
$\rho = 0.9$	93.98	404.59	435774.00	35.72	28558.00	4.95	384.86	2201.95	153.09	25.14	14252.40	10.97	68.00	9.69	16.97

Table 14: MSE of the $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

choice for the sample size n, the number of parameters p, the magnitude of the correlation among the design covariates ρ and also of the magnitude of covariate measurement error σ_u^2. On the other side, one can easily see that the corrected OLS and LASSO estimates are equally affected by the phenomenon of exploding MSE values. This suggests that the reason for the disproportionate MSE values arises during the correction procedure for both methods and estimate types ($TMEV$ and $EMEV$). By examining the distribution of MSE values of the corrected OLS and LASSO estimates in the context of a specific parameter combination ($p = 10$, $\rho = 0.5$, $\sigma_u^2 = 1$, $\epsilon = 1$), we found out that the empirical averages of the MSE values were in fact inflated by a few outliers. Against this background, we had a closer look at the computation of the corresponding correction factors that lead to such outliers in terms of disproportionate MSE values. In doing so, we found that the occurrence of outliers was due to the bad conditioning of the matrix $(\mathbf{C}_{ww} - \boldsymbol{\Sigma}_{uu})$ within specific runs of our MC simulation. As a measure of ill-conditioning of $(\mathbf{C}_{ww} - \boldsymbol{\Sigma}_{uu})$, we considered the condition number $P(A)$ defined in Equation 110.
$P(A)$ quantifies the relation between the largest and the smallest eigenvalue of $(\mathbf{C}_{ww} - \boldsymbol{\Sigma}_{uu})$. If $P(A)$ results in being very large, the difference between the minimal and the maximal eigenvalue of $(\mathbf{C}_{ww} - \boldsymbol{\Sigma}_{uu})$ is very large. Then, $(\mathbf{C}_{ww} - \boldsymbol{\Sigma}_{uu})$ is called ill-conditioned. On the other side, we expect a well-conditioned matrix to have $P(A) \approx 1$. If a system of equations bases on an ill-conditioned matrix, an inaccurate solution can be obtained which, nevertheless, appears to fulfil the underlying system quite well (Riley, 1955). Moreover, ill-conditioned matrices are positive definite, but can become singular and therefore non-invertible if some of its entries are slightly changed. The aforementioned fact implies that problems that involve ill-conditioned matrices are sensitive to very small changes in the data. In fact, the matrices $(\mathbf{C}_{ww} - \boldsymbol{\Sigma}_{uu})$ that belonged to MSE outliers had one or more eigenvalues that were very close to zero. Recall that a matrix loses rank (and therefore becomes non-invertible) if any of its eigenvalues are zero.

$$P(A) = \frac{|\lambda|_{max}}{|\lambda|_{min}} \tag{110}$$

Given the above finding, we can conclude that although using the empirical average as measure to describe the central tendency of the distribution for a given parameter is a common approach for the analysis of MC simulation results, its lack of robustness leads to misleading results if the MC procedure produces outliers. Therefore, the use of the empirical average to evaluate the MSE value distribution for different combinations of n, p, ρ and σ_u^2 represents a clear limitation of our work. We therefore recommend using robust measures, such as the median MSE, for the performance evaluation of the (corrected) OLS and LASSO estimates in future studies.

6.3.2 Results with respect to variable selection consistency

Table 15 and Table 16 summarize useful information on the variable selection consistency of the naive LASSO. More precisely, Table 15 presents the empirical averages of the TPFs, while Table 16 reports the empirical averages of the FPFs for a selection of parameter combinations examined in our simulation study. Note that by considering the above presented global definition of TPF and FPF, we do not check for sign consistency of the LASSO estimates.

If $p = 10$, the TPF of the naive LASSO is clearly larger for $\rho = 0.5$ than for $\rho = 0.1$. Surprisingly, the coincidence of high covariate correlation ($\rho = 0.9$) and large measurement error variance ($\sigma_u^2 = 1$) leads to a TPF of 94.64%. This implies that, in average, the LASSO correctly selects 94.64% of the important covariates under the above conditions. Note that the above TPF is higher than the TPF occurring for any other parameter combination. On the other hand, the coincidence of high covariate correlation and large measurement error variance also results in a dramatic increase of the LASSO's FPF (from approximately 40% for low and moderate covariate correlation and measurement error variance to approximately 90% for high covariate correlation and measurement error variance). This means that in the above case, the LASSO is more or less unable to differ between important and unimportant covariates. As a consequence, it makes a conservative selection by simply including almost all covariates. Note that this finding accords with theoretical results regarding the naive LASSO with measurement error (see Sørensen et al., 2014) and also with analytical findings with respect to the LASSO without measurement error in the design (see, e.g., Bühlmann and Van de Geer, 2011). Bühlmann and Van de Geer (2011) suggest

that, in general, the LASSO performs quite badly for variable selection, but nevertheless achieves good results in variable screening. With respect to the naive LASSO, Tibshirani himself proposed that the second "S" in the acronyme LASSO should rather stand for "screening" than for "selection" (Tibshirani, 2011).

Overall, increasing the sample size for fixed p considerably improves upon the selection capacity of the LASSO. More precisely, the TPFs of the LASSO are about 20%-30% higher, while its FPFs increase for only about 2%-12% if n is raised from 100 to 500 for $p = 50$. However, for fixed sample size, the LASSO shows lower TPFs but also lower FPFs if the number of covariates p is increased.

As far as the sensitivity of the LASSO to the level of covariate measurement error and covariate correlation is concerned, we note that disregarding the exceptional case of $p = 10$, the increase in both of these factors leads the LASSO to perform worse in variable selection. More in particular, both a high level of measurement error and a high level of covariate correlation imply that the LASSO selects a smaller number of important, but a larger number of irrelevant covariates at the same time.
Note that the precise impact of increasing the measurement error variance on the TPF of the naive LASSO is, however, not evident. If $p = 10$ or $p = 50$ (for $n = 100$ and $n = 500$), the TPFs increase with growing measurement error variance if there is a moderate or high level of covariate correlation ($\rho = 0.5$ or $\rho = 0.9$). This increase in the TPFs of the LASSO is accompanied by a markable rise in its FPFs (approximatively 20%-30%). Nevertheless, the above described improvement upon the LASSO's TPFs is somehow surprising, since it seems intuitive that a higher amount of measurement error in the design makes it harder for the LASSO to select the correct covariates. This intuition is also supported by theory (see Sørensen et al. (2014)). Our results indicate that for larger p ($p = 150$ and $p = 250$), the coincidence of high covariate correlation and high measurement error variance entails the worst variable selection performance of the LASSO for a given number of covariates p. Note that, in general, a high level of covariate correlation seems to prevent the LASSO more from including the important covariates than a high level of measurement error does. This aligns with the aforementioned tendency of the LASSO to only select one of the correlated covariates if there are several covariates with high pairwise correlation (Zou and Hastie, 2005; Friedman et al., 2010). On the other side, high levels of measurement error lead the LASSO to include more unimportant covariates as it is the case for only high covariate correlation. For instance, if $n = 500$ and $p = 50$ or $p = 250$, the FPFs given

$\rho = 0.1$ and $\sigma_u^2 = 1$ are about twice the size of the FPFs if $\rho = 0.9$ and $\sigma_u^2 = 0.01$.

We can conclude that for most parameter combinations, the LASSO estimates include the majority of important covariates, even though there are high levels of covariate correlation and/or measurement error variance. This observation is also in line with Bühlmann and Van de Geer (2011) and Tibshirani (2011) who both point towards the LASSO's reliability in terms of variable screening, also in situations where the conditions that are required for it to be consistent in variable selection are violated. Moreover, we find that although the LASSO is shown to perform sub-optimally in the presence of high covariate correlation in theory, it still exhibits a reasonably high TPF if the latter is present. This finding seems to be in line with Raskutti et al. (2010) who state that, in practice, l_1-methods tend to perform very well with regard to variable selection also in settings where the covariates are correlated and non-unitary. However, they also point out that there is no theoretical base so far that proves the performance of l_1-relaxations for dependent designs.

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	80.60	84.24	77.92	67.44	70.72	61.24	64.16	59.52	55.80	89.20	92.60	93.00	80.80	83.12	73.64
$\rho = 0.5$	82.60	88.80	94.04	67.72	69.80	74.40	60.40	57.96	51.00	88.92	88.72	90.80	79.36	77.68	77.16
$\rho = 0.9$	72.08	83.24	94.64	51.68	49.64	57.12	41.92	34.56	26.04	71.72	77.68	84.44	58.44	57.32	52.00

Table 15: TPFs of the naive $LASSO_{\texttt{cv.glmnet}}$ estimator for β_3 (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	44.40	44.96	44.52	16.53	17.72	18.40	8.48	10.23	11.86	19.34	23.36	40.72	7.12	8.20	15.54
$\rho = 0.5$	43.52	51.76	82.44	19.35	19.81	34.32	9.68	10.48	17.50	21.33	28.53	58.99	8.01	9.94	21.87
$\rho = 0.9$	44.28	59.84	90.44	17.49	24.43	44.29	9.24	11.35	20.22	22.19	36.34	70.36	7.22	11.59	26.31

Table 16: FPFs of the naive $LASSO_{\texttt{cv.glmnet}}$ estimator for β_3 (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.00	100.00	100.00	100.00	100.00	100.00	100.00
$\rho = 0.5$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$\rho = 0.9$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 17: TPFs of the $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$\rho = 0.5$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$\rho = 0.9$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 18: FPFs of the $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $TMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.00	100.00	100.00	100.00	100.00	100.00	100.00
$\rho = 0.5$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$\rho = 0.9$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 19: TPFs of the $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

	$n = 100$									$n = 500$					
	$p = 10$			$p = 50$			$p = 150$			$p = 50$			$p = 250$		
	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$	$\sigma_u^2 = 0.01$	$\sigma_u^2 = 0.1$	$\sigma_u^2 = 1$
$\rho = 0.1$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.00	100.00	100.00	100.00	100.00	100.00	100.00
$\rho = 0.5$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$\rho = 0.9$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 20: FPFs of the $LASSO_{\texttt{cv.glmnet}}$ estimator corrected with $EMEV$ for $\boldsymbol{\beta}_3$ (averages of 500 MC simulations, $\epsilon = 1$)

6.3.3 Results with respect to tuning parameter selection

In this section, we examine the distributions of the tuning parameter λ constituted by 500 runs of the simulation performed for each unique parameter combination. In doing so, we mainly focus on how the magnitude of λ is influenced if the values chosen for the noise level ϵ are altered. Moreover, we aim at evaluating if the two functions `cv.glmnet` and `OptL1` differ in terms of the amount of shrinkage applied when estimating the LASSO coefficients. As an example, we consider the parameter combination $p = 10$ and $\sigma_u^2 = 0.01$ for varying values of β_k, ρ and ϵ. Table 21 and Table 22 show some summary statistics (standard deviation (SD), median, range, skewness and kurtosis) for the distributions of 500 MC estimations for the optimal λ obtained by using `cv.glmnet` and `OptL1`, respectively.

As first implication of the results given in Table 21 and Table 22, we state that the general level of penalization is much higher if `OptL1` instead of `cv.glmnet` is used for LASSO coefficient estimation. Note that unlike $\lambda_{\texttt{OptL1}}$, $\lambda_{\texttt{cv.glmnet}}$ is reported on a log scale. This does, however, not explain the extreme difference between both functions in terms of the general level of penalization applied. More in particular, for a designated parameter combination the distributions differ by a factor of approximately 117 (in terms of their SD, medians and ranges).

Overall, we find that both functions show the same behaviour as far as their sensitivity to changes in the level of model noise ϵ and to changes in the degree of covariate correlation ρ is concerned. The general level of penalization measured by the median of the λ-distributions rises with increasing noise level. This observation is evident, since a higher level of model noise implies that the LASSO has to penalize harder and thus must spend more λ in order to correct for the random error induced by the model noise (see also subsubsection 3.3.1). Besides the increase in the general penalization level, the tables presented below suggest that a higher level of model noise leads to wider ranges as well as elevated SDs of the resulting distributions for λ. This indicates that there is more variation in tuning parameter selection for larger ϵ. Furthermore, we note that for $\epsilon = 2.5$ the skewness and the kurtosis of the λ-distributions tend to attempt a higher positive level than for $\epsilon = 1$, while for $\epsilon = 5$ compared to $\epsilon = 2.5$ the reverse effect occurs for some values of β_k and ρ. Recall that the skewness of a distribution quantifies how symmetrical it is. A skewness of zero indicates that the distribution is symmetrical. By contrast, positive skewness points towards the fact that the distribution is right-skewed and negative skewness implies that the considered distibution is left-skewed. On the

other hand, the kurtosis of a distribution quantifies to which extent the form of the latter satisfies the Gaussian distribution. More in particular, a Gaussian distribution has zero kurtosis, while a positive (negative) kurtosis indicates that the considered distribution is more peaked (flatter) than a Gaussian distribution (see Fahrmeir et al., 2011).

In general, we observe that the considered distributions of $\lambda_{\texttt{cv.glmnet}}$ exhibit a low negative or a low positive level of skewness if $\epsilon = 1$. However, they tend to be more right-skewed if $\epsilon = 2.5$ or $\epsilon = 5$. Note that the same trend can be observed for the kurtoses of the distributions which become more peaked if ϵ is raised. The penalization behaviour of `OptL1` is clearly less obvious than the one of `cv.glmnet`. However, overall one can state that the distributions of $\lambda_{\texttt{cv.glmnet}}$ and $\lambda_{\texttt{OptL1}}$ are quite close to being Gaussian if $\epsilon = 1$ (`cv.glmnet`) respectively if $\epsilon = 1$ and $\rho = 0.1$ (`OptL1`). They are substantially skewed and peaked if ϵ is elevated. Moreover, the results suggest that a high degree of covariate correlation ($\rho = 0.9$) implies highly right-skewed and very peaked distributions of $\lambda_{\texttt{cv.glmnet}}$ and $\lambda_{\texttt{OptL1}}$. At the same time, we see that the distributions exhibit mostly considerable lower SDs, smaller ranges as well as lower medians if $\rho = 0.9$ compared to if $\rho = 0.1$. This indicates that a higher degree of covariate correlation generally leads the LASSO to apply a smaller amount of shrinkage. In any case, the considered distributions are far from being Gaussian while the distributions of $\lambda_{\texttt{cv.glmnet}}$ are less skewed and exhibit flatter peaks than the distributions obtained by means of `OptL1`.

Finally, we state that the different underlying coefficients in $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ result in mostly just slightly different levels of penalization. Especially, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are very similar in terms of the amount of penalization applied by the LASSO to estimate them. However, with rising ϵ the difference in penalization applied to estimate the three different parameter vectors becomes larger. For $\boldsymbol{\beta}_2$ the highest amount of shrinkage is used. Note that the above finding seems to be independent of the used `R` function.

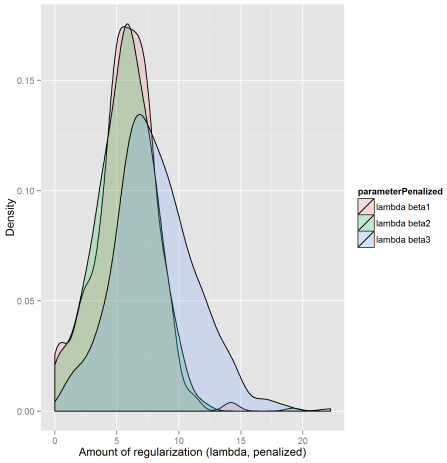
The density plots shown below give a further impression on the sensitivity of tuning parameter estimation to the different parameters involved in our simulation study.

	$\epsilon = 1$					$\epsilon = 2.5$					$\epsilon = 5$				
	Std. Deviation	**Median**	**Range**	**Skewness**	**Kurtosis**	**Std. Deviation**	**Median**	**Range**	**Skewness**	**Kurtosis**	**Std. Deviation**	**Median**	**Range**	**Skewness**	**Kurtosis**
$\beta = \beta_1$															
$\rho = 0.1$	0.02	0.05	0.11	0.02	-0.69	0.06	0.13	0.31	0.08	-0.26	0.27	0.36	1.51	1.54	2.82
$\rho = 0.9$	0.01	0.03	0.07	-0.10	5.05	0.03	0.08	0.22	0.41	1.43	0.09	0.23	0.56	0.23	0.90
$\beta = \beta_2$															
$\rho = 0.1$	0.02	0.05	0.11	-0.15	-0.61	0.12	0.17	0.72	1.45	3.14	0.28	0.68	1.37	0.07	-0.64
$\rho = 0.9$	0.01	0.03	0.07	0.10	-0.14	0.04	0.10	0.30	0.54	1.85	0.12	0.28	1.10	0.91	4.44
$\beta = \beta_3$															
$\rho = 0.1$	0.03	0.07	0.21	0.15	0.09	0.12	0.22	0.59	0.71	0.50	0.27	0.66	1.47	0.26	-0.28
$\rho = 0.9$	0.01	0.04	0.09	-0.04	0.09	0.04	0.12	0.26	0.32	0.72	0.17	0.30	1.21	2.06	6.78

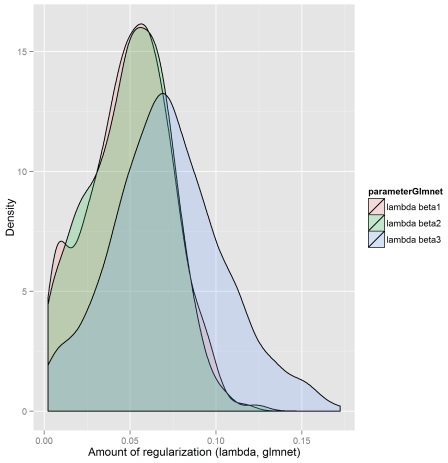
Table 21: Summary statistics for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (10-fold cross-validation, $p = 10$, $\sigma_u^2 = 0.01$)

	$\epsilon = 1$					$\epsilon = 2.5$					$\epsilon = 5$				
	Std. Deviation	**Median**	**Range**	**Skewness**	**Kurtosis**	**Std. Deviation**	**Median**	**Range**	**Skewness**	**Kurtosis**	**Std. Deviation**	**Median**	**Range**	**Skewness**	**Kurtosis**
$\beta = \beta_1$															
$\rho = 0.1$	2.47	5.86	13.41	-0.03	-0.14	5.94	14.85	31.38	-0.25	0.08	30.55	44.10	183.82	1.60	2.93
$\rho = 0.9$	1.63	2.98	16.21	1.71	8.25	4.20	8.09	25.79	1.60	3.39	8.94	23.19	48.68	0.79	0.53
$\beta = \beta_2$															
$\rho = 0.1$	2.53	5.96	13.87	-0.09	-0.05	14.70	19.95	86.10	1.76	3.57	33.11	75.87	179.10	0.27	-0.68
$\rho = 0.9$	1.61	3.31	9.53	1.25	2.41	4.16	9.66	38.18	1.59	5.89	12.56	28.59	122.76	2.59	14.55
$\beta = \beta_3$															
$\rho = 0.1$	3.44	8.22	21.11	0.23	0.25	10.64	25.41	98.02	1.02	3.86	32.09	72.51	157.53	0.44	-0.50
$\rho = 0.9$	1.63	3.58	10.94	1.34	2.92	4.65	10.60	36.15	1.49	3.98	19.70	29.79	171.35	3.32	15.38

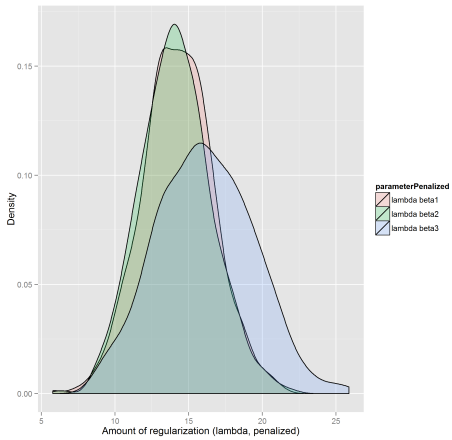
Table 22: Summary statistics for the tuning parameter $\lambda_{\texttt{OptL1}}$($p = 10$, $\sigma_u^2 =$ 0.01,10-fold cross-validation)



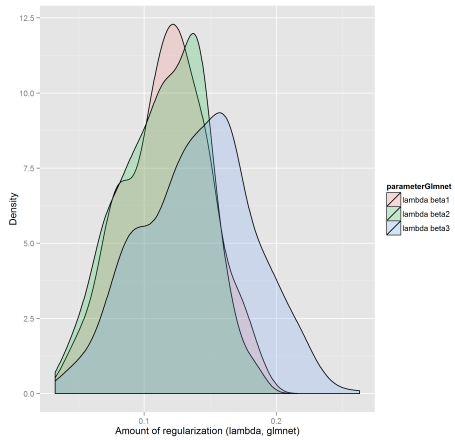
p=10 with `OptL1`



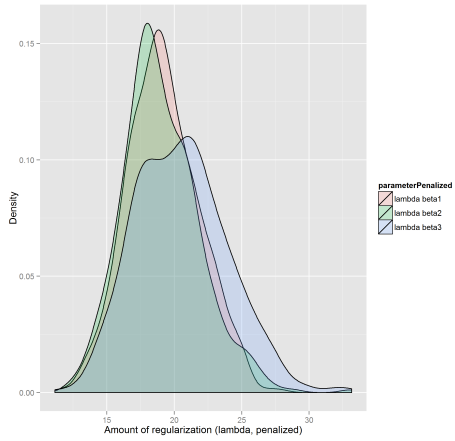
p=10 with `cv.glmnet`



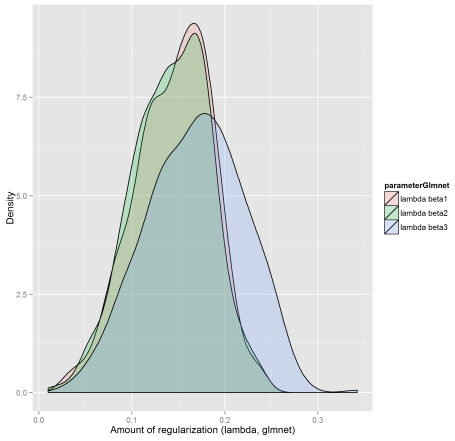
p = 50 with `OptL1`



p = 50 with `cv.glmnet`

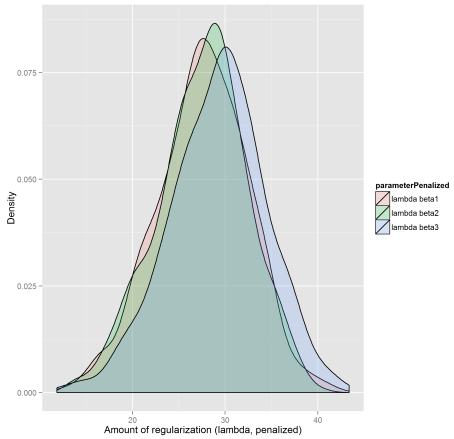


p = 150 with `OptL1`

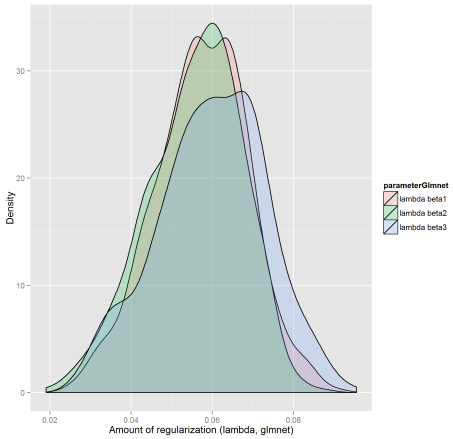


p = 50 with `cv.glmnet`

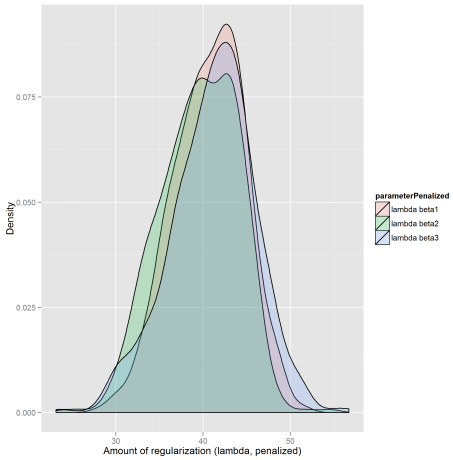
Figure 1: Density plots for the tuning parameters $\lambda_{\texttt{OptL1}}$and $\lambda_{\texttt{cv.glmnet}}$ ($n = 100$, ϵ=1, ρ=0.1, $\sigma_u^2 = 0.01$)



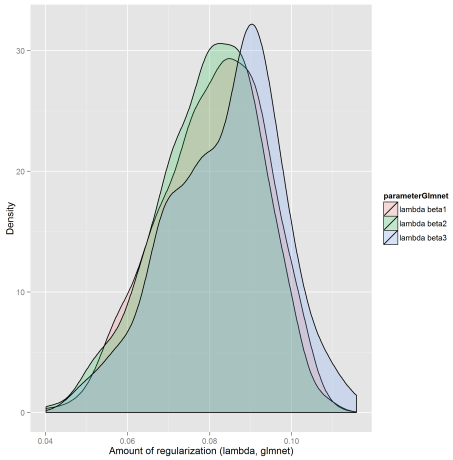
p=50 with `OptL1` function



p=50 with `cv.glmnet` function

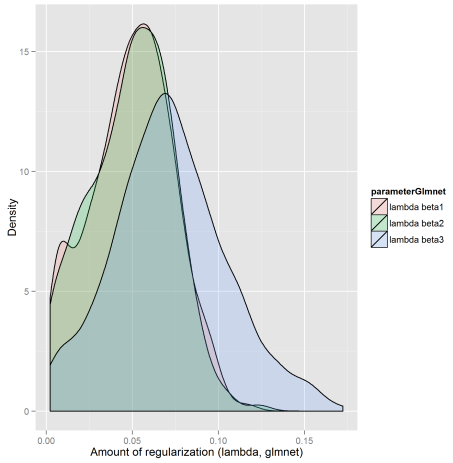


p=250 with `OptL1` function

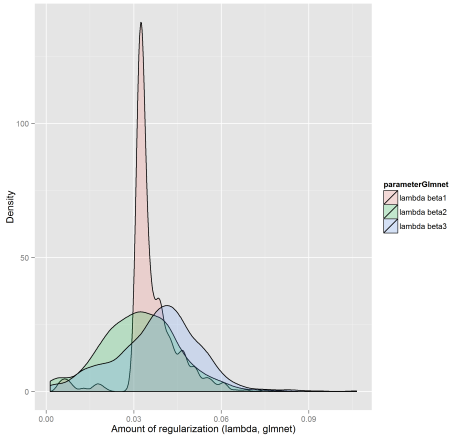


p=50 with `cv.glmnet` function

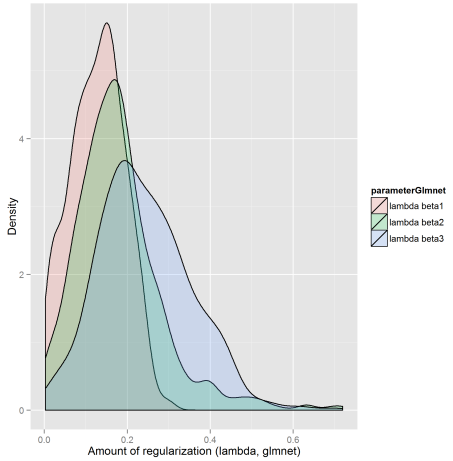
Figure 2: Density plots for the tuning parameters $\lambda_{\texttt{OptL1}}$and $\lambda_{\texttt{cv.glmnet}}$ ($n = 500$, ϵ=1, ρ=0.1, $\sigma_u^2 = 0.01$)



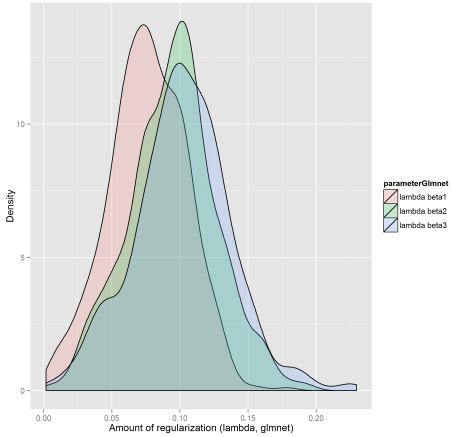
ϵ=1, ρ=0.1, σ_u^2 = 0.01



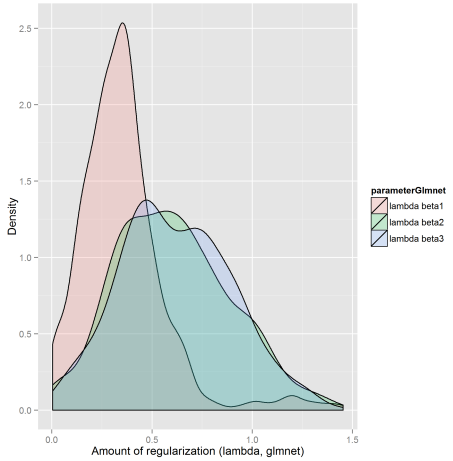
ϵ=1, ρ=0.9, σ_u^2 =0.01



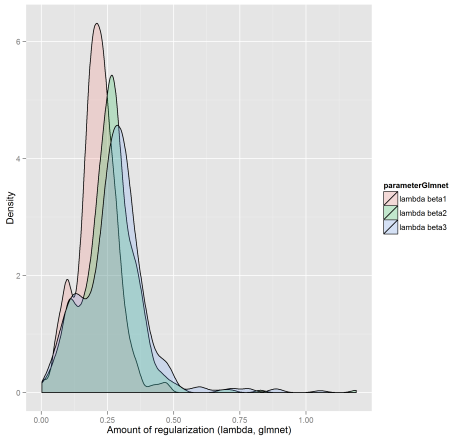
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



ϵ=2.5, ρ=0.9, σ_u^2 = 0.01

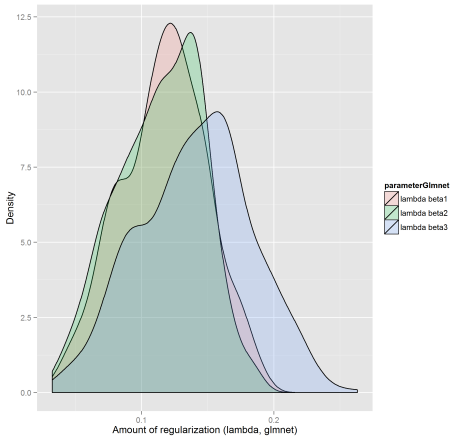


ϵ=5, ρ=0.1, σ_u^2 = 0.01

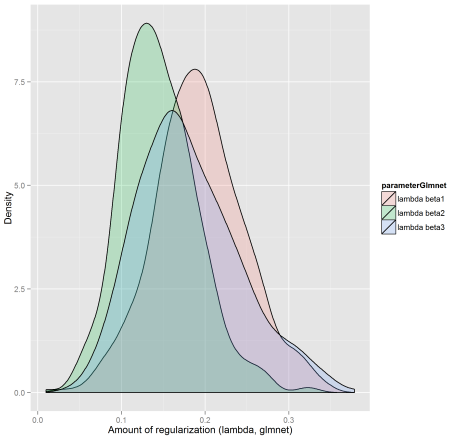


ϵ=5, ρ=0.9, σ_u^2 = 0.01

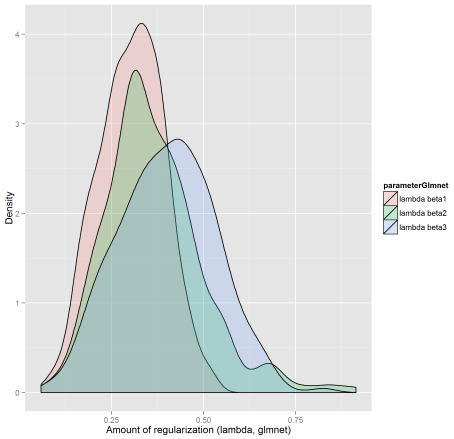
Figure 3: Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=10)



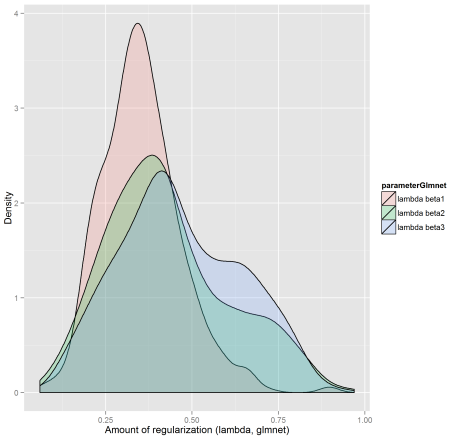
ϵ=1, ρ=0.1, $\sigma_u^2 = 0.01$



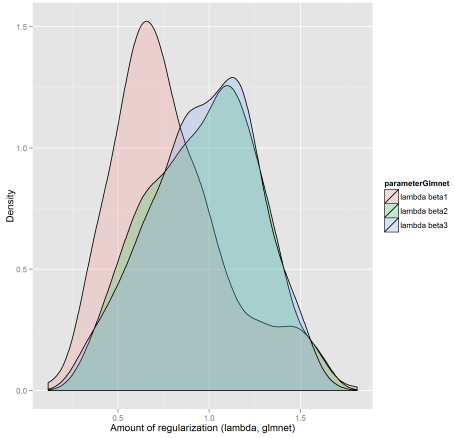
ϵ=1, ρ=0.1, $\sigma_u^2 = 1$



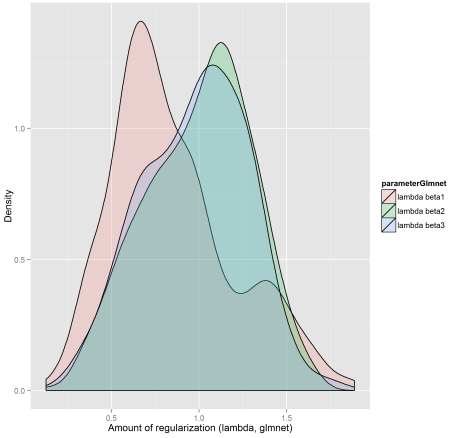
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



ϵ=2.5, ρ=0.1, $\sigma_u^2 = 1$

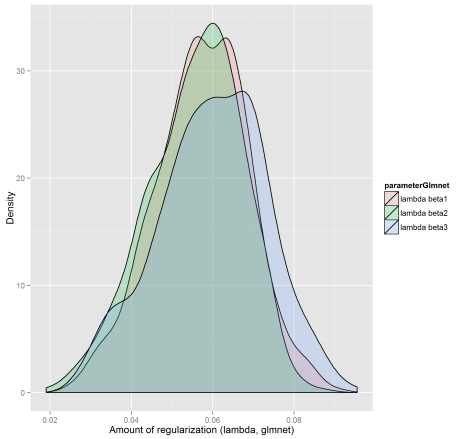


ϵ=5, ρ=0.1, $\sigma_u^2 = 0.01$

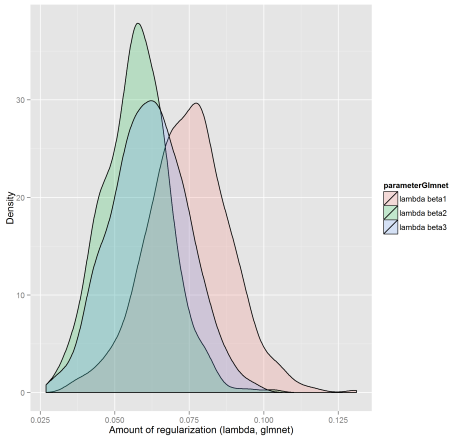


ϵ=5, ρ=0.1, $\sigma_u^2 = 1$

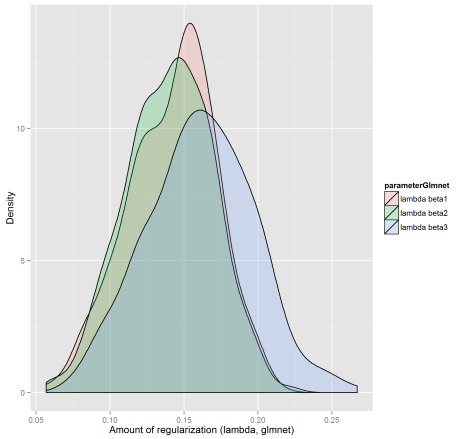
Figure 4: Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=50)



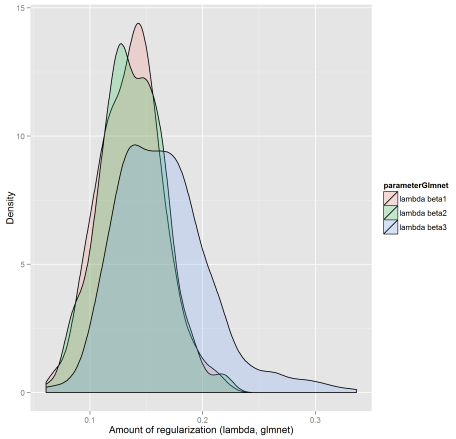
ϵ=1, ρ=0.1, $\sigma_u^2 = 0.01$



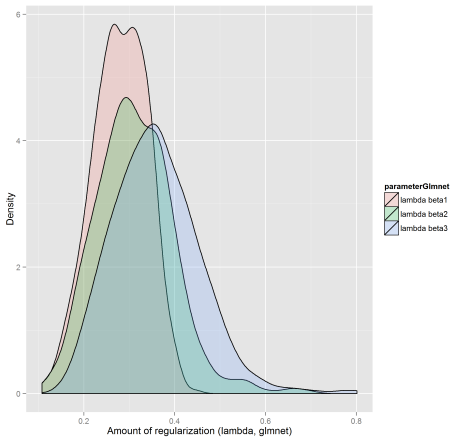
ϵ=1, ρ=0.1, $\sigma_u^2 = 1$



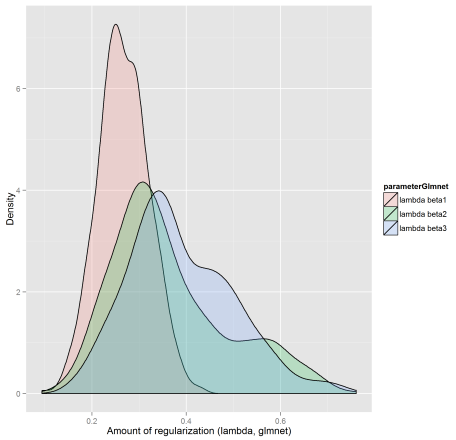
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



ϵ=2.5, ρ=0.1, $\sigma_u^2 = 1$

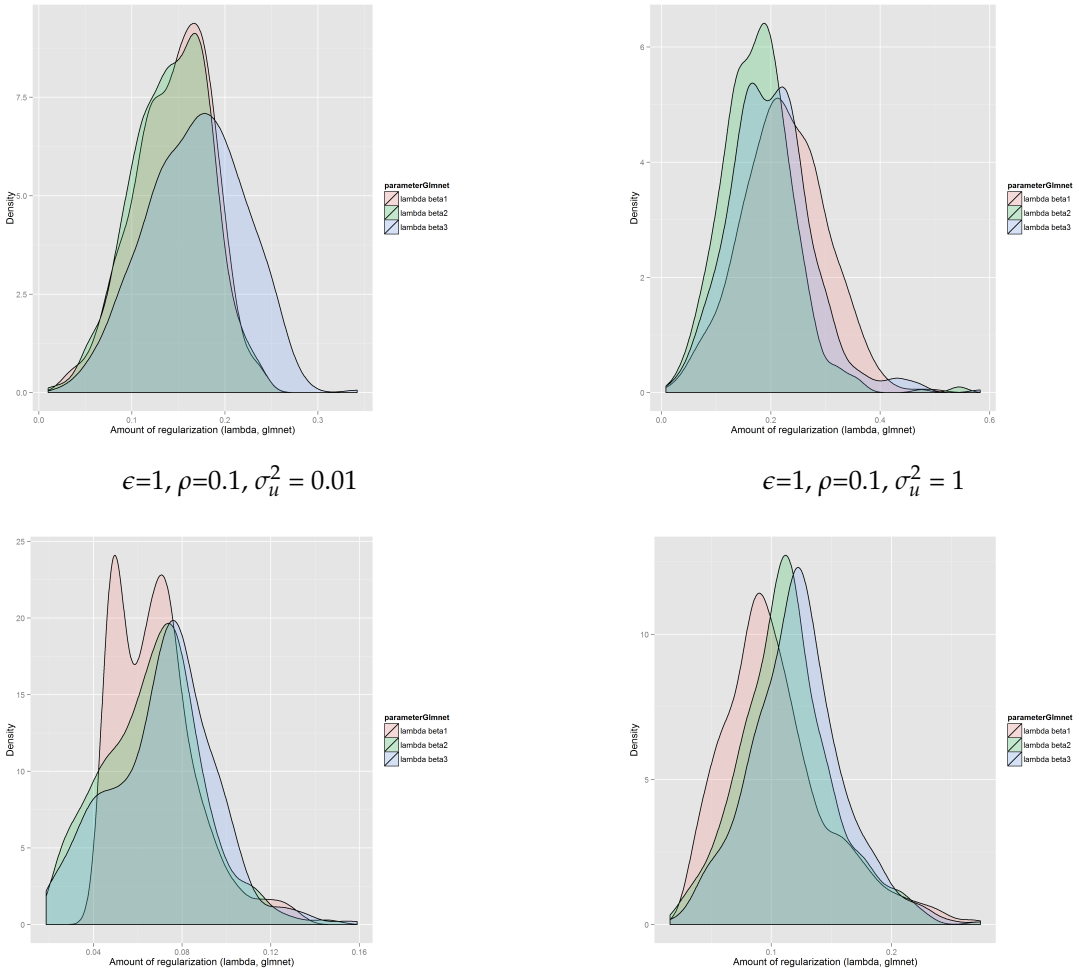


ϵ=5, ρ=0.1, $\sigma_u^2 = 0.01$



ϵ=5, ρ=0.1, $\sigma_u^2 = 1$

Figure 5: Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=500, p=50)



ϵ=1, ρ=0.1, $\sigma_u^2 = 0.01$ | ϵ=1, ρ=0.1, $\sigma_u^2 = 1$

ϵ=1, ρ=0.9, $\sigma_u^2 = 0.01$ | ϵ=1, ρ=0.9, $\sigma_u^2 = 1$

Figure 6: Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=150)

7 Discussion and conclusion

The LASSO (Tibshirani, 1996) is a powerful method which due to the fact that is uses a l_1-penalty allows for the estimation of regression coefficients and variable selection at the same time. It does variable selection in the sense that depending on the respective choice of the regularization parameter, the estimated coefficients can exactly be zero for some variables. Its ability to perform variable selection is the main feature which distinguishes it from the well-known Ridge regression method (Hoerl and Kennard, 1970). Another important property of the LASSO is that it can be applied even when the number of covariates p is larger than the number of observations n. This differs the LASSO from the popular OLS method which typically requires $p < n$. From a practical point of view, the LASSO produces sparse solutions and thus facilitates model interpretation when the number of covariates exceeds the number of observations.

In this work, we studied the LASSO in the presence of additive measurement error in the design matrix. In doing so, we allowed for analytical results on the estimation and variable selection consistency of both the LASSO with perfect design and the naive LASSO with additive covariate measurement error. Moreover, we performed a MC simulation study to assess the finite sample performance of the OLS and the LASSO under matrix uncertainty. Thereby, we also computed the corresponding corrected (in terms of adjusted by the effect of additive measurement error) estimates. In particular, we used the well-known reliability ratio (Fuller, 1987) to correct the OLS estimates and a reliability ratio-like factor according to Sørensen et al. (2014) to correct the naive LASSO estimates.

With regard to the theoretical background of the LASSO with perfect design, we introduced two basic conditions that need to be satisfied for the LASSO to be consistent in estimation and prediction. More precisely, we assumed a condition on the noise together with the CC (Bühlmann and Van de Geer, 2011) which imposes a lower bound on the RE of the Gram matrix $\mathbf{C}_{xx}$. The latter conditions lead the LASSO to be consistent in prediction and estimation if, in addition, the tuning parameter is chosen appropriately. Moreover, they are crucial for establishing oracle results which provide bounds for the prediction and estimation error of the LASSO. Overall, we illustrated that the LASSO enjoys good theoretical properties in the sense that its prediction error is about the same magnitude as the prediction error one would have if one knew a priori which covariates have an influence on the response (Bühlmann and Van de Geer, 2011). As far as the variable selection consistency of the LASSO is concerned, we introduced the IC (Zhao and Yu, 2006) which - under certain regularity conditions - is sufficient and essentially necessary for the LASSO

to select the true model both in the classical fixed p setting and in the large p setting where the number of covariates increases with n. As detailed in subsubsection 3.3.2, the term *essentially* refers to the fact that the necessary condition requires a quantity to be equal less than one, while the sufficient condition requires strict < 1 (Bühlmann and Van de Geer, 2011). The IC basically postulates that the covariance matrix of the design may not exhibit too strong degrees of linear dependence within smaller sub-matrices, in particular within the covariance sub-matrix of the irrelevant and the truly important covariates. If the IC is violated, the LASSO is unable to consistently recover the underlying true model. The reason therefore is that in order to produce sparse models, the LASSO shrinks the coefficient estimates belonging to important covariates too heavily. If the IC fails, the irrelevant covariates are enough correlated with the important covariates to be selected by the LASSO to outweigh the over-shrinkage of the non-zero coefficient estimates (Zhao and Yu, 2006). For statistical practice, this implies that one has to be aware of the fact that if unimportant variables are strongly correlated with covariates that are part of the true model, the LASSO is unlikely to select the true model. As mentioned before, the IC requires strong assumptions that are not verifiable in statistical practice. Although the LASSO might not be able to infer the correct set of covariates with non-zero coefficients from the data if the IC is violated, in practice it can nevertheless be used for variable screening which means that the model estimated by the LASSO includes the substantial covariates with high probability (Bühlmann and Van de Geer, 2011). Bühlmann and Van de Geer (2009, 2011) formally showed that the IC for variable consistency implies the CC for prediction and estimation accuracy. This indicates that using the LASSO for variable selection constitutes a more sophisticated problem than using it for parameter estimation and prediction.

With respect to the naive LASSO with corrupted covariate measurements, it can be stated that corresponding analytical results can barely be translated in words or conditions which are applicable in practice. We presented analytical results according to Sørensen et al. (2014). The results with regard to the estimation accuracy of the naive LASSO with corrupted design imply that in the presence of measurement error the LASSO does not consistently estimate the regression coefficients of the model. This means that the resulting estimates are clearly biased (see subsection 5.2). However, one can use the asymptotic results presented by Sørensen et al. (2014) to correct the naive LASSO for the bias induced by additive measurement error. Note that the correction factor for the naive LASSO basically equals the well-known reliability ratio used to correct for the attenuation bias caused by additive measurement error in the common linear model. This also implies

that with λ being scaled properly, the bias induced by additive measurement error is the same for the LASSO as for a multivariate linear model (Sørensen et al., 2014; Carroll et al., 2006). While in the absence of measurement error, the IC forms a sufficient and essentially necessary condition for consistent covariate selection, it does not guarantee consistent variable selection for the LASSO with error-prone design. For the latter, the IC-ME which is equivalent to the IC in the error-free case, serves to establish a lower bound on the probability of consistent covariate selection. In addition, the MEC, which is a much stronger condition than the IC-ME, must also be satisfied for the LASSO with error-prone design to perform consistently in covariate selection. Nevertheless, the MEC is not necessary for asymptotically sign consistent covariate selection, which means that the LASSO can still perform sign consistent covariate selection, even though the MEC might not be satisfied (Sørensen et al., 2014). In summary, it can be stated that the LASSO with measurement error requires a much stronger condition for sign consistent covariate selection than the LASSO with perfect design. Moreover, while we already found it hard to transfer the SIC into conditions for practical application, translating the population variances-based MEC into practically applicable terms appears almost impossible.

From the results of our MC simulation study (500 runs for each dedicated parameter combination) we conclude that in the absence of matrix uncertainty the LASSO generally outperforms the naive OLS in terms of exhibiting considerably smaller MSE values. This reflects that despite the fact that the LASSO exhibits a larger global bias than the naive OLS, it is clearly favourable in terms of exhibiting comparatively low MSE values. This advantage over the OLS appears particularly strong if the number of covariates is raised. However, our results also suggest that the presence of a reasonably high amount of error in the covariates for a quite small number of covariates makes the LASSO estimates lose their advantage in terms of low variance, while it makes the OLS estimates lose their advantage in terms of low bias. As a result, both methods approach in terms of estimation accuracy measured by the MSE of their estimates. Interestingly, the degree of correlation among covariates does not seem to play a role here. Overall, the MSE values of both the naive LASSO and the naive OLS increase with growing measurement error variance and covariate correlation. It is known that the problems of measurement error in the design and correlation among the model covariates by themselves already impede both estimation methods to correctly identify the underlying true model. Thus, it is not surprising that the combination of both adversities makes it especially hard for both methods to

consistently estimate the regression coefficients. With respect to the corrected LASSO and OLS estimates, our results suggest that there does not exist any overall evidence of the efficacy of the applied measurement error correction factors. Especially, the MSE values of the corrected estimates tend to be larger than the ones for the naive estimators, regardless of the estimation method used. Furthermore, for the majority of parameter combinations the increase in terms of MSE values exhibited by the corrected estimates as compared to their naive analogues is substantial. By examining the distribution of MSE values for corrected OLS and LASSO estimates in the context of a specific combination of parameters, we found out that the empirical averages of the MSE values were inflated by a few outliers and that the occurrence of outliers was due to the bad conditioning of the matrix $(\mathbf{C}_{ww} - \mathbf{\Sigma}_{uu})$ within specific runs of our MC simulation. In fact, those matrices $(\mathbf{C}_{ww} - \mathbf{\Sigma}_{uu})$ that belonged to MSE outliers had one or more eigenvalues that were very close to zero. Given the above finding, we conclude that although using the empirical average as measure to describe the central tendency of the distribution for a given parameter is a common approach for the analysis of MC simulation results, its lack of robustness leads to misleading results if the MC procedure produces outliers. Therefore, the use of empirical averages to evaluate the MSE value distributions represents a clear limitation of our work. We therefore recommend using robust measures, such as the median MSE, for the performance evaluation of the (corrected) OLS and LASSO estimates in future studies. As far as the variable selection consistency of the LASSO is concerned, our simulation results suggest that the coincidence of high covariate correlation and large measurement error variance leads the LASSO to be more or less unable to differ between important and unimportant covariates which is why it includes nearly all proposed covariates. This finding is also supported by theory which suggests that the LASSO performs quite badly for variable selection, but nevertheless achieves good results in variable screening. Moreover, Bühlmann and Van de Geer (2011) state that the LASSO for variable selection only works in a rather narrow range of problems excluding many cases where there are strong empirical correlations between the covariates. In general, the LASSO tends to select too many covariates, but on the other hand a certain amount of false negative selections cannot be avoided either if the absolute value of some β_j is below the LASSO's detection limit (Bühlmann and Van de Geer, 2011). As far as the sensitivity of the LASSO to covariate measurement error and covariate correlation is concerned, we note that the increase in both of these factors leads the LASSO to perform worse in variable selection. More in particular, high levels of both of these factors lead the LASSO to select a smaller number of important, but a larger number of irrelevant covariates at the same time. Note that,

in general, a high level of covariate correlation seems to prevent the LASSO more from including the important covariates than a high level of measurement error does. This aligns with the aforementioned tendency of the LASSO to only select one of the correlated covariates if there are several covariates with high pairwise correlation (Zou and Hastie, 2005; Friedman et al., 2010). On the other side, high levels of measurement error lead the LASSO to include more unimportant covariates as it is the case for only high covariate correlation. However, we can conclude that for most combinations, the LASSO estimates include the majority of important covariates, even though there are high levels of covariate correlation and/or measurement error variance. This observation is also in line with Bühlmann and Van de Geer (2011) and Tibshirani (2011) who both point towards the LASSO's reliability in terms of variable screening, also in situations where the conditions that are required for it to be consistent in variable selection are violated.

Finally, we also examined the distributions of the tuning parameter constituted by 500 runs of the simulation performed for each unique parameter combination. We mainly focused on how the magnitude of λ is influenced if the values chosen for the noise level ϵ are altered. We found that the general level of penalization measured by the median of the λ-distributions rises with increasing noise level. This observation is evident since a higher level of model noise implies that the LASSO has to penalize harder and thus must spend more λ in order to correct for the random error induced by the model noise. Besides the increase in the general penalization level, our results also indicate that there is more variation in tuning parameter selection for larger ϵ. The results suggest that a high degree of covariate correlation implies highly right-skewed and very peaked distributions of $\lambda_{\texttt{cv.glmnet}}$ and $\lambda_{\texttt{OptL1}}$. At the same time, our results imply that a higher degree of covariate correlation generally leads the LASSO to apply a smaller amount of shrinkage. In any case, the considered distributions of the tuning parameter are far from being Gaussian while the distributions of $\lambda_{\texttt{cv.glmnet}}$ are less skewed and exhibit flatter peaks than the distributions obtained by means of `OptL1`.

Overall, the above presented simulation results support the theory on the LASSO outlined in section 3 and section 5. However, due to their lack of robustness the use of empirical averages in our MC simulation study was inappropriate. This lead to results which were not meaningful but highly influenced by a very small number of outliers. Hence, further studies are needed to appropriately examine the performance of a relia-

bility ratio-like measurement error correction factor for the naive LASSO. The question of how to determine an amount of shrinkage which simultaneously satisfies the requirements of estimation and variable selection consistency has not been addressed by the prevalent literature so far. We find that the above aspects are interesting topics for future research.

8 References

Bickel, P., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann Statist* **37** 1705-1732.

Boyd, S., Vandenberghe, L. (2004). Convex Optimization. Cambridge: Cambridge University Press.

Brent, R. P. (1973). Algorithms for Minimization without Derivatives. Englewood Cliffs: Prentice-Hall.

Bühlmann, P. and Van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Berlin, Heidelberg: Springer.

Bunea, F., Tsybakov, A.B. and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron J Statist* **1** 169–194.

Buonaccorsi, J.P. (2010). Measurement Error: Models, Methods and Applications. Boca Raton: Chapman & Hall/CRC.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C. (2006). Measurement Error in Nonlinear Models: A Modern Perspective. 2nd edn. Boca Raton: Chapman & Hall/CRC.

Chand, S. (2012). On tuning parameter selection of lasso-type methods- A Monte Carlo study. *The 9th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, Islamabad, Pakistan.

Cheng, C.-L. and Van Ness, J.W. (1999). Statistical Regression with Measurement Error. London: Arnold.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann Statist* **32** 407-451.

Fahrmeir, L., Künstler, R., Pigeot, I. and Tutz, G. (2009). Statistik - Der Weg zur Datenanalyse. 5th edn. Berlin, Heidelberg: Springer.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013). Regression: Models, Methods and Applications. Berlin: Springer.

Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *J Am Stat Assoc* **96** 1348-1360.

Friedman, J., Hastie, T. and Tibshirani R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33** 1-22.

Fuller, W. (1987). Measurement Error Models. New York: Wiley.

Genz, A., Bretz, F. (2009). Computation of Multivariate Normal and t Probabilities. *Lecture Notes in Statistics* **195**, Heidelberg: Springer.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2014). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-2.

Goeman, J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biom J* **52** 70-84.

Goeman, J., Meijer, R. and Chaturvedi, N. (2014). penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. R package version 0.9-45.

Hansen, B.E. The Risk of James-Stein and Lasso Shrinkage," *Econom Reviews*, forthcoming.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd edn. New York: Springer.

Heid, I.M., Küchenhoff, H., Miles, J., Kreienbrock, L. and Wichmann, H.E. (2004). Two dimensions of measurement error: Classical and Berkson error in residential radon exposure assessment. *J Expo Sci Environ Epidemiol* **14** 365–377.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12** 55-67.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Ann Statist* **28** 1356–1378.

Loh, P-L. and Wainwright, M.J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Ann Statist* **40** 1637–1664.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann Statist* **34** 1436-1462.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann Statist* **37** 246-270.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raskutti, G., Wainwright, M.J. and Yu, B. (2010). Restricted Eigenvalue Properties for Correlated Gaussian Designs. *J Machine Learning Research* **11** 2241-2259.

Riley, J. D. (1955). Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix. *Math Comp* **9** 96-101.

Rosenbaum, M. and Tsybakov, A.B. (2010). Sparse recovery under matrix uncertainty. *Ann Statist* **38** 2620–2651.

Rosenbaum, M. and Tsybakov , A.B. (2013). Improved matrix uncertainty selector. *IMS Collections* **9** 276-290.

Schneeweiß, H. and Mittag, H.-J. (1986). Lineare Modelle mit fehlerbehafteten Daten. Heidelberg, Wien, Wurzburg: Physica-Verlag.

She, Y. (2010). Sparse regression with exact clustering. *Electron J Statist* **4** 1055–1096.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw* **39** 1-13.

Sørensen, Ø., Frigessi, A. and Thoresen, M. (2014). Measurement error in LASSO: impact and likelihood bias correction. To appear in *Statistica Sinica*.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *J. Roy. Statist. Soc. Ser. B* **73** 273–282.

Van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron J Statist* **3** 1360-1392.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann Statist* **36** 1567–1594.

Zhang, T. (2009). Some Sharp Performance Bounds for Least Squares Regression with L1 Regularization. *Ann Statist* **37** 2109-2114.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J Machine Learning Research* **7** 2541–2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J Am Stat Assoc* **101** 1418–1429.

A Appendix

A.1 Some important results from matrix algebra

This part of the appendix gives a summary of basic definitions and results in matrix algebra which are applied throughout this work. We restrict ourselves to important definitions and theorems which we state without allowing for the corresponding proofs. All the following matrix algebra results are quoted from Fahrmeir et al. (2013, Chapter A).

Theorem A.10 Existence and Uniqueness of the Inverse
The inverse A^{-1} of a square $n \times n$-matrix A exists if and only if $rk(A) = n$, i.e., if A is regular. The inverse is unique, and A is called invertible, regular, or nonsingular. If no inverse of A exists, it is called singular.

Theorem A.22 Properties of the Eigenvalues
The eigenvalues λ_i of an $n \times n$-matrix A have (amongst others) the following properties:

- **A** is regular if and only if all eigenvalues are non-zero.
- The matrices A and A^T have the same eigenvalues.
- If λ is an eigenvalue of a regular matrix A, then $\frac{1}{\lambda}$ is an eigenvalue of A^{-1}.
- The eigenvalues of a diagonal matrix are the elements of the diagonal.
- The eigenvalues of an orthogonal matrix A are either 1 or -1.
- The eigenvalues of an idempotent matrix A are either 1 or 0.

Definition A.27 Definite Matrices
The quadratic form $x^T Ax$ and the matrix A are called:

- Positive definite (p.d.), if $x^T Ax > 0$ for all $x \neq 0$, notation:$A > 0$.
- Positive semidefinite (p.s.d.), if $x^T Ax \geq 0$ and $x^T Ax = 0$ for at least one $x \neq 0$.
- Nonnegative definite, if $x^T Ax$ and A are either p.d. or p.s.d., notation: $A \geq 0$.
- Negative definite (n.d.), if $-A$ is positive definite.
- Negative semidefinite (n.s.d.), if $-A$ is p.s.d.

- Indefinite in all other cases.

Theorem A.27 Criteria for Definite Matrices
Let A be a symmetric matrix with real eigenvalues $\lambda_1, ..., \lambda_n$. It then follows that A is:

- Positive definite, if and only if $\lambda_i > 0$ for $i = 1, ..., n$.
- Positive semidefinite, if and only if $\lambda_i \geq 0$ for $i = 1, ..., n$ and $\lambda_i = 0$ for at least one eigenvalue.
- Negative definite, if and only if $\lambda_i < 0$ for $i = 1, ..., n$.
- Negative semidefinite, if and only if $\lambda_i \leq 0$ for $i = 1, ..., n$ and at least one $\lambda_i = 0$.
- Indefinite, if and only if A has at least one positive and one negative eigenvalue.

Theorem A.28 Properties of Positive Definite Matrices
For any positive definite matrix A the following properties hold:

- A is regular and thus invertible.
- A^{-1} is positive definite.
- The diagonal elements $a_{ii}, i = 1, ..., n$ are positive, i.e., $a_{ii} > 0$.
- $tr(A) > 0$.
- If B is positive semidefinite, then $A + B$ is positive definite.

A.2 Simulation results

All results presented in the following represent averages over 500 MC simulations.

Note that the `cv.glmnet` LASSO estimates are denoted by $Lasso1, naive$, $Lasso1, TMEV$ and $Lasso1, EMEV$, respectively, while the `penalized` LASSO estimates are denoted by $Lasso2, naive$, $Lasso2, TMEV$ and $Lasso2, EMEV$.

A.2.1 Tables containing MSE, bias, TPF and FPF (n=100)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.06	0.06			10.04	0.01						
OLS_{TMEV}	1.33	0.00			6450.41	-0.05						
OLS_{EMEV}	216.91	0.72			260.04	0.11						
$Lasso_{1,naive}$	0.82	0.12	99.56 (0.15)	51.36 (1.11)	1.72	0.03	98.44 (0.28)	24.84 (4.55)	2.29	0.02	94.84 (0.49)	11.37 (4.32)
$Lasso_{1,TMEV}$	0.93	0.07	100 (0.00)	100 (0.00)	1295.65	0.02	100 (0.00)	100 (0.00)	416.86	0.02	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	54.55	0.76	100 (0.00)	100 (0.00)	110.05	0.11	100 (0.00)	100 (0.00)	2642.88	0.03	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.81	0.12	99.60 (0.14)	51.60 (1.08)	1.57	0.04	98.36 (0.29)	21.11 (2.58)	2.25	0.02	94.48 (0.50)	10.44 (3.61)
$Lasso_{2,TMEV}$	0.92	0.08	100 (0.00)	100 (0.00)	714.17	0.02	100 (0.00)	100 (0.00)	393.46	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	53.94	0.77	100 (0.00)	100 (0.00)	103.01	0.11	100 (0.00)	100 (0.00)	2479.52	0.03	100 (0.00)	100 (0.00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	1.21	0.22			7.67	0.05						
OLS_{TMEV}	132084.00	0.25			162999.00	0.10						
OLS_{EMEV}	2929.78	0.68			3742.98	-0.11						
$Lasso_{1,naive}$	1.15	0.23	99.76 (0.11)	79.48 (0.83)	2.85	0.06	94.08 (0.50)	24.43 (3.89)	3.91	0.02	81.32 (0.86)	18.36 (5.44)
$Lasso_{1,TMEV}$	67300.40	0.15	100 (0.00)	100 (0.00)	34243.30	0.07	100 (0.00)	100 (0.00)	67519.39	0.02	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	2889.55	0.68	100 (0.00)	100 (0.00)	5912.83	-0.11	100 (0.00)	100 (0.00)	1315.93	0.05	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	1.15	0.23	99.76 (0.11)	79.16 (0.82)	2.77	0.06	93.84 (0.51)	22.05 (2.56)	3.82	0.02	80.24 (0.86)	16.52 (3.47)
$Lasso_{2,TMEV}$	67327.10	0.17	100 (0.00)	100 (0.00)	31683.50	0.08	100 (0.00)	100 (0.00)	76993.97	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	2843.70	0.67	100 (0.00)	100 (0.00)	7044.70	-0.11	100 (0.00)	100 (0.00)	225.61	0.04	100 (0.00)	100 (0.00)
$\sigma_u^2 = 1$												
OLS_{naive}	2.39	0.41			5.53	0.09						
OLS_{TMEV}	3160.62	0.04			13481.62	0.14						
OLS_{EMEV}	3174680.00	-23.43			29.52	0.00						
$Lasso_{1,naive}$	2.39	0.41	99.80 (0.10)	98.96 (0.22)	4.54	0.09	83.96 (0.82)	64.45 (2.92)	5.15	0.03	52.92 (1.11)	33.44 (6.11)
$Lasso_{1,TMEV}$	3118.31	0.05	100 (0.00)	100 (0.00)	6703.77	0.13	100 (0.00)	100 (0.00)	18771.61	0.03	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	3128850.00	-23.26	100 (0.00)	100 (0.00)	28.51	0.00	100 (0.00)	100 (0.00)	1295.84	-0.11	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	2.39	0.41	99.76 (0.11)	98.88 (0.23)	4.51	0.09	82.24 (0.86)	61.13 (2.67)	5.16	0.02	36.4 (1.09)	22.69 (3.78)
$Lasso_{2,TMEV}$	3116.66	0.05	100 (0.00)	100 (0.00)	7808.80	0.13	100 (0.00)	100 (0.00)	2146.58	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	3170360.00	-23.41	100 (0.00)	100 (0.00)	26.54	0.00	100 (0.00)	100 (0.00)	903.24	-0.10	100 (0.00)	100 (0.00)

Table 23: MSE, bias, TPF and FPF for β_1 (n=100, ρ = 0.9, ϵ = 1)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.01	0.03			9.68	0.00						
OLS_{TMEV}	1.29	0.01			7295.33	-0.04						
OLS_{EMEV}	123.63	0.34			145.62	0.06						
$Lasso_{1,naive}$	0.66	0.09	89.88 (0.61)	49.80 (1.16)	1.13	0.03	73.36 (0.88)	20.74 (3.75)	1.47	0.01	56.44 (1.07)	10.86 (6.96)
$Lasso_{1,TMEV}$	0.80	0.07	100 (0.00)	100 (0.00)	257.96	0.02	100 (0.00)	100 (0.00)	263.74	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	55.68	0.39	100 (0.00)	100 (0.00)	37.75	0.05	100 (0.00)	100 (0.00)	343.56	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.62	0.10	89.52 (0.61)	46.24 (1.06)	1.02	0.03	72.28 (0.89)	18.24 (2.27)	1.26	0.01	55.32 (1.07)	8.59 (3.06)
$Lasso_{2,TMEV}$	0.74	0.08	100 (0.00)	100 (0.00)	209.53	0.02	100 (0.00)	100 (0.00)	243.25	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	46.21	0.38	100 (0.00)	100 (0.00)	34.35	0.05	100 (0.00)	100 (0.00)	283.05	0.01	100 (0.00)	100 (0.00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.70	0.11			5.82	0.02						
OLS_{TMEV}	5855.60	0.00			83064.30	0.06						
OLS_{EMEV}	720.75	0.34			10621.80	-0.06						
$Lasso_{1,naive}$	0.58	0.13	94.68 (0.49)	66.04 (1.02)	1.18	0.04	68.36 (0.95)	27.68 (3.72)	1.49	0.01	45.36 (1.03)	13.52 (6.81)
$Lasso_{1,TMEV}$	1927.45	-0.08	100 (0.00)	100 (0.00)	10256.10	0.04	100 (0.00)	100 (0.00)	43108.06	0.02	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	702.51	0.33	100 (0.00)	100 (0.00)	10649.90	-0.06	100 (0.00)	100 (0.00)	1809.57	0.03	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.57	0.13	94.48 (0.49)	64.12 (1.00)	1.13	0.04	67.64 (0.94)	25.51 (2.60)	1.38	0.01	44.48 (1.03)	11.80 (3.16)
$Lasso_{2,TMEV}$	1685.33	-0.07	100 (0.00)	100 (0.00)	8334.89	0.04	100 (0.00)	100 (0.00)	38505.26	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	691.09	0.33	100 (0.00)	100 (0.00)	11420.40	-0.06	100 (0.00)	100 (0.00)	6.02	0.02	100 (0.00)	100 (0.00)
$\sigma_u^2 = 1$												
OLS_{naive}	0.66	0.21			2.09	0.05						
OLS_{TMEV}	2072.25	-0.08			13326.84	0.07						
OLS_{EMEV}	801725.00	-11.78			7.64	0.00						
$Lasso_{1,naive}$	0.66	0.20	97.08 (0.38)	95.00 (0.46)	1.27	0.05	65.92 (1.05)	50.78 (3.26)	1.40	0.02	35.60 (1.10)	24.10 (6.39)
$Lasso_{1,TMEV}$	2060.59	-0.07	100 (0.00)	100 (0.00)	3815.19	0.08	100 (0.00)	100 (0.00)	5828.87	0.02	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	746017.00	-11.36	100 (0.00)	100 (0.00)	7.11	0.00	100 (0.00)	100 (0.00)	283.98	-0.05	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.66	0.20	97.2 (0.37)	94.8 (0.47)	1.25	0.04	64.00 (1.06)	47.74 (2.68)	1.36	0.01	27.96 (1.00)	19.20 (3.65)
$Lasso_{2,TMEV}$	2069.34	-0.07	100 (0.00)	100 (0.00)	5707.37	0.08	100 (0.00)	100 (0.00)	5553.56	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	745299.00	-11.35	100 (0.00)	100 (0.00)	6.79	0.00	100 (0.00)	100 (0.00)	274.13	-0.05	100 (0.00)	100 (0.00)

Table 24: MSE, bias, TPF and FPF for β_2 (n=100, ρ = 0.9, ϵ = 1)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.01	0.03			9.68	0.00						
OLS_{TMEV}	1.29	0.01			8066.53	-0.04						
OLS_{EMEV}	424.72	0.21			158.60	0.05						
$Lasso_{1,naive}$	0.52	0.07	72.08 (0.92)	44.28 (1.13)	0.82	0.02	51.68 (0.92)	17.49 (4.16)	1.17	0.01	41.92 (0.89)	9.24 (7.73)
$Lasso_{1,TMEV}$	0.61	0.06	100 (0.00)	100 (0.00)	152.09	0.02	100 (0.00)	100 (0.00)	253.21	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	93.98	0.26	100 (0.00)	100 (0.00)	35.72	0.04	100 (0.00)	100 (0.00)	384.86	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.48	0.08	70.92 (0.90)	41.28 (1.03)	0.70	0.02	50.04 (0.87)	14.81 (2.28)	0.95	0.01	39.48 (0.83)	6.90 (2.93)
$Lasso_{2,TMEV}$	0.56	0.06	100 (0.00)	100 (0.00)	55.42	0.02	100 (0.00)	100 (0.00)	243.22	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	94.95	0.26	100 (0.00)	100 (0.00)	32.58	0.04	100 (0.00)	100 (0.00)	380.22	0.01	100 (0.00)	100 (0.00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.84	0.09			5.88	0.02						
OLS_{TMEV}	76754.30	0.36			130568.00	0.09						
OLS_{EMEV}	427.14	0.26			43537.80	-0.04						
$Lasso_{1,naive}$	0.67	0.11	83.24 (0.77)	59.84 (1.07)	1.10	0.03	49.64 (0.96)	24.43 (3.886)	1.32	0.01	34.56 (0.86)	11.35 (6.62)
$Lasso_{1,TMEV}$	22890.90	0.23	100 (0.00)	100 (0.00)	6618.64	0.05	100 (0.00)	100 (0.00)	13554.06	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	404.59	0.25	100 (0.00)	100 (0.00)	28558.00	-0.04	100 (0.00)	100 (0.00)	2201.95	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.66	0.11	82.68 (0.77)	57.84 (1.04)	1.04	0.03	48.28 (0.94)	22.05 (2.56)	1.22	0.01	33.2 (0.84)	9.57 (2.93)
$Lasso_{2,TMEV}$	22624.20	0.25	100 (0.00)	100 (0.00)	6726.20	0.05	100 (0.00)	100 (0.00)	12911.92	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	403.09	0.25	100 (0.00)	100 (0.00)	28114.30	-0.04	100 (0.00)	100 (0.00)	6.22	0.02	100 (0.00)	100 (0.00)
$\sigma_u^2 = 1$												
OLS_{naive}	0.93	0.15			2.16	0.04						
OLS_{TMEV}	2409.62	-0.13			27793.11	0.04						
OLS_{EMEV}	470450.00	-9.02			5.48	0.00						
$Lasso_{1,naive}$	0.92	0.15	94.64 (0.51)	90.44 (0.62)	1.32	0.03	57.12 (1.04)	44.29 (3.20)	1.44	0.01	26.04 (0.97)	20.22 (6.86)
$Lasso_{1,TMEV}$	2269.63	-0.12	100 (0.00)	100 (0.00)	3085.89	0.06	100 (0.00)	100 (0.00)	2506.51	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	435774.00	-8.68	100 (0.00)	100 (0.00)	4.95	0.00	100 (0.00)	100 (0.00)	153.09	-0.04	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.92	0.15	94.24 (0.52)	89.80 (0.64)	1.30	0.03	55.52 (1.03)	41.56 (2.55)	1.41	0.01	22.00 (0.92)	16.30 (3.39)
$Lasso_{2,TMEV}$	2228.35	-0.12	100 (0.00)	100 (0.00)	2973.26	0.06	100 (0.00)	100 (0.00)	1644.63	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	434083.00	-8.67	100 (0.00)	100 (0.00)	4.70	0.00	100 (0.00)	100 (0.00)	141.36	-0.04	100 (0.00)	100 (0.00)

Table 25: MSE, bias, TPF and FPF for β_3 (n=100, ρ = 0.9, ϵ = 1)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.21	0.01			2.11	0.00						
OLS_{TMEV}	0.22	0.00			2.57	0.00						
OLS_{EMEV}	3662.08	1.37			11493.90	0.13						
$Lasso_{1,naive}$	0.16	0.04	100 (0.00)	51.72 (1.29)	0.36	0.01	100 (0.00)	23.80 (4.94)	0.69	0.01	99.96 (0.04)	12.92 (8.62)
$Lasso_{1,TMEV}$	0.16	0.03	100 (0.00)	100 (0.00)	0.35	0.01	100 (0.00)	100 (0.00)	2.08	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	3277.89	1.35	100 (0.00)	100 (0.00)	6928.36	0.13	100 (0.00)	100 (0.00)	38184.80	0.06	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.16	0.04	100 (0.00)	46.40 (1.14)	0.33	0.01	100 (0.00)	19.45 (2.70)	0.66	0.01	99.96 (0.045)	9.74 (3.48)
$Lasso_{2,TMEV}$	0.15	0.03	100 (0.00)	100 (0.00)	0.31	0.01	100 (0.00)	100 (0.00)	2.05	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	3294.78	1.35	100 (0.00)	100 (0.00)	11059.40	0.14	100 (0.00)	100 (0.00)	30814.00	0.06	100 (0.00)	100 (0.00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.30	0.06			2.49	0.01						
OLS_{TMEV}	0.44	-0.01			65576.80	-0.05						
OLS_{EMEV}	542.93	-0.24			264.35	-0.03						
$Lasso_{1,naive}$	0.26	0.07	100 (0.00)	62.16 (1.16)	0.72	0.02	100 (0.00)	27.96 (4.25)	0.93	0.01	100 (0.00)	14.21 (7.81)
$Lasso_{1,TMEV}$	0.31	0.00	100 (0.00)	100 (0.00)	17451.30	-0.04	100 (0.00)	100 (0.00)	3677.29	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	532.05	-0.23	100 (0.00)	100 (0.00)	240.67	-0.03	100 (0.00)	100 (0.00)	2740.81	0.06	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.25	0.08	100 (0.00)	59.16 (1.11)	0.69	0.02	100 (0.00)	24.74 (2.75)	0.86	0.01	100 (0.00)	10.82 (3.44)
$Lasso_{2,TMEV}$	0.29	0.00	100 (0.00)	100 (0.00)	12613.50	-0.03	100 (0.00)	100 (0.00)	3507.62	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	522.30	-0.24	100 (0.00)	100 (0.00)	239.31	-0.04	100 (0.00)	100 (0.00)	2687.71	0.05	100 (0.00)	100 (0.00)
$\sigma_u^2 = 1$												
OLS_{naive}	1.46	0.27			3.81	0.06						
OLS_{TMEV}	477460000.00	-57.90			64574.20	-0.15						
OLS_{EMEV}	8.61	-0.05			4074.59	0.25						
$Lasso_{1,naive}$	1.48	0.26	100 (0.00)	94.6 (0.50)	2.48	0.06	99.16 (0.20)	47.77 (3.63)	3.73	0.02	90.40 (0.66)	26.60 (5.85)
$Lasso_{1,TMEV}$	440394000.00	-55.60	100 (0.00)	100 (0.00)	36308.60	-0.12	100 (0.00)	100 (0.00)	13460.39	0.02	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	8.44	-0.04	100 (0.00)	100 (0.00)	1287.60	0.26	100 (0.00)	100 (0.00)	639.05	-0.12	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	1.48	0.26	100 (0.00)	94.32 (0.50)	2.48	0.05	99.00 (0.23)	42. 68 (2.83)	3.94	0.02	78.32 (0.88)	17.78 (3.59)
$Lasso_{2,TMEV}$	445852000.00	-55.95	100 (0.00)	100 (0.00)	35297.60	-0.13	100 (0.00)	100 (0.00)	4863.49	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	8.51	-0.05	100 (0.00)	100 (0.00)	1020.00	0.24	100 (0.00)	100 (0.00)	401.51	-0.09	100 (0.00)	100 (0.00)

Table 26: MSE, bias, TPF and FPF for β_1 (n=100, ρ = 0.5, ϵ = 1)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.20	0.01			2.04	0.00						
OLS_{TMEV}	0.21	0.00			2.48	0.00						
OLS_{EMEV}	1100.51	0.67			17277.40	0.05						
$Lasso_{1,naive}$	0.16	0.03	99.88 (0.08)	50.24 (1.31)	0.34	0.01	99.44 (0.17)	23.28 (5.32)	0.61	0.01	93.48 (0.55)	12.66 (10.28)
$Lasso_{1,TMEV}$	0.16	0.03	100 (0.00)	100 (0.00)	0.35	0.01	100 (0.00)	100 (0.00)	0.78	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	894.90	0.66	100 (0.00)	100 (0.00)	1093.94	0.06	100 (0.00)	100 (0.00)	2537.28	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.15	0.04	99.88 (0.08)	45.76 (1.19)	0.30	0.01	99.44 (0.17)	18.74 (2.77)	0.57	0.01	90.88 (0.64)	8.92 (3.17)
$Lasso_{2,TMEV}$	0.15	0.03	100 (0.00)	100 (0.00)	0.30	0.01	100 (0.00)	100 (0.00)	0.74	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	892.17	0.65	100 (0.00)	100 (0.00)	1025.83	0.06	100 (0.00)	100 (0.00)	2413.81	0.02	100 (0.00)	100 (0.00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.20	0.03			1.88	0.01						
OLS_{TMEV}	0.33	-0.01			36218.40	-0.03						
OLS_{EMEV}	138.09	-0.12			85.79	-0.02						
$Lasso_{1,naive}$	0.16	0.05	100 (0.00)	56.4 (1.25)	0.43	0.02	98.56 (0.26)	25.46 (4.93)	0.56	0.01	95.76 (0.44)	12.76 (8.09)
$Lasso_{1,TMEV}$	0.22	0.01	100 (0.00)	100 (0.00)	1642.52	0.01	100 (0.00)	100 (0.00)	1344.26	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	127.73	-0.12	100 (0.00)	100 (0.00)	59.51	-0.02	100 (0.00)	100 (0.00)	1134.39	0.03	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.16	0.05	100 (0.00)	52.56 (1.14)	0.40	0.02	98.48 (0.27)	21.30 (2.76)	0.51	0.01	95.72 (0.45)	9.50 (3.48)
$Lasso_{2,TMEV}$	0.21	0.02	100 (0.00)	100 (0.00)	561.49	0.01	100 (0.00)	100 (0.00)	1232.52	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	128.07	-0.12	100 (0.00)	100 (0.00)	63.01	-0.02	100 (0.00)	100 (0.00)	951.13	0.03	100 (0.00)	100 (0.00)
$\sigma_u^2 = 1$												
OLS_{naive}	0.41	0.13			1.46	0.03						
OLS_{TMEV}	189011000.00	-36.46			25881.80	-0.09						
OLS_{EMEV}	2.19	-0.02			320.58	0.13						
$Lasso_{1,naive}$	0.42	0.13	99.64 (0.13)	88.6 (0.70)	0.76	0.03	94.76 (0.50)	40.10 (3.61)	1.11	0.01	71.76 (0.99)	21.72 (7.22)
$Lasso_{1,TMEV}$	183940000.00	-36.00	100 (0.00)	100 (0.00)	11302.70	-0.07	100 (0.00)	100 (0.00)	2589.74	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	2.12	-0.02	100 (0.00)	100 (0.00)	1327.36	0.14	100 (0.00)	100 (0.00)	146.19	-0.05	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.42	0.13	99.68 (0.13)	88.0 (0.70)	0.75	0.03	93.84 (0.53)	35.79 (2.74)	1.11	0.01	61.08 (1.02)	15.39 (3.71)
$Lasso_{2,TMEV}$	131874000.00	-30.42	100 (0.00)	100 (0.00)	9946.66	-0.07	100 (0.00)	100 (0.00)	3504.93	0.00	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	2.13	-0.02	100 (0.00)	100 (0.00)	991.04	0.11	100 (0.00)	100 (0.00)	73.66	-0.04	100 (0.00)	100 (0.00)

Table 27: MSE, bias, TPF and FPF for β_2 (n=100, ρ = 0.5, ϵ = 1)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.20	0.00			2.04	0.00						
OLS_{TMEV}	0.21	0.00			2.48	0.00						
OLS_{EMEV}	546.77	0.50			9906.70	0.05						
$Lasso_{1,naive}$	0.12	0.03	82.60 (0.77)	43.52 (1.26)	0.23	0.01	67.72 (0.80)	19.35 (5.11)	0.34	0.00	60.40 (0.67)	9.68 (9.24)
$Lasso_{1,TMEV}$	0.12	0.03	100 (0.00)	100 (0.00)	0.24	0.01	100 (0.00)	100 (0.00)	0.71	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	484.68	0.48	100 (0.00)	100 (0.00)	10332.30	0.06	100 (0.00)	100 (0.00)	19998.80	0.03	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.11	0.03	81.76 (0.75)	40.64 (1.16)	0.20	0.01	66.20 (0.78)	15.43 (2.52)	0.29	0.00	58.16 (0.62)	6.71 (3.25)
$Lasso_{2,TMEV}$	0.12	0.03	100 (0.00)	100 (0.00)	0.20	0.01	100 (0.00)	100 (0.00)	0.66	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	484.42	0.48	100 (0.00)	100 (0.00)	12370.60	0.06	100 (0.00)	100 (0.00)	18173.40	0.03	100 (0.00)	100 (0.00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.22	0.02			1.89	0.00						
OLS_{TMEV}	0.33	-0.01			76129.60	-0.07						
OLS_{EMEV}	87.50	-0.10			71.77	-0.01						
$Lasso_{1,naive}$	0.16	0.04	88.8 (0.64)	51.76 (1.27)	0.30	0.01	69.80 (0.78)	19.81 (4.38)	0.39	0.01	57.96 (0.80)	10.48 (8.60)
$Lasso_{1,TMEV}$	0.19	0.01	100 (0.00)	100 (0.00)	35195.40	-0.06	100 (0.00)	100 (0.00)	1466.48	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	81.75	-0.10	100 (0.00)	100 (0.00)	56.57	-0.01	100 (0.00)	100 (0.00)	29426.40	0.03	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.15	0.04	88.24 (0.64)	47.16 (1.15)	0.28	0.01	68.44 (0.76)	16.42 (2.56)	0.34	0.00	56.56 (0.78)	7.58 (3.35)
$Lasso_{2,TMEV}$	0.17	0.02	100 (0.00)	100 (0.00)	37782.50	-0.06	100 (0.00)	100 (0.00)	1508.91	0.00		100 (0.00)
$Lasso_{2,EMEV}$	80.00	-0.10	100 (0.00)	100 (0.00)	56.48	-0.01	100 (0.00)	100 (0.00)	28973.70	0.03		100 (0.00)
$\sigma_u^2 = 1$												
OLS_{naive}	0.56	0.10			1.52	0.02						
OLS_{TMEV}	70245900.00	-22.19			23230.00	-0.10						
OLS_{EMEV}	2.01	-0.02			435.45	0.11						
$Lasso_{1,naive}$	0.56	0.10	94.04 (0.49)	82.44 (0.86)	0.75	0.02	74.40 (0.87)	34.32 (4.21)	1.04	0.01	51.00 (0.78)	17.50 (7.37)
$Lasso_{1,TMEV}$	67750200.00	-21.80	100 (0.00)	100 (0.00)	7469.23	-0.06	100 (0.00)	100 (0.00)	6933.58	0.00	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	1.93	-0.02	100 (0.00)	100 (0.00)	570.64	0.10	100 (0.00)	100 (0.00)	72.36	-0.04	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.56	0.10	93.80 (0.50)	81.44 (0.85)	0.74	0.02	72.24 (0.92)	29.45 (2.68)	1.06	0.01	47.08 (0.78)	12.92 (3.61)
$Lasso_{2,TMEV}$	70240800.00	-22.21	100 (0.00)	100 (0.00)	6418.93	-0.05	100 (0.00)	100 (0.00)	4433.17	0.00	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	1.95	-0.02	100 (0.00)	100 (0.00)	273.52	0.08	100 (0.00)	100 (0.00)	69.11	-0.04	100 (0.00)	100 (0.00)

Table 28: MSE, bias, TPF and FPF for β_3 (n=100, ρ = 0.5, ϵ = 1)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.12	0.00			1.20	0.00						
OLS_{TMEV}	0.13	0.00			1.33	0.00						
OLS_{EMEV}	4516.21	-0.25			11309.60	0.24						
$Lasso_{1,naive}$	0.11	0.01	100.0 (00)	56.24 (1.46)	0.23	0.01	100.0 (00)	23.02 (5.80)	0.33	0.00	100.0 (00)	11.74 (10.03)
$Lasso_{1,TMEV}$	0.11	0.01	100.0 (00)	100.0 (00)	0.21	0.00	100.0 (00)	100.0 (00)	1.68	0.01	100.0 (00)	100.0 (00)
$Lasso_{1,EMEV}$	4269.92	-0.25	100.0 (00)	100.0 (00)	9590.31	0.23	100.0 (00)	100.0 (00)	8819.94	0.03	100.0 (00)	100.0 (00)
$Lasso_{2,naive}$	0.11	0.01	100.0 (00)	52.28 (1.37)	0.22	0.01	100.0 (00)	15.64 (2.79)	0.32	0.00	100.0 (00)	6.36 (3.17)
$Lasso_{2,TMEV}$	0.11	0.01	100.0 (00)	100.0 (00)	0.20	0.00	100.0 (00)	100.0 (00)	1.67	0.01	100.0 (00)	100.0 (00)
$Lasso_{2,EMEV}$	4067.98	-0.27	100.0 (00)	100.0 (00)	9477.97	0.22	100.0 (00)	100.0 (00)	7998.17	0.03	100.0 (00)	100.0 (00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.22	0.04			1.54	0.01						
OLS_{TMEV}	0.25	-0.01			32775.00	-0.01						
OLS_{EMEV}	39.87	-0.04			1424.33	0.04						
$Lasso_{1,naive}$	0.23	0.04	100.0 (00)	59.72 (1.34)	0.37	0.01	100.0 (00)	22.94 (5.63)	0.76	0.01	100.0 (00)	14.78 (11.75)
$Lasso_{1,TMEV}$	0.19	0.00	100.0 (00)	100.0 (00)	4434.29	0.01	100.0 (00)	100.0 (00)	105.47	0.02	100.0 (00)	100.0 (00)
$Lasso_{1,EMEV}$	35.00	-0.04	100.0 (00)	100.0 (00)	1274.84	0.04	100.0 (00)	100.0 (00)	847.31	0.03	100.0 (00)	100.0 (00)
$Lasso_{2,naive}$	0.23	0.04	100.0 (00)	55.48 (1.25)	0.36	0.01	100.0 (00)	16.50 (2.74)	0.75	0.00	100.0 (00)	7.92 (3.21)
$Lasso_{2,TMEV}$	0.18	0.00	100.0 (00)	100.0 (00)	4625.53	0.00	100.0 (00)	100.0 (00)	1131.05	0.02	100.0 (00)	100.0 (00)
$Lasso_{2,EMEV}$	34.98	-0.04	100.0 (00)	100.0 (00)	1312.80	0.04	100.0 (00)	100.0 (00)	566.52	0.03	100.0 (00)	100.0 (00)
$\sigma_u^2 = 1$												
OLS_{naive}	1.25	0.09			2.96	0.03						
OLS_{TMEV}	2992.95	-0.22			101118.00	-0.05						
OLS_{EMEV}	8.98	0.00			24.49	0.00						
$Lasso_{1,naive}$	1.41	0.08	99.8 (0.10)	73.0 (1.25)	2.11	0.02	99.56 (0.16)	31.86 (5.17)	2.88	0.01	96.6 (0.41)	19.36 (11.06)
$Lasso_{1,TMEV}$	2744.61	-0.22	100.0 (00)	100.0 (00)	67969.60	-0.07	100.0 (00)	100.0 (00)	21790.20	-0.05	100.0 (00)	100.0 (00)
$Lasso_{1,EMEV}$	9.26	0.00	100.0 (00)	100.0 (00)	31.36	0.00	100.0 (00)	100.0 (00)	98.84	0.00	100.0 (00)	100.0 (00)
$Lasso_{2,naive}$	1.45	0.07	99.64 (0.13)	68.08 (1.25)	2.26	0.02	98.56 (0.30)	21.45 (2.95)	3.12	0.01	90.52 (0.69)	9.15 (3.70)
$Lasso_{2,TMEV}$	2594.29	-0.20	100.0 (00)	100.0 (00)	34666.40	-0.04	100.0 (00)	100.0 (00)	10189.50	-0.03	100.0 (00)	100.0 (00)
$Lasso_{2,EMEV}$	9.10	0.00	100.0 (00)	100.0 (00)	24.25	0.00	100.0 (00)	100.0 (00)	17.38	0.00	100.0 (00)	100.0 (00)

Table 29: MSE, bias, TPF and FPF for β_1 (n=100, ρ = 0.1, ϵ = 1)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.12	0.00			1.15	0.00						
OLS_{TMEV}	0.12	0.00			1.28	0.00						
OLS_{EMEV}	1760.36	-0.05			2925.08	0.12						
$Lasso_{1,naive}$	0.11	0.01	100.0 (00)	55.52 (1.44)	0.22	0.01	99.92 (0.06)	23.32 (6.07)	0.32	0.00	99.40 (0.17)	11.86 (10.47)
$Lasso_{1,TMEV}$	0.11	0.01	100.0 (00)	100.0 (00)	0.22	0.00	100.0 (00)	100.0 (00)	0.57	0.01	100.0 (00)	100.0 (00)
$Lasso_{1,EMEV}$	1534.19	-0.04	100.0 (00)	100.0 (00)	2256.59	0.11	100.0 (00)	100.0 (00)	1964.05	0.01	100.0 (00)	100.0 (00)
$Lasso_{2,naive}$	0.11	0.01	100.0 (00)	50.72 (1.37)	0.21	0.01	99.88 (0.08)	15.28 (2.79)	0.30	0.00	98.96 (0.22)	6.18 (3.07)
$Lasso_{2,TMEV}$	0.11	0.01	100.0 (00)	100.0 (00)	0.20	0.00	100.0 (00)	100.0 (00)	0.56	0.01	100.0 (00)	100.0 (00)
$Lasso_{2,EMEV}$	1500.20	-0.06	100.0 (00)	100.0 (00)	2125.01	0.10	100.0 (00)	100.0 (00)	1602.02	0.01	100.0 (00)	100.0 (00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.14	0.02			1.17	0.00						
OLS_{TMEV}	0.18	0.00			7609.42	0.00						
OLS_{EMEV}	10.10	-0.02			207.97	0.02						
$Lasso_{1,naive}$	0.14	0.03	100.0 (00)	56.84 (1.36)	0.24	0.01	100.0 (00)	22.44 (5.63)	0.48	0.00	98.2 (0.31)	13.86 (11.41)
$Lasso_{1,TMEV}$	0.14	0.01	100.0 (00)	100.0 (00)	771.92	0.01	100.0 (00)	100.0 (00)	177.87	0.01	100.0 (00)	100.0 (00)
$Lasso_{1,EMEV}$	8.20	-0.02	100.0 (00)	100.0 (00)	249.79	0.02	100.0 (00)	100.0 (00)	193.80	0.01	100.0 (00)	100.0 (00)
$Lasso_{2,naive}$	0.14	0.03	100.0 (00)	51.88 (1.24)	0.23	0.01	100.0 (00)	15.32 (2.59)	0.47	0.00	96.76 (0.38)	6.97 (3.32)
$Lasso_{2,TMEV}$	0.14	0.01	100.0 (00)	100.0 (00)	948.99	0.01	100.0 (00)	100.0 (00)	210.52	0.01	100.0 (00)	100.0 (00)
$Lasso_{2,EMEV}$	8.17	-0.02	100.0 (00)	100.0 (00)	189.35	0.02	100.0 (00)	100.0 (00)	128.39	0.01	100.0 (00)	100.0 (00)
$\sigma_u^2 = 1$												
OLS_{naive}	0.36	0.05			1.14	0.02						
OLS_{TMEV}	717.23	-0.12			30241.80	-0.04						
OLS_{EMEV}	1.78	0.00			5.81	0.00						
$Lasso_{1,naive}$	0.42	0.04	97.92 (0.32)	64.6 (1.37)	0.67	0.01	93.92 (0.56)	26.92 (5.66)	0.90	0.01	83.64 (0.94)	15.61 (11.27)
$Lasso_{1,TMEV}$	438.00	-0.07	100.0 (00)	100.0 (00)	14602.30	-0.04	100.0 (00)	100.0 (00)	13531.90	-0.04	99.40 (0.39)	99.40 (11.21)
$Lasso_{1,EMEV}$	1.86	0.00	100.0 (00)	100.0 (00)	8.55	0.00	100.0 (00)	100.0 (00)	24.79	0.00	99.40 (0.39)	99.40 (11.21)
$Lasso_{2,naive}$	0.44	0.04	97.48 (0.34)	58.92 (1.37)	0.72	0.01	88.64 (0.82)	3.21 (16.66)	0.94	0.00	69.84 (1.20)	6.88 (4.17)
$Lasso_{2,TMEV}$	520.86	-0.10	100.0 (00)	100.0 (00)	5723.56	-0.02	99.6 (0.32)	99.6 (2.84)	7009.80	-0.03	98.20 (0.67)	98.20 (19.30)
$Lasso_{2,EMEV}$	1.84	0.00	100.0 (00)	100.0 (00)	6.45	0.00	99.6 (0.32)	99.6 (2.84)	2.77	0.00	98.20 (0.67)	98.20 (19.30)

Table 30: MSE, bias, TPF and FPF for β_2 (n=100, ρ = 0.1, ϵ = 1)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.12	0.00			1.15	0.00						
OLS_{TMEV}	0.12	0.00			1.28	0.00						
OLS_{EMEV}	5461.30	0.33			4331.30	0.10						
$Lasso_{1,naive}$	0.09	0.01	80.6 (0.87)	44.4 (1.49)	0.16	0.00	67.44 (0.83)	16.53 (5.76)	0.21	0.00	64.16 (0.77)	8.48 (9.92)
$Lasso_{1,TMEV}$	0.09	0.01	100.0 (00)	100.0 (00)	0.15	0.00	100.0 (00)	100.0 (00)	0.57	0.00	100.0 (00)	100.0 (00)
$Lasso_{1,EMEV}$	4546.07	0.30	100.0 (00)	100.0 (00)	3658.34	0.08	100.0 (00)	100.0 (00)	2194.10	0.01	100.0 (00)	100.0 (00)
$Lasso_{2,naive}$	0.09	0.01	78.76 (0.89)	38.48 (1.30)	0.14	0.00	64.32 (0.75)	11.25 (3.04)	0.19	0.00	60.68 (0.68)	4.46 (3.28)
$Lasso_{2,TMEV}$	0.09	0.01	100.0 (00)	100.0 (00)	0.14	0.00	100.0 (00)	100.0 (00)	0.55	0.00	100.0 (00)	100.0 (00)
$Lasso_{2,EMEV}$	4555.08	0.31	100.0 (00)	100.0 (00)	3405.47	0.08	100.0 (00)	100.0 (00)	2041.64	0.01	100.0 (00)	100.0 (00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.15	0.01			1.18	0.00						
OLS_{TMEV}	0.18	0.00			5718.10	-0.01						
OLS_{EMEV}	7.63	0.00			24.82	0.01						
$Lasso_{1,naive}$	0.13	0.01	84.24 (0.76)	46.96 (1.43)	0.20	0.00	70.72 (0.82)	17.72 (5.48)	0.38	0.00	59.52 (0.73)	10.23 (10.80)
$Lasso_{1,TMEV}$	0.12	0.01	100.0 (00)	100.0 (00)	381.95	0.01	100.0 (00)	100.0 (00)	1532.49	-0.01	100.0 (00)	100.0 (00)
$Lasso_{1,EMEV}$	6.04	0.00	100.0 (00)	100.0 (00)	65.41	0.01	100.0 (00)	100.0 (00)	256.00	0.01	100.0 (00)	100.0 (00)
$Lasso_{2,naive}$	0.13	0.02	82.72 (0.75)	40.44 (1.27)	0.19	0.00	67.60 (0.78)	11.80 (2.75)	0.37	0.00	55.12 (0.69)	5.13 (3.46)
$Lasso_{2,TMEV}$	0.12	0.01	100.0 (00)	100.0 (00)	346.39	0.01	100.0 (00)	100.0 (00)	1095.80	-0.01	100.0 (00)	100.0 (00)
$Lasso_{2,EMEV}$	6.08	0.00	100.0 (00)	100.0 (00)	62.16	0.01	100.0 (00)	100.0 (00)	159.83	0.01	100.0 (00)	100.0 (00)
$\sigma_u^2 = 1$												
OLS_{naive}	0.42	0.03			1.19	0.01						
OLS_{TMEV}	725.67	-0.21			12193.50	-0.02						
OLS_{EMEV}	2.93	0.00			7.82	0.00						
$Lasso_{1,naive}$	0.49	0.02	77.92 (1.03)	44.52 (1.53)	0.63	0.01	61.24 (1.02)	18.40 (5.59)	0.85	0.00	55.8 (0.88)	11.86 (12.03)
$Lasso_{1,TMEV}$	281.20	-0.12	100.0 (00)	100.0 (00)	4162.58	-0.01	100.0 (00)	100.0 (00)	32888.60	-0.07	99.00 (0.50)	99.00 (14.44)
$Lasso_{1,EMEV}$	3.02	0.00	100.0 (00)	100.0 (00)	12.48	0.00	100.0 (00)	100.0 (00)	30.43	0.00	99.00 (0.50)	99.00 (14.44)
$Lasso_{2,naive}$	0.51	0.02	75.16 (1.06)	33.84 (1.43)	0.65	0.00	56.52 (0.95)	11.95 (3.21)	0.88	0.00	48.88 (0.89)	5.58 (4.09)
$Lasso_{2,TMEV}$	283.29	-0.14	100.0 (00)	100.0 (00)	5778.71	0.00	100.0 (00)	100.0 (00)	17615.30	-0.05	97.40 (0.80)	97.40 (23.1)
$Lasso_{2,EMEV}$	2.98	0.00	100.0 (00)	100.0 (00)	12.45	0.00	100.0 (00)	100.0 (00)	1.78	0.00	97.40 (0.80)	97.40 (23.1)

Table 31: MSE, bias, TPF and FPF for β_3 (n=100, ρ = 0.1, ϵ = 1)

	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.32	0.01			12.19	0.00						
OLS_{TMEV}	1.40	-0.01			14.36	0.00						
OLS_{EMEV}	12138.10	0.82			2409170.00	-0.45						
$Lasso_{1,naive}$	1.00	0.08	99.44 (0.18)	50.64 (1.36)	1.99	0.03	95.60 (0.46)	21.97 (5.04)	2.69	0.01	91.72 (0.60)	11.02 (8.37)
$Lasso_{1,TMEV}$	1.02	0.07	100 (0.00)	100 (0.00)	2.07	0.03	100 (0.00)	100 (0.00)	3.35	0.02	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	11120.30	0.80	100 (0.00)	100 (0.00)	1270120.00	-0.31	100 (0.00)	100 (0.00)	81442.57	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.94	0.09	99.36 (0.19)	46.44 (1.24)	1.78	0.03	95.36 (0.48)	17.38 (2.74)	2.41	0.01	90.60 (0.64)	8.27 (3.14)
$Lasso_{2,TMEV}$	0.96	0.08	100 (0.00)	100 (0.00)	1.83	0.03	100 (0.00)	100 (0.00)	3.12	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	6009.30	0.91	100 (0.00)	100 (0.00)	1306990.00	-0.32	100 (0.00)	100 (0.00)	78185.12	0.01	100 (0.00)	100 (0.00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	1.29	0.10			11.46	0.02						
OLS_{TMEV}	1.94	0.00			23121.50	-0.04						
OLS_{EMEV}	1535.01	0.05			33397.30	1.32						
$Lasso_{1,naive}$	1.12	0.13	99.20 (0.20)	59.84 (1.25)	2.79	0.05	91.28 (0.59)	27.74 (4.76)	2.95	0.02	89.52 (0.65)	12.32 (8.49)
$Lasso_{1,TMEV}$	1.41	0.05	100 (0.00)	100 (0.00)	2300.03	0.03	100 (0.00)	100 (0.00)	2625.56	0.02	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	1668.48	0.03	100 (0.00)	100 (0.00)	30061.50	1.27	100 (0.00)	100 (0.00)	152.79	0.06	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	1.08	0.14	99.28 (0.19)	55.04 (1.15)	2.62	0.05	90.48 (0.61)	23.59 (2.55)	2.63	0.02	88.52 (0.68)	9.19 (3.08)
$Lasso_{2,TMEV}$	1.30	0.06	100 (0.00)	100 (0.00)	1635.92	0.03	100 (0.00)	100 (0.00)	2243.34	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	1794.98	0.03	100 (0.00)	100 (0.00)	29329.60	1.25	100 (0.00)	100 (0.00)	135.39	0.05	100 (0.00)	100 (0.00)
$\sigma_u^2 = 1$												
OLS_{naive}	1.98	0.29			7.46	0.06						
OLS_{TMEV}	4260.46	0.00			10459800.00	-0.03						
OLS_{EMEV}	647.63	0.27			13.66	-0.02						
$Lasso_{1,naive}$	2.00	0.29	98.76 (0.24)	87.56 (0.74)	3.53	0.06	89.00 (0.66)	38.92 (3.73)	4.72	0.03	67.64 (1.01)	20.50 (7.27)
$Lasso_{1,TMEV}$	4103.93	0.02	100 (0.00)	100 (0.00)	605100.00	0.07	100 (0.00)	100 (0.00)	8315.12	0.02	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	637.54	0.27	100 (0.00)	100 (0.00)	10.88	-0.02	100 (0.00)	100 (0.00)	35247.95	0.79	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	2.00	0.29	98.52 (0.26)	86.56 (0.77)	3.46	0.06	87.32 (0.69)	34.30 (2.73)	4.55	0.02	60.20 (1.10)	14.95 (3.55)
$Lasso_{2,TMEV}$	4090.72	0.03	100 (0.00)	100 (0.00)	659905.00	0.04	100 (0.00)	100 (0.00)	8196.30	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	642.74	0.27	100 (0.00)	100 (0.00)	10.31	-0.02	100 (0.00)	100 (0.00)	23505.36	0.64	100 (0.00)	100 (0.00)

Table 32: MSE, bias, TPF and FPF for β_1 (n=100, ρ = 0.5, ϵ = 2.5)

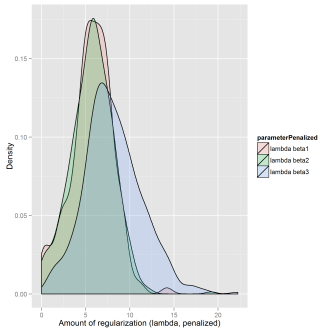
	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.32	0.00			12.10	0.00						
OLS_{TMEV}	1.40	0.00			14.25	0.00						
OLS_{EMEV}	5941.39	0.30			836879.00	-0.27						
$Lasso_{1,naive}$	0.80	0.07	86.64 (0.71)	45.68 (1.31)	1.29	0.02	66.12 (1.01)	18.31 (4.20)	1.66	0.01	51.56 (1.02)	9.40 (8.96)
$Lasso_{1,TMEV}$	0.83	0.07	100 (0.00)	100 (0.00)	1.36	0.02	100 (0.00)	100 (0.00)	1.55	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	3485.15	0.33	100 (0.00)	100 (0.00)	178228.00	-0.10	100 (0.00)	100 (0.00)	14557.01	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.73	0.08	86.04 (0.71)	41.00 (1.18)	1.12	0.03	63.92 (1.01)	14.69 (2.42)	1.33	0.01	48.28 (1.00)	6.71 (3.02)
$Lasso_{2,TMEV}$	0.75	0.08	100 (0.00)	100 (0.00)	1.17	0.02	100 (0.00)	100 (0.00)	1.29	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	1426.41	0.42	100 (0.00)	100 (0.00)	192647.00	-0.11	100 (0.00)	100 (0.00)	12910.46	0.00	100 (0.00)	100 (0.00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	1.14	0.05			10.77	0.01						
OLS_{TMEV}	1.82	0.01			28252.90	-0.03						
OLS_{EMEV}	588.05	0.01			8201.38	0.64						
$Lasso_{1,naive}$	0.76	0.10	85.56 (0.76)	48.64 (1.17)	1.44	0.03	56.24 (1.06)	21.19 (4.51)	1.60	0.01	50.88 (1.03)	9.87 (8.82)
$Lasso_{1,TMEV}$	1.02	0.07	100 (0.00)	100 (0.00)	207.43	0.02	100 (0.00)	100 (0.00)	2389.30	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	578.55	-0.02	100 (0.00)	100 (0.00)	6482.06	0.59	100 (0.00)	100 (0.00)	49.47	0.03	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.74	0.10	85.48 (0.73)	47.12 (1.10)	1.27	0.03	53.40 (1.04)	17.63 (2.40)	1.29	0.01	47.64 (1.02)	7.02 (2.98)
$Lasso_{2,TMEV}$	0.99	0.07	100 (0.00)	100 (0.00)	168.16	0.02	100 (0.00)	100 (0.00)	2618.79	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	837.00	-0.03	100 (0.00)	100 (0.00)	4861.15	0.52	100 (0.00)	100 (0.00)	40.50	0.02	100 (0.00)	100 (0.00)
$\sigma_u^2 = 1$												
OLS_{naive}	0.84	0.15			5.06	0.03						
OLS_{TMEV}	2186.81	-0.10			9254600.00	-0.06						
OLS_{EMEV}	172.96	0.14			5.48	-0.01						
$Lasso_{1,naive}$	0.75	0.14	85.24 (0.79)	68.64 (1.09)	1.25	0.03	58.76 (1.07)	27.84 (3.96)	1.51	0.01	34.04 (1.04)	13.67 (6.90)
$Lasso_{1,TMEV}$	1714.15	-0.07	100 (0.00)	100 (0.00)	192051.00	0.04	100 (0.00)	100 (0.00)	4536.52	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	159.20	0.14	100 (0.00)	100 (0.00)	3.01	-0.01	100 (0.00)	100 (0.00)	8431.41	0.38	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.75	0.14	84.08 (0.83)	66.68 (1.08)	1.18	0.03	54.20 (1.07)	22.69 (2.86)	1.39	0.01	30.08 (0.97)	10.28 (3.42)
$Lasso_{2,TMEV}$	1548.32	-0.02	100 (0.00)	100 (0.00)	113626.00	0.02	99.8 (0.22)	99.80 (2.01)	3530.41	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	112.32	0.11	100 (0.00)	100 (0.00)	2.92	-0.01	99.8 (0.22)	99.80 (2.01)	7013.98	0.34	100 (0.00)	100 (0.00)

Table 33: MSE, bias, TPF and FPF for β_2 (n=100, ρ = 0.5, ϵ = 2.5)

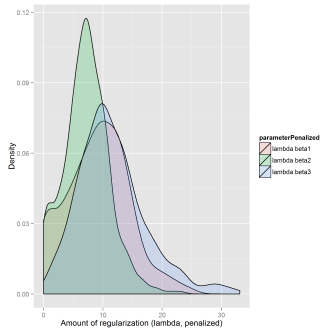
	$p = 10$				$p = 50$				$p = 150$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.31	0.00			12.11	0.00						
OLS_{TMEV}	1.39	0.00			14.26	0.00						
OLS_{EMEV}	3986.47	0.23			42066.20	-0.02						
$Lasso_{1,naive}$	0.55	0.05	64.76 (1.00)	36.44 (1.26)	1.02	0.02	46.24 (0.91)	15.08 (4.41)	1.32	0.01	36.52 (0.86)	7.56 (7.78)
$Lasso_{1,TMEV}$	0.57	0.05	100 (0.00)	100 (0.00)	1.05	0.02	99.80 (0.22)	99.80 (2.01)	1.34	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	202.82	0.40	100 (0.00)	100 (0.00)	226949.00	0.20	99.80 (0.22)	99.80 (2.01)	12846.23	0.00	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.49	0.06	62.40 (0.95)	33.08 (1.12)	0.88	0.02	43.64 (0.85)	11.72 (2.43)	1.10	0.01	34.36 (0.81)	5.46 (3.09)
$Lasso_{2,TMEV}$	0.50	0.05	100 (0.00)	100 (0.00)	0.90	0.02	99.80 (0.22)	99.80 (2.01)	1.17	0.01	100 (0.00)	100 (0.00)
$Lasso_{2,EMEV}$	331.38	0.38	100 (0.00)	100 (0.00)	238399.00	0.20	99.80 (0.22)	99.80 (2.01)	10694.51	0.00	100 (0.00)	100 (0.00)
$\sigma_u^2 = 0.1$												
OLS_{naive}	1.14	0.04			10.78	0.01						
OLS_{TMEV}	1.82	0.01			28066.70	-0.05						
OLS_{EMEV}	900.00	-0.02			4952.02	0.50						
$Lasso_{1,naive}$	0.61	0.07	66.60 (1.04)	41.56 (1.24)	1.13	0.02	41.28 (0.95)	17.52 (4.78)	1.14	0.01	37.32 (0.83)	7.76 (7.83)
$Lasso_{1,TMEV}$	0.78	0.05	100 (0.00)	100 (0.00)	310.33	0.01	100 (0.00)	100 (0.00)	770.41	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	961.73	-0.02	100 (0.00)	100 (0.00)	3786.58	0.46	100 (0.00)	100 (0.00)	27.45	0.02	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.57	0.07	64.24 (0.97)	36.76 (1.09)	0.97	0.02	38.44 (0.87)	13.92 (2.36)	0.94	0.01	35.40 (0.78)	5.38 (3.01)
$Lasso_{2,TMEV}$	0.70	0.06	100 (0.00)	100 (0.00)	278.91	0.01	99.80 (0.22)	99.80 (2.01)	947.64	0.01	99.60 (0.32)	99.60 (9.16)
$Lasso_{2,EMEV}$	837.07	-0.01	100 (0.00)	100 (0.00)	3377.47	0.43	99.80 (0.22)	99.80 (2.01)	24.78	0.02	99.60 (0.32)	99.60 (9.16)
$\sigma_u^2 = 1$												
OLS_{naive}	0.98	0.12			5.11	0.02						
OLS_{TMEV}	2666.22	0.09			8784010.00	-0.11						
OLS_{EMEV}	115.84	0.12			5.35	-0.01						
$Lasso_{1,naive}$	0.86	0.11	73.64 (0.96)	61.76 (1.13)	1.19	0.02	45.44 (1.09)	22.33 (4.24)	1.41	0.01	25.88 (0.90)	11.09 (7.48)
$Lasso_{1,TMEV}$	1841.29	0.17	100 (0.00)	100 (0.00)	204592.00	0.01	99.60 (0.32)	99.60 (2.84)	2826.18	0.01	100 (0.00)	100 (0.00)
$Lasso_{1,EMEV}$	69.52	0.09	100 (0.00)	100 (0.00)	2.67	-0.01	99.60 (0.32)	99.60 (2.84)	4204.56	0.26	100 (0.00)	100 (0.00)
$Lasso_{2,naive}$	0.85	0.11	72.04 (0.94)	59.00 (1.12)	1.15	0.02	41.04 (1.05)	17.66 (3.01)	1.33	0.01	21.68 (0.85)	7.67 (3.55)
$Lasso_{2,TMEV}$	2177.50	0.17	100 (0.00)	100 (0.00)	119330.00	0.01	98.8 (0.54)	98.80 (4.91)	1581.06	0.01	99.20 (0.45)	99.20 (12.93)
$Lasso_{2,EMEV}$	35.85	0.05	100 (0.00)	100 (0.00)	2.55	-0.01	98.8 (0.54)	98.80 (4.91)	2074.63	0.18	99.20 (0.45)	99.20 (12.93)

Table 34: MSE, bias, TPF and FPF for β_3 (n=100, ρ = 0.5, ϵ = 2.5)

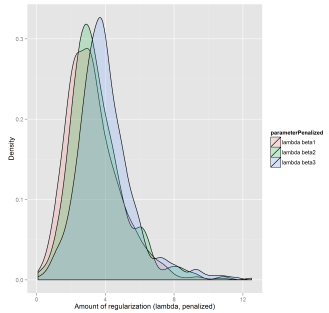
A.2.2 Density plots for the tuning parameter λ (n=100)



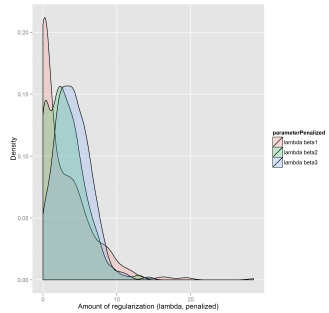
ϵ=1, ρ=0.1, σ_u^2 = 0.01



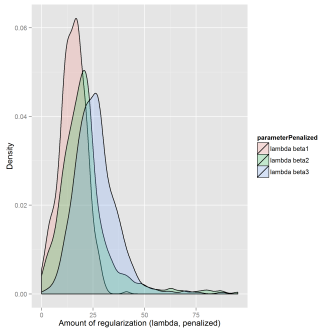
ϵ=1, ρ=0.1, σ_u^2 = 1



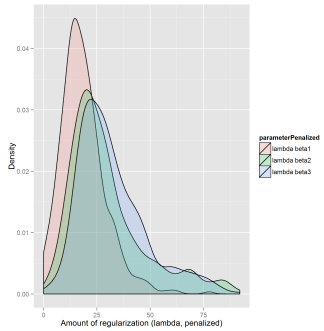
ϵ=1, ρ=0.9, σ_u^2 =0.01



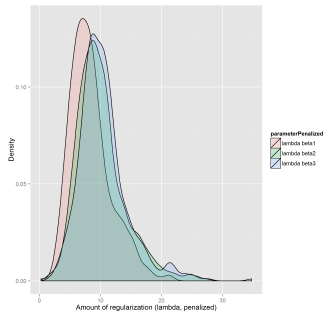
ϵ=1, ρ=0.9, σ_u^2 = 1



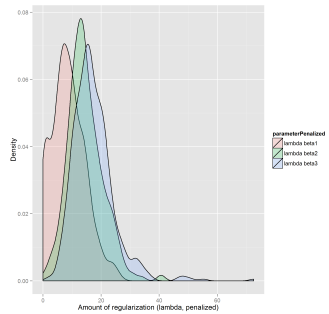
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



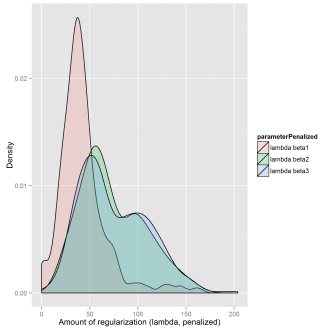
ϵ=2.5, ρ=0.1, σ_u^2 = 1



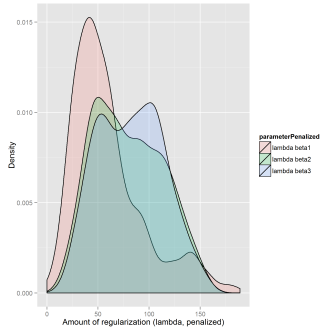
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



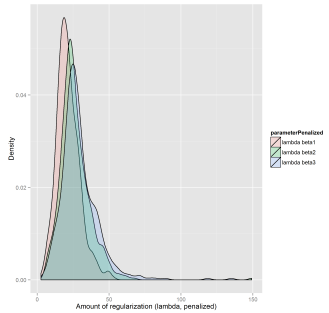
ϵ=2.5, ρ=0.9, σ_u^2 = 1



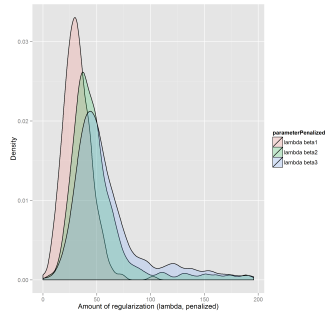
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1

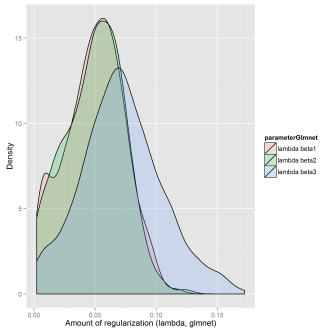


ϵ=5, ρ=0.9, σ_u^2 = 0.01

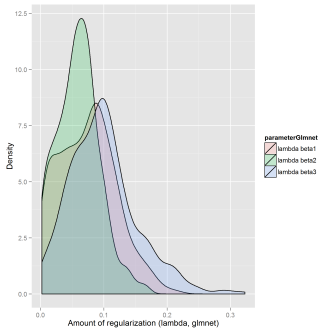


ϵ=5, ρ=0.9, σ_u^2 = 1

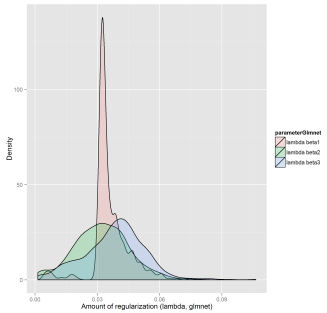
Figure 7: Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=100, p=10)



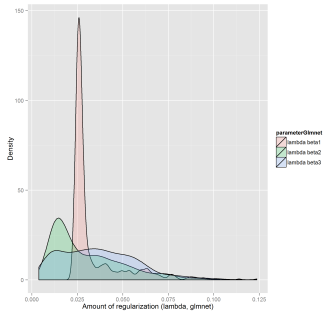
ϵ=1, ρ=0.1, σ_u^2 = 0.01



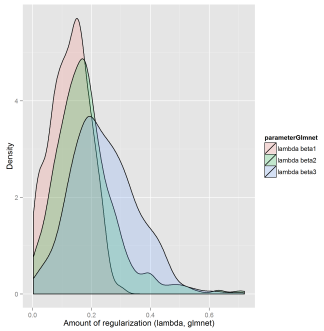
ϵ=1, ρ=0.1, σ_u^2 = 1



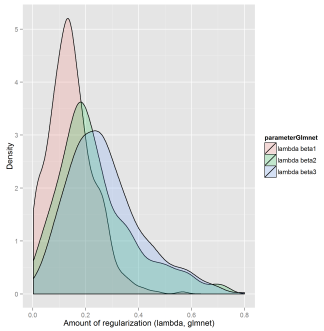
ϵ=1, ρ=0.9, σ_u^2 =0.01



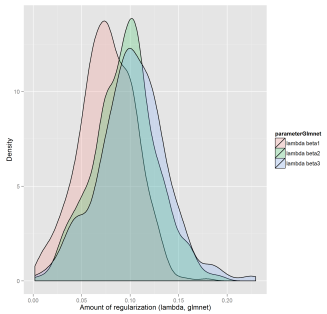
ϵ=1, ρ=0.9, σ_u^2 = 1



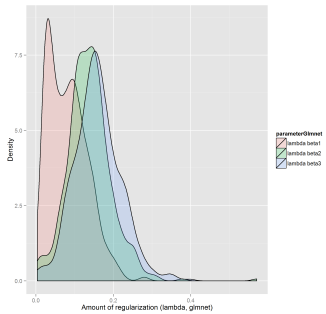
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



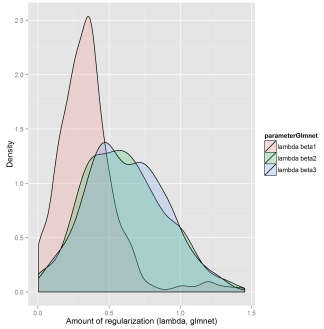
ϵ=2.5, ρ=0.1, σ_u^2 = 1



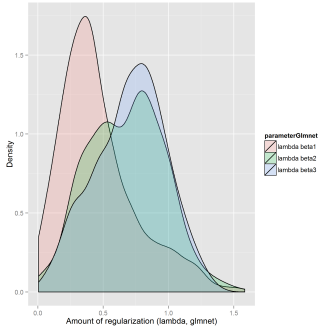
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



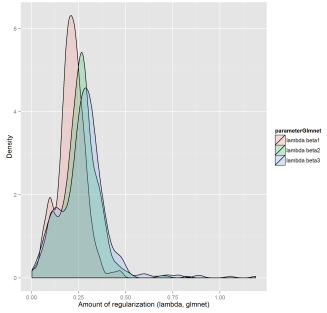
ϵ=2.5, ρ=0.9, σ_u^2 = 1



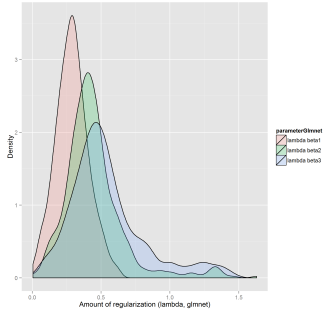
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1

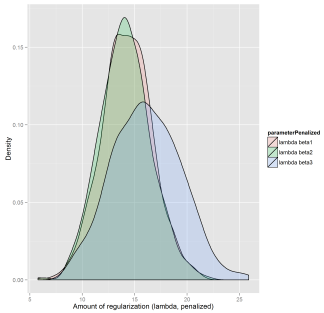


ϵ=5, ρ=0.9, σ_u^2 = 0.01

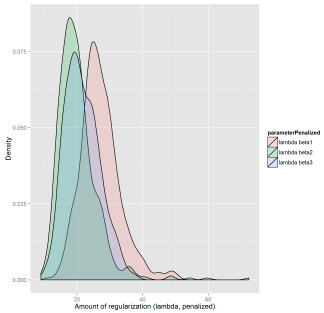


ϵ=5, ρ=0.9, σ_u^2 = 1

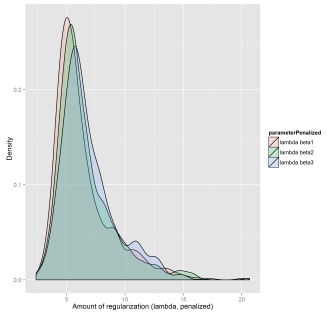
Figure 8: Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=10)



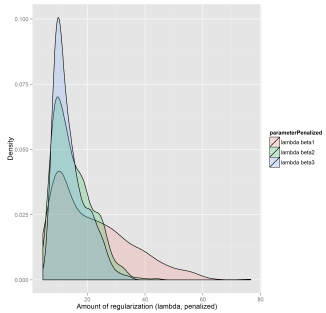
ϵ=1, ρ=0.1, σ_u^2 = 0.01



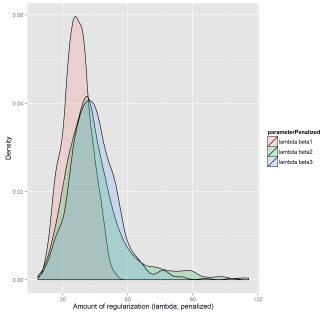
ϵ=1, ρ=0.1, σ_u^2 = 1



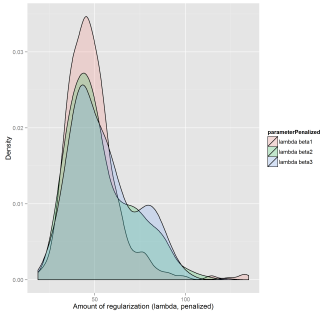
ϵ=1, ρ=0.9, σ_u^2 =0.01



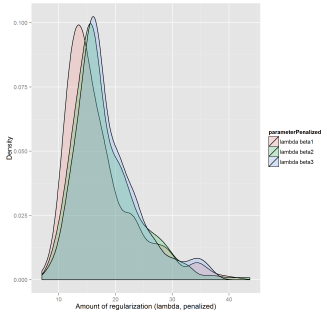
ϵ=1, ρ=0.9, σ_u^2 = 1



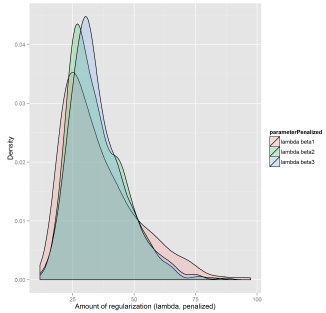
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



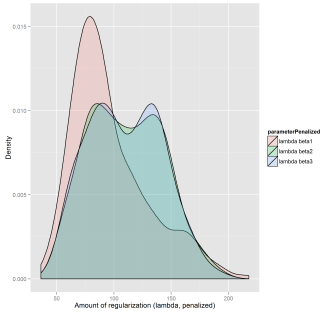
ϵ=2.5, ρ=0.1, σ_u^2 = 1



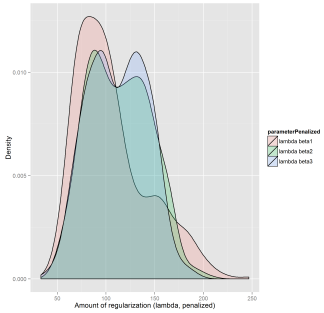
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



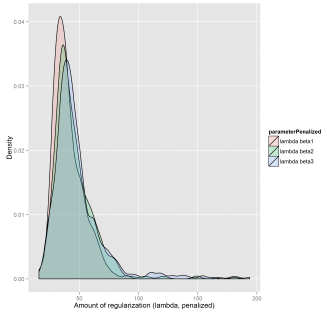
ϵ=2.5, ρ=0.9, σ_u^2 = 1



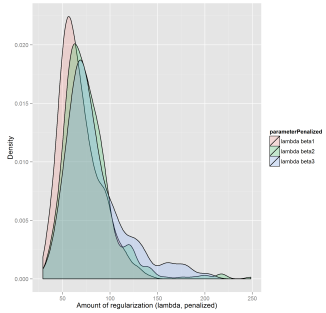
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1

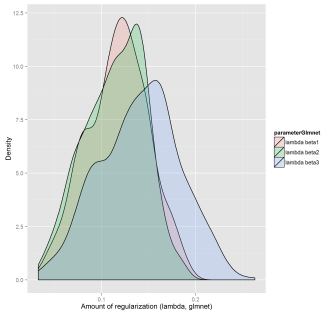


ϵ=5, ρ=0.9, σ_u^2 = 0.01

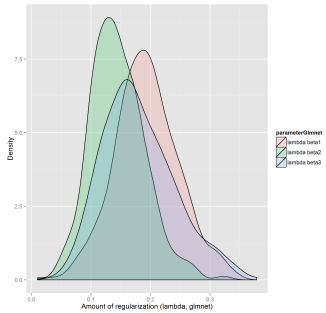


ϵ=5, ρ=0.9, σ_u^2 = 1

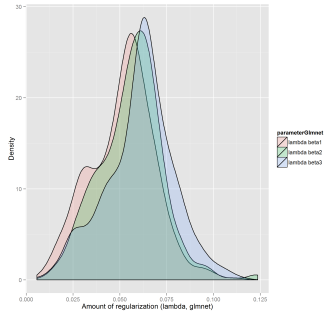
Figure 9: Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=100, p=50)



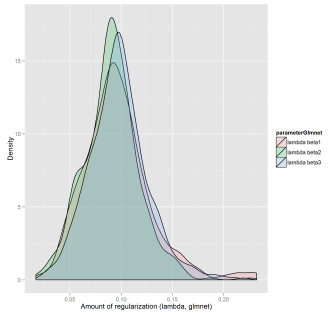
ϵ=1, ρ=0.1, σ_u^2 = 0.01



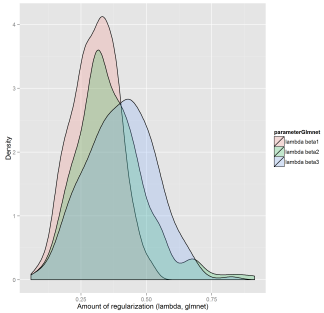
ϵ=1, ρ=0.1, σ_u^2 = 1



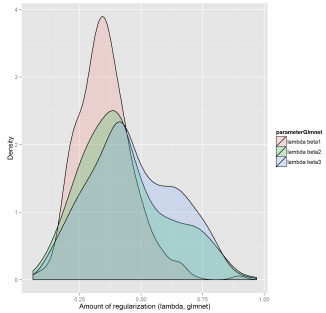
ϵ=1, ρ=0.9, σ_u^2 =0.01



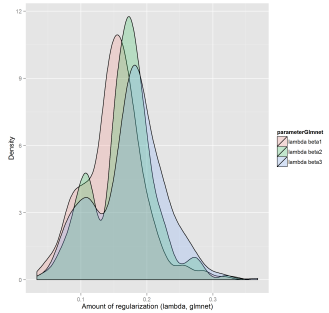
ϵ=1, ρ=0.9, σ_u^2 = 1



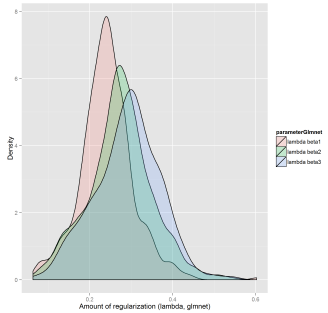
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



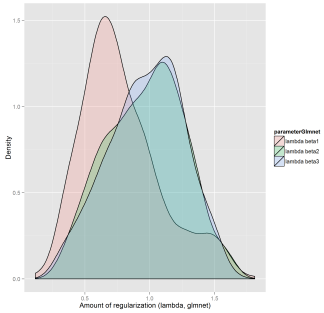
ϵ=2.5, ρ=0.1, σ_u^2 = 1



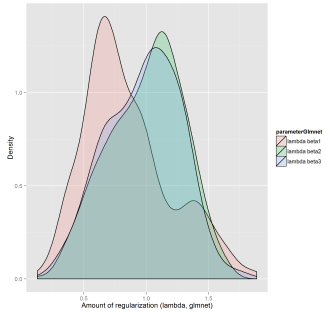
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



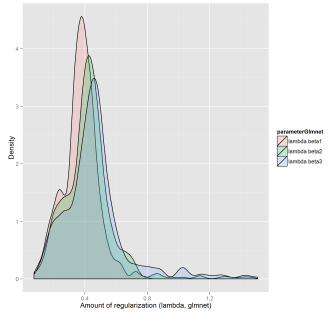
ϵ=2.5, ρ=0.9, σ_u^2 = 1



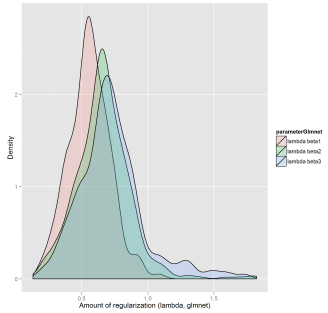
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1

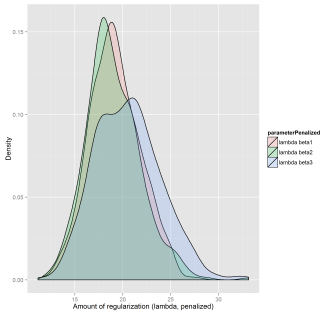


ϵ=5, ρ=0.9, σ_u^2 = 0.01

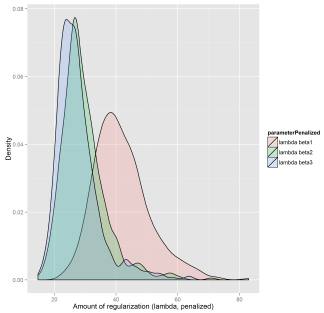


ϵ=5, ρ=0.9, σ_u^2 = 1

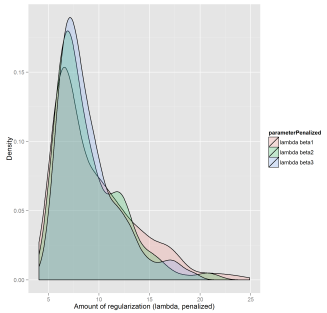
Figure 10: Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=50)



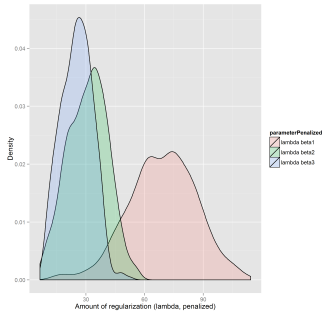
ϵ=1, ρ=0.1, σ_u^2 = 0.01



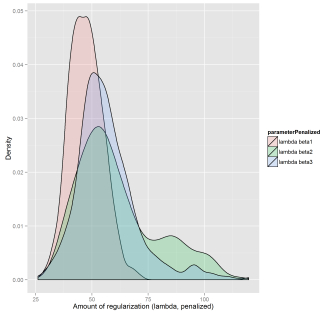
ϵ=1, ρ=0.1, σ_u^2 = 1



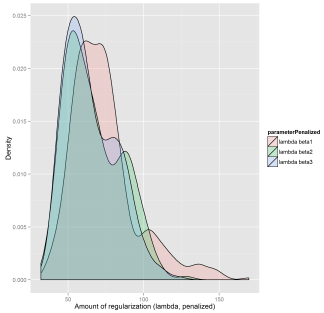
ϵ=1, ρ=0.9, σ_u^2 =0.01



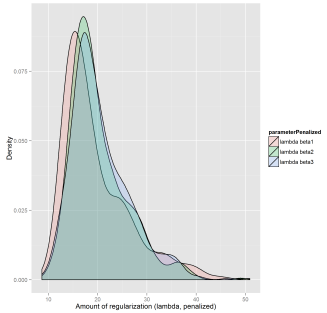
ϵ=1, ρ=0.9, σ_u^2 = 1



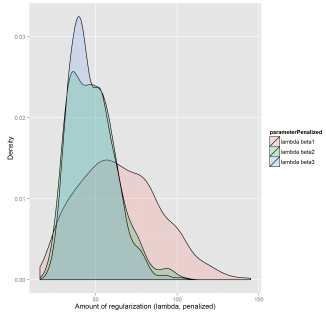
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



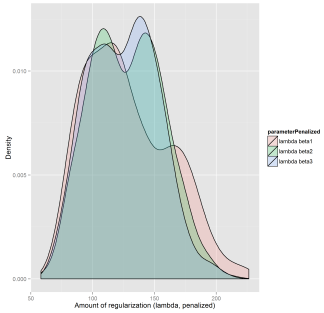
ϵ=2.5, ρ=0.1, σ_u^2 = 1



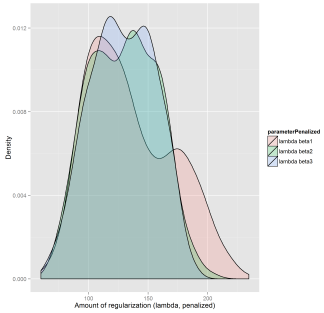
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



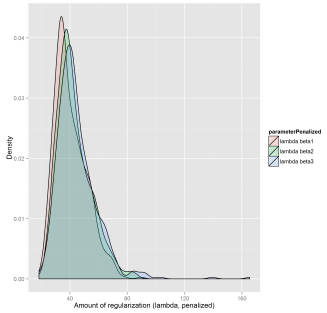
ϵ=2.5, ρ=0.9, σ_u^2 = 1



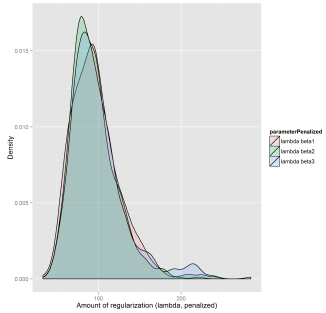
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1

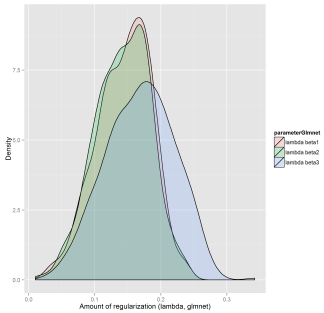


ϵ=5, ρ=0.9, σ_u^2 = 0.01

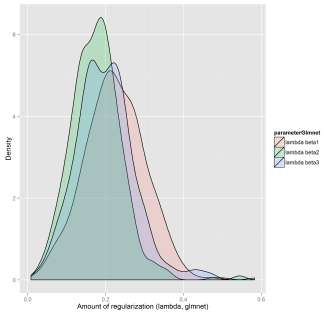


ϵ=5, ρ=0.9, σ_u^2 = 1

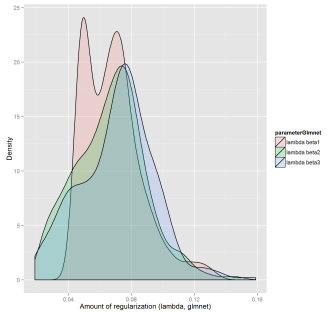
Figure 11: Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=100, p=150)



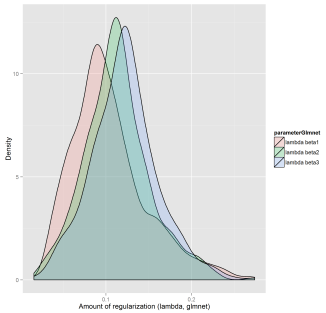
ϵ=1, ρ=0.1, σ_u^2 = 0.01



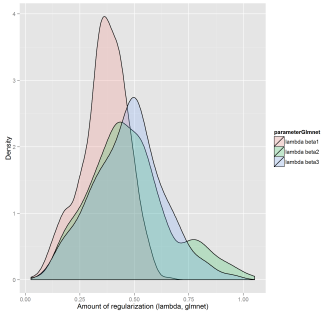
ϵ=1, ρ=0.1, σ_u^2 = 1



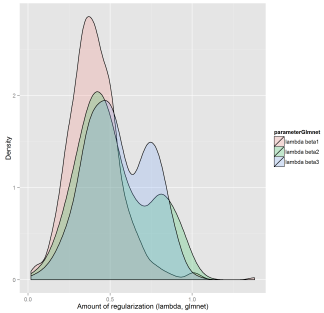
ϵ=1, ρ=0.9, σ_u^2 =0.01



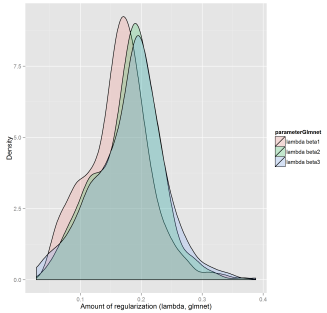
ϵ=1, ρ=0.9, σ_u^2 = 1



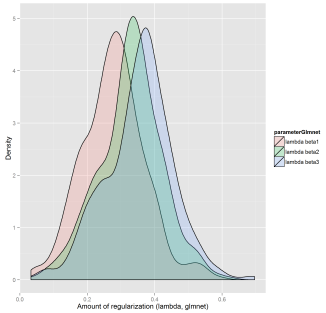
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



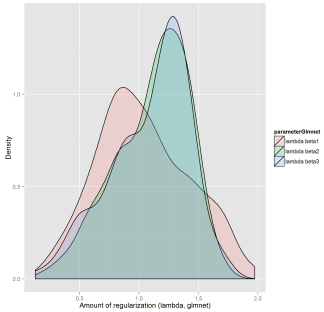
ϵ=2.5, ρ=0.1, σ_u^2 = 1



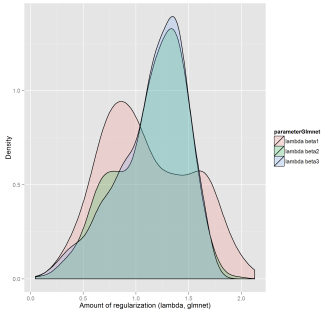
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



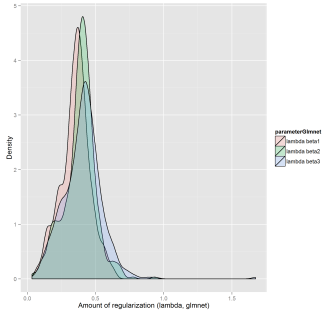
ϵ=2.5, ρ=0.9, σ_u^2 = 1



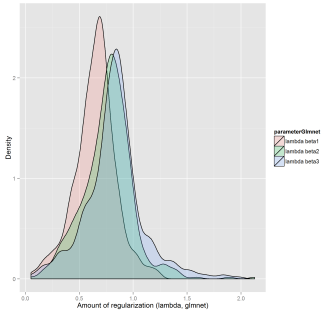
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1



ϵ=5, ρ=0.9, σ_u^2 = 0.01



ϵ=5, ρ=0.9, σ_u^2 = 1

Figure 12: Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=100, p=150)

A.2.3 Tables containing MSE, bias, TPF and FPF (n=500)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.08	0.01			9.55	0.00						
OLS_{TMEV}	1.34	0.00			20005.50	0.00						
OLS_{EMEV}	217.52	0.12			971.13	0.02						
$Lasso_{1,naive}$	0.39	0.02	100 (0.00)	24.91 (2.87)	0.62	0.01	100 (0.00)	8.79 (4.72)				
$Lasso_{1,TMEV}$	0.31	0.01	100 (0.00)	100 (0.00)	1605.81	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	67.92	0.12	100 (0.00)	100 (0.00)	301.87	0.02	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.41	0.02	100 (0.00)	27.02 (3.41)	0.65	0.01	100 (0.00)	9.19 (5.07)				
$Lasso_{2,TMEV}$	0.34	0.01	100 (0.00)	100 (0.00)	1571.35	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	68.25	0.12	100 (0.00)	100 (0.00)	298.87	0.02	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	1.77	0.05			7.52	0.01						
OLS_{TMEV}	196349.00	0.09			236860.00	0.00						
OLS_{EMEV}	49891.70	-0.21			76.05	0.02						
$Lasso_{1,naive}$	1.54	0.05	100 (0.00)	55.09 (2.77)	2.18	0.01	99.92 (0.06)	19.92 (6.86)				
$Lasso_{1,TMEV}$	89466.00	0.03	100 (0.00)	100 (0.00)	66942.60	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	37565.90	-0.21	100 (0.00)	100 (0.00)	23.13	0.02	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	1.56	0.05	100 (0.00)	55.51 (2.81)	2.20	0.01	99.92 (0.06)	19.35 (5.39)				
$Lasso_{2,TMEV}$	88112.00	0.03	100 (0.00)	100 (0.00)	64469.90	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	37568.50	-0.21	100 (0.00)	100 (0.00)	20.53	0.02	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	3.87	0.09			5.45	0.02						
OLS_{TMEV}	7099.59	0.11			1999470.00	-0.04						
OLS_{EMEV}	84.27	-0.03			114.59	0.01						
$Lasso_{1,naive}$	3.86	0.09	99.80 (0.10)	91.96 (1.72)	4.51	0.02	93.48 (0.56)	44.00 (6.11)				
$Lasso_{1,TMEV}$	7743.47	0.10	100 (0.00)	100 (0.00)	2327700.00	-0.06	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	87.13	-0.03	100 (0.00)	100 (0.00)	111.46	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	3.86	0.09	99.80 (0.10)	91.83 (1.69)	4.51	0.02	93.32 (0.57)	43.48 (5.75)				
$Lasso_{2,TMEV}$	7634.88	0.10	100 (0.00)	100 (0.00)	2323210.00	-0.06	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	85.88	-0.03	100 (0.00)	100 (0.00)	111.24	0.01	100 (0.00)	100 (0.00)				

Table 35: MSE, bias, TPF and FPF for β_1 (n=500, ρ = 0.9, ϵ = 1)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.02	0.00			9.20	0.00						
OLS_{TMEV}	1.29	0.00			21966.20	0.00						
OLS_{EMEV}	104.57	0.06			548.69	0.01						
$Lasso_{1,naive}$	0.32	0.02	99.52 (0.15)	25.43 (4.21)	0.51	0.00	97.28 (0.35)	8.84 (7.33)				
$Lasso_{1,TMEV}$	0.32	0.01	100 (0.00)	100 (0.00)	306.36	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	15.27	0.06	100 (0.00)	100 (0.00)	56.31	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.32	0.02	99.44 (0.17)	24.20 (3.31)	0.51	0.00	96.96 (0.39)	8.26 (4.95)				
$Lasso_{2,TMEV}$	0.32	0.01	100 (0.00)	100 (0.00)	296.72	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	15.72	0.06	100 (0.00)	100 (0.00)	52.35	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.85	0.02			5.64	0.00						
OLS_{TMEV}	105284.00	0.05			145567.00	0.00						
OLS_{EMEV}	17976.10	-0.09			67.87	0.01						
$Lasso_{1,naive}$	0.57	0.03	98.6 (0.26)	41.34 (3.04)	0.86	0.01	90.80 (0.64)	14.53 (6.30)				
$Lasso_{1,TMEV}$	13141.90	0.00	100 (0.00)	100 (0.00)	15687.00	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	9671.54	-0.11	100 (0.00)	100 (0.00)	8.78	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.58	0.03	98.52 (0.27)	40.83 (2.82)	0.86	0.01	90.60 (0.64)	14.13 (5.02)				
$Lasso_{2,TMEV}$	12818.40	0.00	100 (0.00)	100 (0.00)	15061.60	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	9612.76	-0.11	100 (0.00)	100 (0.00)	4.95	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	1.04	0.04			2.04	0.01						
OLS_{TMEV}	3454.02	0.06			1488650.00	-0.03						
OLS_{EMEV}	13.67	-0.01			29.23	0.00						
$Lasso_{1,naive}$	1.02	0.04	96.28 (0.40)	78.75 (2.56)	1.22	0.01	69.08 (1.01)	31.27 (6.31)				
$Lasso_{1,TMEV}$	3443.35	0.05	100 (0.00)	100 (0.00)	794474.00	-0.03	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	17.02	-0.01	100 (0.00)	100 (0.00)	27.31	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	1.02	0.04	96.32 (0.40)	78.46 (2.44)	1.22	0.01	69.04 (1.02)	30.78 (5.77)				
$Lasso_{2,TMEV}$	3253.21	0.05	100 (0.00)	100 (0.00)	788765.00	-0.03	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	17.03	-0.01	100 (0.00)	100 (0.00)	27.02	0.00	100 (0.00)	100 (0.00)				

Table 36: MSE, bias, TPF and FPF for β_2 (n=500, ρ = 0.9, ϵ = 1)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.02	0.00			9.21	0.00						
OLS_{TMEV}	1.29	0.00			19331.20	0.00						
OLS_{EMEV}	134.41	0.05			648.46	0.01						
$Lasso_{1,naive}$	0.25	0.01	71.72 (0.79)	22.19 (3.80)	0.37	0.00	58.44 (0.80)	7.22 (6.51)				
$Lasso_{1,TMEV}$	0.24	0.01	100 (0.00)	100 (0.00)	787.35	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	25.14	0.04	100 (0.00)	100 (0.00)	68.00	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.24	0.01	70.84 (0.80)	20.96 (3.20)	0.36	0.00	57.84 (0.80)	6.78 (4.86)				
$Lasso_{2,TMEV}$	0.23	0.01	100 (0.00)	100 (0.00)	770.55	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	23.95	0.04	100 (0.00)	100 (0.00)	65.88	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.90	0.02			5.66	0.00						
OLS_{TMEV}	114371.00	0.08			113155.00	-0.01						
OLS_{EMEV}	23287.90	-0.06			90.02	0.01						
$Lasso_{1,naive}$	0.58	0.02	77.68 (0.77)	36.34 (3.02)	0.71	0.01	57.32 (0.77)	11.59 (5.41)				
$Lasso_{1,TMEV}$	23566.90	0.04	100 (0.00)	100 (0.00)	20747.50	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	14252.40	-0.07	100 (0.00)	100 (0.00)	9.69	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.58	0.02	77.44 (0.77)	35.82 (2.89)	0.71	0.01	56.92 (0.77)	11.29 (4.56)				
$Lasso_{2,TMEV}$	23396.00	0.04	100 (0.00)	100 (0.00)	20427.70	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	14279.80	-0.07	100 (0.00)	100 (0.00)	9.69	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	1.15	0.03			2.12	0.01						
OLS_{TMEV}	5365.51	0.06			1616120.00	0.00						
OLS_{EMEV}	30.64	-0.01			18.82	0.00						
$Lasso_{1,naive}$	1.12	0.03	84.44 (0.74)	70.36 (2.80)	1.27	0.01	52.00 (1.02)	26.31 (6.14)				
$Lasso_{1,TMEV}$	4457.17	0.05	100 (0.00)	100 (0.00)	417329.00	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	10.97	-0.01	100 (0.00)	100 (0.00)	16.97	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	1.12	0.03	84.44 (0.74)	70.24 (2.71)	1.27	0.01	52.20 (1.02)	25.91 (5.57)				
$Lasso_{2,TMEV}$	4434.26	0.05	100 (0.00)	100 (0.00)	430842.00	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	10.95	-0.01	100 (0.00)	100 (0.00)	17.04	0.00	100 (0.00)	100 (0.00)				

Table 37: MSE, bias, TPF and FPF for β_3 (n=500, ρ = 0.9, ϵ = 1)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.22	0.00			2.03	0.00						
OLS_{TMEV}	0.23	0.00			2.41	0.00						
OLS_{EMEV}	2804.63	0.11			6910.83	0.02						
$Lasso_{1,naive}$	0.07	0.01	100 (0.00)	23.59 (4.12)	0.12	0.00	100 (0.00)	9.18 (7.97)				
$Lasso_{1,TMEV}$	0.06	0.00	100 (0.00)	100 (0.00)	0.09	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	2440.68	0.11	100 (0.00)	100 (0.00)	4473.30	0.02	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.06	0.01	100 (0.00)	23.12 (3.51)	0.12	0.00	100 (0.00)	8.46 (5.51)				
$Lasso_{2,TMEV}$	0.05	0.01	100 (0.00)	100 (0.00)	0.09	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	2442.05	0.11	100 (0.00)	100 (0.00)	4447.36	0.02	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.37	0.02			2.49	0.00						
OLS_{TMEV}	0.45	0.00			206700.00	0.00						
OLS_{EMEV}	240073.00	-3.16			256.47	0.03						
$Lasso_{1,naive}$	0.25	0.02	100 (0.00)	40.46 (3.47)	0.36	0.00	100 (0.00)	13.53 (6.29)				
$Lasso_{1,TMEV}$	0.14	0.00	100 (0.00)	100 (0.00)	53837.70	0.01	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	233381.00	-3.12	100 (0.00)	100 (0.00)	976.72	0.02	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.25	0.02	100 (0.00)	39.94 (3.17)	0.36	0.00	100 (0.00)	13.10 (5.20)				
$Lasso_{2,TMEV}$	0.14	0.00	100 (0.00)	100 (0.00)	53342.50	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	235073.00	-3.13	100 (0.00)	100 (0.00)	976.43	0.03	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	2.25	0.06			3.95	0.01						
OLS_{TMEV}	428608.00	-0.27			16219100.00	-0.07						
OLS_{EMEV}	55.94	-0.02			136.77	0.01						
$Lasso_{1,naive}$	2.24	0.06	100 (0.00)	80.13 (2.41)	2.54	0.01	100 (0.00)	32.34 (6.11)				
$Lasso_{1,TMEV}$	329976.00	-0.24	100 (0.00)	100 (0.00)	1299210.00	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	85.45	-0.02	100 (0.00)	100 (0.00)	131.34	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	2.24	0.06	100 (0.00)	79.98 (2.35)	2.54	0.01	100 (0.00)	31.55 (5.25)				
$Lasso_{2,TMEV}$	331967.00	-0.24	100 (0.00)	100 (0.00)	1399780.00	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	85.30	-0.02	100 (0.00)	100 (0.00)	129.00	0.01	100 (0.00)	100 (0.00)				

Table 38: MSE, bias, TPF and FPF for β_1 (n=500, ρ = 0.5, ϵ = 1)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.21	0.00			1.95	0.00						
OLS_{TMEV}	0.22	0.00			2.32	0.00						
OLS_{EMEV}	685.51	0.06			2489.49	0.01						
$Lasso_{1,naive}$	0.06	0.01	100 (0.00)	22.91 (4.31)	0.11	0.00	100 (0.00)	9.24 (10.06)				
$Lasso_{1,TMEV}$	0.05	0.01	100 (0.00)	100 (0.00)	0.10	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	524.47	0.06	100 (0.00)	100 (0.00)	980.10	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.06	0.01	100 (0.00)	22.00 (3.57)	0.10	0.00	100 (0.00)	8.24 (5.79)				
$Lasso_{2,TMEV}$	0.05	0.01	100 (0.00)	100 (0.00)	0.09	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	518.09	0.06	100 (0.00)	100 (0.00)	913.43	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.23	0.01			1.87	0.00						
OLS_{TMEV}	0.33	0.00			151581.00	0.00						
OLS_{EMEV}	62946.60	-1.62			89.16	0.01						
$Lasso_{1,naive}$	0.12	0.01	100 (0.00)	32.32 (3.73)	0.18	0.00	100 (0.00)	11.05 (6.93)				
$Lasso_{1,TMEV}$	0.08	0.00	100 (0.00)	100 (0.00)	14073.90	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	59041.90	-1.56	100 (0.00)	100 (0.00)	200.86	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.11	0.01	100 (0.00)	31.62 (3.37)	0.17	0.00	100 (0.00)	10.48 (5.07)				
$Lasso_{2,TMEV}$	0.08	0.00	100 (0.00)	100 (0.00)	14128.90	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	59769.40	-1.57	100 (0.00)	100 (0.00)	200.35	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	0.62	0.03			1.49	0.01						
OLS_{TMEV}	148100.00	-0.13			5156790.00	-0.05						
OLS_{EMEV}	7.94	-0.01			34.51	0.01						
$Lasso_{1,naive}$	0.60	0.03	100 (0.00)	68.83 (2.88)	0.71	0.01	99.96 (0.04)	25.70 (6.60)				
$Lasso_{1,TMEV}$	90447.00	-0.12	100 (0.00)	100 (0.00)	770733.00	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	19.09	-0.01	100 (0.00)	100 (0.00)	31.51	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.60	0.03	100 (0.00)	68.47 (2.83)	0.71	0.01	99.96 (0.04)	25.06 (5.49)				
$Lasso_{2,TMEV}$	90182.40	-0.12	100 (0.00)	100 (0.00)	782889.00	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	18.49	-0.01	100 (0.00)	100 (0.00)	31.51	0.01	100 (0.00)	100 (0.00)				

Table 39: MSE, bias, TPF and FPF for β_2 (n=500, ρ = 0.5, ϵ = 1)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.21	0.00			1.95	0.00						
OLS_{TMEV}	0.22	0.00			2.32	0.00						
OLS_{EMEV}	1714.72	0.05			2935.00	0.01						
$Lasso_{1,naive}$	0.05	0.01	88.92 (0.61)	21.33 (4.16)	0.09	0.00	79.36 (0.63)	8.01 (8.53)				
$Lasso_{1,TMEV}$	0.05	0.00	100 (0.00)	100 (0.00)	0.08	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	1463.75	0.05	100 (0.00)	100 (0.00)	1327.20	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.05	0.01	88.76 (0.62)	20.57 (3.56)	0.09	0.00	79.12 (0.63)	7.56 (6.32)				
$Lasso_{2,TMEV}$	0.05	0.01	100 (0.00)	100 (0.00)	0.08	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	1462.88	0.05	100 (0.00)	100 (0.00)	1241.50	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.23	0.01			1.89	0.00						
OLS_{TMEV}	0.34	0.00			155438.00	0.00						
OLS_{EMEV}	38488.00	-1.26			1143.08	0.01						
$Lasso_{1,naive}$	0.11	0.01	88.72 (0.61)	28.53 (3.80)	0.16	0.00	77.68 (0.71)	9.94 (7.70)				
$Lasso_{1,TMEV}$	0.08	0.00	100 (0.00)	100 (0.00)	15638.00	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	35428.10	-1.21	100 (0.00)	100 (0.00)	75.73	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.11	0.01	88.48 (0.62)	28.00 (3.53)	0.15	0.00	77.64 (0.70)	9.28 (5.25)				
$Lasso_{2,TMEV}$	0.07	0.00	100 (0.00)	100 (0.00)	15746.70	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	36899.00	-1.24	100 (0.00)	100 (0.00)	76.64	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	0.66	0.02			1.55	0.01						
OLS_{TMEV}	222620.00	-0.15			6495340.00	-0.06						
OLS_{EMEV}	2.92	0.00			21.83	0.00						
$Lasso_{1,naive}$	0.63	0.02	90.80 (0.57)	58.99 (2.98)	0.73	0.01	77.16 (0.78)	21.87 (6.73)				
$Lasso_{1,TMEV}$	275087.00	-0.14	100 (0.00)	100 (0.00)	1098690.00	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	4.13	-0.01	100 (0.00)	100 (0.00)	19.13	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.63	0.02	90.84 (0.56)	58.67 (2.87)	0.73	0.01	77.08 (0.78)	21.19 (5.18)				
$Lasso_{2,TMEV}$	272544.00	-0.14	100 (0.00)	100 (0.00)	1091400.00	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	4.11	-0.01	100 (0.00)	100 (0.00)	19.29	0.00	100 (0.00)	100 (0.00)				

Table 40: MSE, bias, TPF and FPF for β_3 (n=500, ρ = 0.5, ϵ = 1)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.13	0.00			1.18	0.00						
OLS_{TMEV}	0.13	0.00			1.30	0.00						
OLS_{EMEV}	16961.00	0.25			47335.20	0.02						
$Lasso_{1,naive}$	0.04	0.00	100 (0.00)	21.53 (5.02)	0.08	0.00	100 (0.00)	8.58 (10.97)				
$Lasso_{1,TMEV}$	0.03	0.00	100 (0.00)	100 (0.00)	0.06	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	15380.70	0.23	100 (0.00)	100 (0.00)	42052.20	0.02	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.04	0.00	100 (0.00)	22.53 (4.44)	0.07	0.00	100 (0.00)	7.69 (6.34)				
$Lasso_{2,TMEV}$	0.03	0.00	100 (0.00)	100 (0.00)	0.06	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	15812.20	0.23	100 (0.00)	100 (0.00)	41935.60	0.02	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.20	0.01			1.52	0.00						
OLS_{TMEV}	0.22	0.00			19001.70	-0.01						
OLS_{EMEV}	740.65	-0.05			1889.31	-0.06						
$Lasso_{1,naive}$	0.12	0.01	100 (0.00)	29.92 (4.51)	0.21	0.00	100 (0.00)	11.04 (9.50)				
$Lasso_{1,TMEV}$	0.06	0.00	100 (0.00)	100 (0.00)	13716.30	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	593.35	-0.05	100 (0.00)	100 (0.00)	1375.32	-0.06	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.12	0.01	100 (0.00)	29.68 (3.95)	0.21	0.00	100 (0.00)	10.48 (6.70)				
$Lasso_{2,TMEV}$	0.06	0.00	100 (0.00)	100 (0.00)	13696.50	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	516.12	-0.04	100 (0.00)	100 (0.00)	1379.58	-0.06	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	1.36	0.04			3.26	0.00						
OLS_{TMEV}	436518.00	-0.07			28998.60	0.00						
OLS_{EMEV}	13.95	-0.01			20.03	0.00						
$Lasso_{1,naive}$	1.35	0.03	100 (0.00)	57.01 (4.01)	2.04	0.00	100 (0.00)	26.32 (8.53)				
$Lasso_{1,TMEV}$	367342.00	-0.08	100 (0.00)	100 (0.00)	12422.00	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	11.88	-0.01	100 (0.00)	100 (0.00)	11.92	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	1.35	0.03	100 (0.00)	56.57 (3.72)	2.04	0.00	100 (0.00)	23.68 (6.28)				
$Lasso_{2,TMEV}$	362068.00	-0.08	100 (0.00)	100 (0.00)	11938.70	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	11.86	-0.01	100 (0.00)	100 (0.00)	11.65	0.00	100 (0.00)	100 (0.00)				

Table 41: MSE, bias, TPF and FPF for β_1 (n=500, ρ = 0.1, ϵ = 1)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.12	0.00			1.13	0.00						
OLS_{TMEV}	0.13	0.00			1.25	0.00						
OLS_{EMEV}	4605.95	0.12			13916.00	0.01						
$Lasso_{1,naive}$	0.04	0.00	100 (0.00)	22.05 (5.52)	0.07	0.00	100 (0.00)	8.29 (10.74)				
$Lasso_{1,TMEV}$	0.03	0.00	100 (0.00)	100 (0.00)	0.06	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	3721.43	0.11	100 (0.00)	100 (0.00)	10253.60	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.04	0.00	100 (0.00)	21.71 (4.43)	0.07	0.00	100 (0.00)	7.87 (7.20)				
$Lasso_{2,TMEV}$	0.03	0.00	100 (0.00)	100 (0.00)	0.06	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	3762.88	0.11	100 (0.00)	100 (0.00)	10485.00	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.13	0.00			1.14	0.00						
OLS_{TMEV}	0.16	0.00			7713.11	-0.01						
OLS_{EMEV}	282.19	-0.03			657.93	-0.03						
$Lasso_{1,naive}$	0.06	0.00	100 (0.00)	25.96 (4.80)	0.11	0.00	100 (0.00)	9.30 (10.11)				
$Lasso_{1,TMEV}$	0.04	0.00	100 (0.00)	100 (0.00)	2787.45	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	182.49	-0.02	100 (0.00)	100 (0.00)	322.44	-0.03	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.06	0.01	100 (0.00)	26.25 (4.33)	0.11	0.00	100 (0.00)	8.78 (6.32)				
$Lasso_{2,TMEV}$	0.04	0.00	100 (0.00)	100 (0.00)	2859.44	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	181.21	-0.02	100 (0.00)	100 (0.00)	325.46	-0.03	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	0.38	0.02			1.21	0.01						
OLS_{TMEV}	368930.00	-0.03			34442.70	0.01						
OLS_{EMEV}	3.65	0.00			7.37	0.00						
$Lasso_{1,naive}$	0.37	0.01	100 (0.00)	46.63 (4.58)	0.58	0.00	100 (0.00)	20.76 (9.67)				
$Lasso_{1,TMEV}$	137136.00	-0.03	100 (0.00)	100 (0.00)	2690.65	0.01	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	2.91	0.00	100 (0.00)	100 (0.00)	3.06	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.37	0.01	100 (0.00)	46.22 (4.23)	0.58	0.00	100 (0.00)	18.69 (6.87)				
$Lasso_{2,TMEV}$	136981.00	-0.03	100 (0.00)	100 (0.00)	2741.03	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	2.90	0.00	100 (0.00)	100 (0.00)	2.97	0.00	100 (0.00)	100 (0.00)				

Table 42: MSE, bias, TPF and FPF for β_2 (n=500, ρ = 0.1, ϵ = 1)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	0.12	0.00			1.13	0.00						
OLS_{TMEV}	0.13	0.00			1.25	0.00						
OLS_{EMEV}	2161.83	0.10			18051.90	0.01						
$Lasso_{1,naive}$	0.03	0.00	89.20 (0.58)	19.34 (5.48)	0.06	0.00	80.8 (0.62)	7.12 (10.40)				
$Lasso_{1,TMEV}$	0.03	0.00	100 (0.00)	100 (0.00)	0.05	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	2138.80	0.08	100 (0.00)	100 (0.00)	12344.30	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.03	0.00	89.28 (0.58)	19.16 (4.38)	0.06	0.00	80.92 (0.62)	6.81 (6.88)				
$Lasso_{2,TMEV}$	0.03	0.00	100 (0.00)	100 (0.00)	0.05	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	2194.52	0.08	100 (0.00)	100 (0.00)	12027.50	0.00	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	0.14	0.00			1.14	0.00						
OLS_{TMEV}	0.17	0.00			7440.42	-0.01						
OLS_{EMEV}	25.71	-0.01			443.34	-0.02						
$Lasso_{1,naive}$	0.06	0.00	92.60 (0.53)	23.36 (4.82)	0.10	0.00	83.12 (0.59)	8.20 (9.74)				
$Lasso_{1,TMEV}$	0.04	0.00	100 (0.00)	100 (0.00)	2138.64	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	33.11	-0.01	100 (0.00)	100 (0.00)	110.49	-0.02	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.06	0.00	92.56 (0.53)	23.64 (4.22)	0.10	0.00	83.20 (0.59)	7.83 (6.62)				
$Lasso_{2,TMEV}$	0.04	0.00	100 (0.00)	100 (0.00)	2174.98	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	32.15	-0.01	100 (0.00)	100 (0.00)	112.24	-0.02	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	0.46	0.02			1.24	0.00						
OLS_{TMEV}	231422.00	-0.03			38465.90	0.00						
OLS_{EMEV}	2.67	0.00			7.83	0.00						
$Lasso_{1,naive}$	0.43	0.01	93.00 (0.49)	40.72 (5.02)	0.56	0.00	73.64 (0.74)	15.54 (9.84)				
$Lasso_{1,TMEV}$	37340.20	-0.04	100 (0.00)	100 (0.00)	3280.45	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	3.38	0.00	100 (0.00)	100 (0.00)	3.41	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.43	0.01	92.92 (0.50)	40.29 (4.24)	0.56	0.00	72.64 (0.76)	14.04 (7.45)				
$Lasso_{2,TMEV}$	10156.10	-0.04	100 (0.00)	100 (0.00)	3256.88	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	3.40	0.00	100 (0.00)	100 (0.00)	3.26	0.00	100 (0.00)	100 (0.00)				

Table 43: MSE, bias, TPF and FPF for β_3 (n=500, ρ = 0.1, ϵ = 1)

	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.36	0.00			12.34	0.00						
OLS_{TMEV}	1.43	0.00			14.72	0.00						
OLS_{EMEV}	30980.90	0.20			33223.90	0.02						
$Lasso_{1,naive}$	0.39	0.02	100 (0.00)	23.80 (4.46)	0.64	0.00	100 (0.00)	8.52 (8.82)				
$Lasso_{1,TMEV}$	0.38	0.01	100 (0.00)	100 (0.00)	0.60	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	35505.80	0.17	100 (0.00)	100 (0.00)	12309.10	0.02	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.38	0.02	100 (0.00)	22.42 (3.77)	0.63	0.00	100 (0.00)	7.92 (6.02)				
$Lasso_{2,TMEV}$	0.36	0.01	100 (0.00)	100 (0.00)	0.58	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	31211.70	0.18	100 (0.00)	100 (0.00)	12385.60	0.02	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	1.34	0.02			11.25	0.00						
OLS_{TMEV}	2.04	0.00			66192000.00	0.04						
OLS_{EMEV}	3945.74	0.29			143943.00	0.01						
$Lasso_{1,naive}$	0.61	0.02	100 (0.00)	29.76 (3.98)	1.12	0.01	100 (0.00)	11.24 (7.32)				
$Lasso_{1,TMEV}$	0.49	0.01	100 (0.00)	100 (0.00)	149039.00	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	3488.24	0.28	100 (0.00)	100 (0.00)	38141.60	0.02	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.60	0.02	100 (0.00)	29.03 (3.54)	1.11	0.01	100 (0.00)	10.69 (5.12)				
$Lasso_{2,TMEV}$	0.47	0.01	100 (0.00)	100 (0.00)	141324.00	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	3404.13	0.28	100 (0.00)	100 (0.00)	38483.50	0.02	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	2.68	0.06			7.49	0.01						
OLS_{TMEV}	943795.00	-0.17			1735990.00	0.03						
OLS_{EMEV}	234.45	-0.02			51.82	-0.01						
$Lasso_{1,naive}$	2.57	0.06	100 (0.00)	63.99 (2.79)	3.20	0.01	99.44 (0.17)	23.07 (7.10)				
$Lasso_{1,TMEV}$	53602.10	0.02	100 (0.00)	100 (0.00)	3346780.00	0.03	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	213.92	-0.02	100 (0.00)	100 (0.00)	17.90	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	2.56	0.06	100 (0.00)	63.97 (2.78)	3.19	0.01	99.48 (0.16)	22.49 (5.74)				
$Lasso_{2,TMEV}$	54393.30	0.02	100 (0.00)	100 (0.00)	3756530.00	0.03	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	211.26	-0.02	100 (0.00)	100 (0.00)	18.74	-0.01	100 (0.00)	100 (0.00)				

Table 44: MSE, bias, TPF and FPF for β_1 (n=500, ρ = 0.5, ϵ = 2.5)

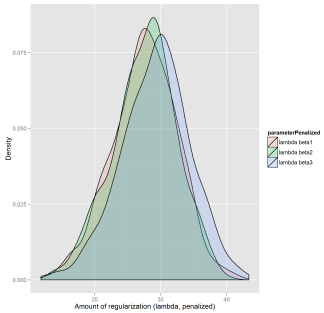
	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.35	0.00			12.27	0.00						
OLS_{TMEV}	1.42	0.00			14.63	0.00						
OLS_{EMEV}	8978.69	0.11			18467.80	0.01						
$Lasso_{1,naive}$	0.37	0.01	99.12 (0.21)	23.05 (4.44)	0.59	0.00	94.80 (0.48)	8.26 (8.78)				
$Lasso_{1,TMEV}$	0.37	0.01	100 (0.00)	100 (0.00)	0.60	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	4179.08	0.09	100 (0.00)	100 (0.00)	3049.61	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.36	0.01	99.08 (0.21)	22.12 (3.86)	0.57	0.00	94.68 (0.49)	7.49 (5.25)				
$Lasso_{2,TMEV}$	0.36	0.01	100 (0.00)	100 (0.00)	0.57	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	4330.68	0.09	100 (0.00)	100 (0.00)	2743.65	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	1.20	0.01			10.58	0.00						
OLS_{TMEV}	1.92	0.00			73392600.00	0.05						
OLS_{EMEV}	1016.39	0.15			74382.50	0.00						
$Lasso_{1,naive}$	0.41	0.02	98.40 (0.29)	25.20 (3.82)	0.74	0.01	91.20 (0.63)	9.66 (7.35)				
$Lasso_{1,TMEV}$	0.43	0.01	100 (0.00)	100 (0.00)	23004.10	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	795.37	0.14	100 (0.00)	100 (0.00)	7396.17	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.40	0.02	98.44 (0.28)	24.52 (3.48)	0.73	0.01	91.04 (0.64)	9.13 (5.30)				
$Lasso_{2,TMEV}$	0.42	0.01	100 (0.00)	100 (0.00)	23698.40	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	776.86	0.14	100 (0.00)	100 (0.00)	7428.92	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	1.01	0.03			5.03	0.01						
OLS_{TMEV}	827259.00	-0.24			1367390.00	0.02						
OLS_{EMEV}	122.88	-0.01			33.01	-0.01						
$Lasso_{1,naive}$	0.83	0.03	94.48 (0.52)	46.90 (3.40)	1.08	0.01	75.88 (0.97)	15.80 (6.83)				
$Lasso_{1,TMEV}$	47656.10	-0.04	100 (0.00)	100 (0.00)	1690530.00	0.01	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	72.11	-0.01	100 (0.00)	100 (0.00)	5.69	-0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.83	0.03	94.40 (0.53)	46.63 (3.13)	1.07	0.01	75.80 (0.98)	15.37 (5.38)				
$Lasso_{2,TMEV}$	47295.00	-0.04	100 (0.00)	100 (0.00)	1653040.00	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	73.27	-0.01	100 (0.00)	100 (0.00)	5.35	-0.01	100 (0.00)	100 (0.00)				

Table 45: MSE, bias, TPF and FPF for β_2 (n=500, ρ = 0.5, ϵ = 2.5)

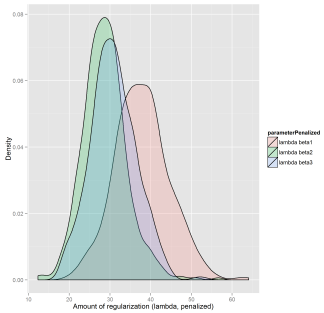
	$p = 50$				$p = 250$				$p = 750$			
	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF	MSE	Bias	TPF	FPF
$\sigma_u^2 = 0.01$												
OLS_{naive}	1.35	0.00			12.27	0.00						
OLS_{TMEV}	1.42	0.00			14.63	0.00						
OLS_{EMEV}	15970.50	0.10			23870.70	0.01						
$Lasso_{1,naive}$	0.27	0.01	68.00 (0.80)	19.06 (4.07)	0.35	0.00	52.36 (0.71)	6.09 (7.92)				
$Lasso_{1,TMEV}$	0.27	0.01	100 (0.00)	100 (0.00)	0.35	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	4055.44	0.07	100 (0.00)	100 (0.00)	6933.42	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.26	0.01	67.72 (0.79)	18.08 (3.44)	0.34	0.00	52.52 (0.70)	5.72 (5.27)				
$Lasso_{2,TMEV}$	0.26	0.01	100 (0.00)	100 (0.00)	0.34	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	4269.23	0.07	100 (0.00)	100 (0.00)	6862.84	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 0.1$												
OLS_{naive}	1.20	0.01			10.58	0.00						
OLS_{TMEV}	1.92	0.00			72127100.00	0.05						
OLS_{EMEV}	792.52	0.12			14186.90	0.01						
$Lasso_{1,naive}$	0.31	0.01	70.48 (0.81)	21.01 (3.79)	0.51	0.00	51.76 (0.77)	7.32 (7.11)				
$Lasso_{1,TMEV}$	0.31	0.01	100 (0.00)	100 (0.00)	636841.00	0.00	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	615.80	0.12	100 (0.00)	100 (0.00)	740.69	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.30	0.01	70.40 (0.80)	20.59 (3.17)	0.50	0.00	51.40 (0.76)	6.99 (4.91)				
$Lasso_{2,TMEV}$	0.30	0.01	100 (0.00)	100 (0.00)	638748.00	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	576.94	0.12	100 (0.00)	100 (0.00)	691.39	0.01	100 (0.00)	100 (0.00)				
$\sigma_u^2 = 1$												
OLS_{naive}	1.05	0.03			5.10	0.00						
OLS_{TMEV}	409337.00	-0.19			100505.00	0.01						
OLS_{EMEV}	58.80	-0.01			31.75	0.00						
$Lasso_{1,naive}$	0.80	0.02	71.68 (0.87)	39.44 (3.36)	1.00	0.01	48.84 (0.87)	12.13 (6.31)				
$Lasso_{1,TMEV}$	107702.00	0.07	100 (0.00)	100 (0.00)	433724.00	0.01	100 (0.00)	100 (0.00)				
$Lasso_{1,EMEV}$	18.86	-0.01	100 (0.00)	100 (0.00)	3.91	0.00	100 (0.00)	100 (0.00)				
$Lasso_{2,naive}$	0.80	0.02	71.88 (0.87)	39.36 (3.11)	0.99	0.01	48.60 (0.86)	11.95 (5.06)				
$Lasso_{2,TMEV}$	84104.60	0.06	100 (0.00)	100 (0.00)	450765.00	0.01	100 (0.00)	100 (0.00)				
$Lasso_{2,EMEV}$	19.31	-0.01	100 (0.00)	100 (0.00)	2.80	0.00	100 (0.00)	100 (0.00)				

Table 46: MSE, bias, TPF and FPF for β_3 (n=500, ρ = 0.5, ϵ = 2.5)

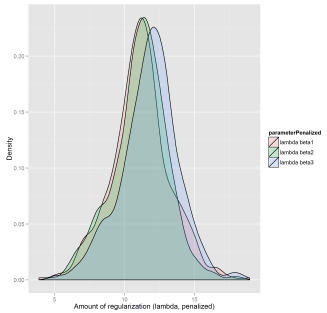
A.2.4 Density plots for the tuning parameter λ (n=500)



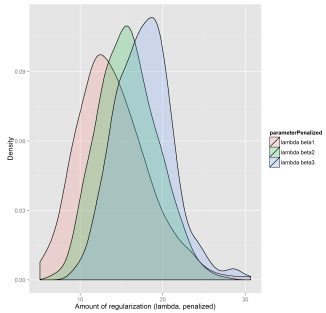
ϵ=1, ρ=0.1, σ_u^2 = 0.01



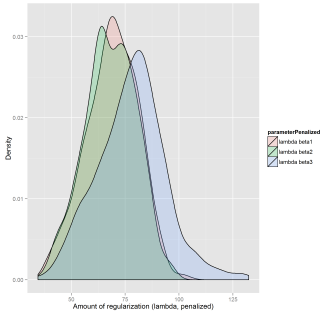
ϵ=1, ρ=0.1, σ_u^2 = 1



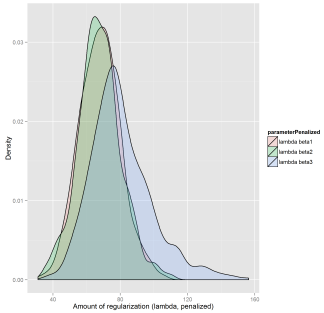
ϵ=1, ρ=0.9, σ_u^2 =0.01



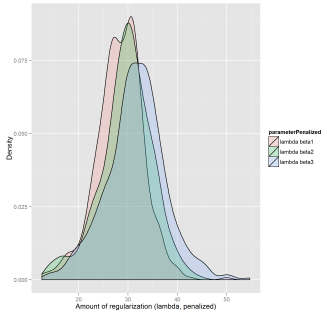
ϵ=1, ρ=0.9, σ_u^2 = 1



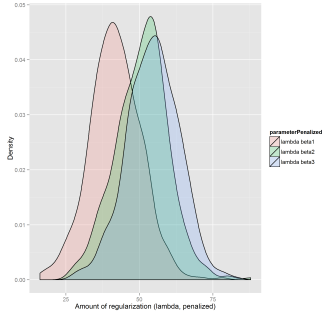
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



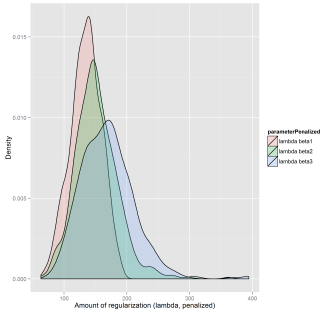
ϵ=2.5, ρ=0.1, σ_u^2 = 1



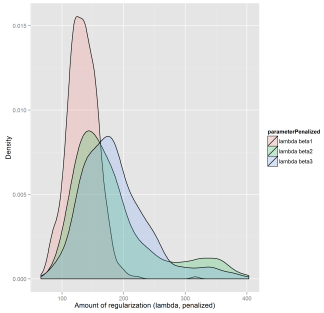
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



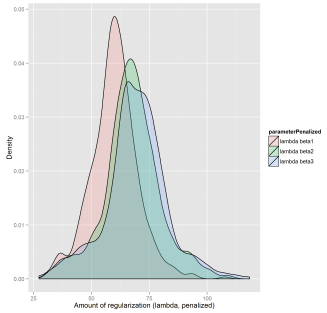
ϵ=2.5, ρ=0.9, σ_u^2 = 1



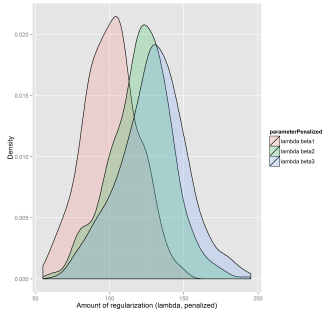
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1

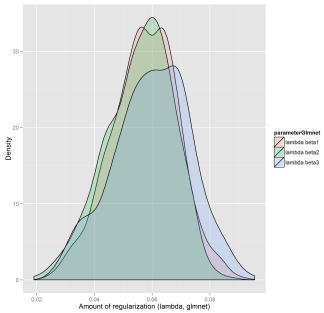


ϵ=5, ρ=0.9, σ_u^2 = 0.01

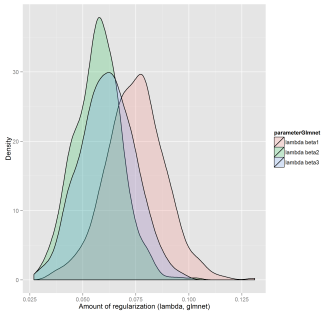


ϵ=5, ρ=0.9, σ_u^2 = 1

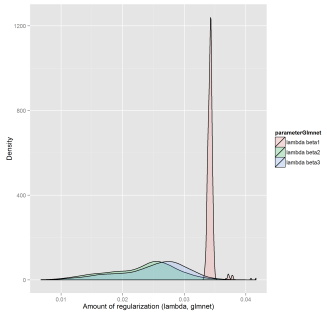
Figure 13: Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=500, p=50)



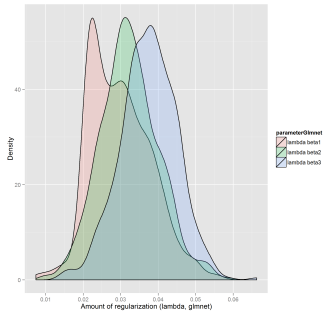
ϵ=1, ρ=0.1, σ_u^2 = 0.01



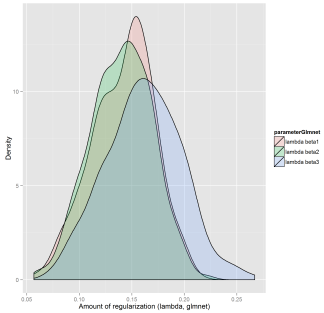
ϵ=1, ρ=0.1, σ_u^2 = 1



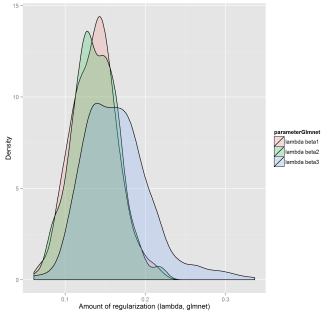
ϵ=1, ρ=0.9, σ_u^2 =0.01



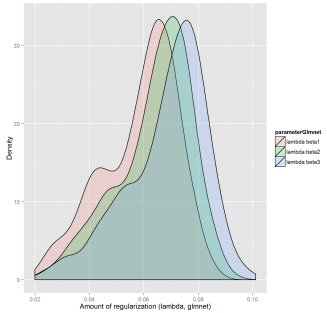
ϵ=1, ρ=0.9, σ_u^2 = 1



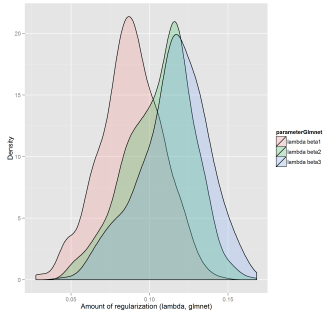
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



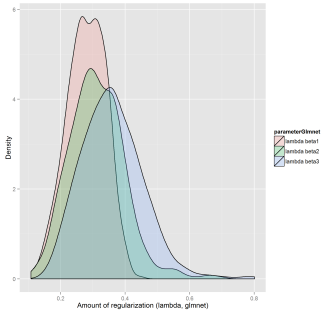
ϵ=2.5, ρ=0.1, σ_u^2 = 1



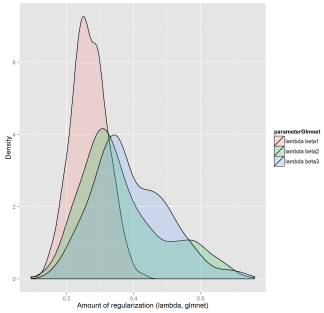
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



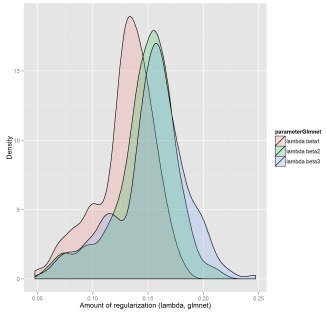
ϵ=2.5, ρ=0.9, σ_u^2 = 1



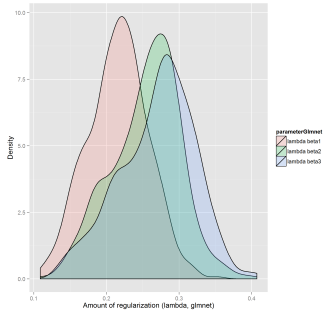
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1

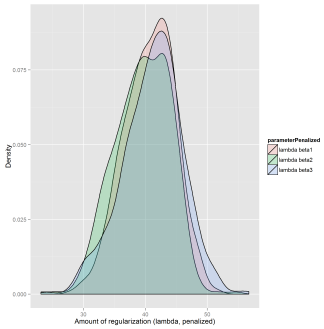


ϵ=5, ρ=0.9, σ_u^2 = 0.01

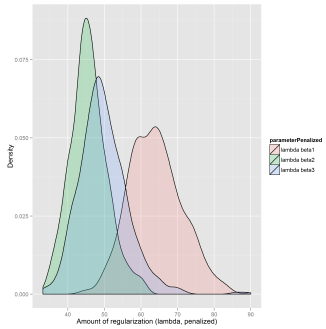


ϵ=5, ρ=0.9, σ_u^2 = 1

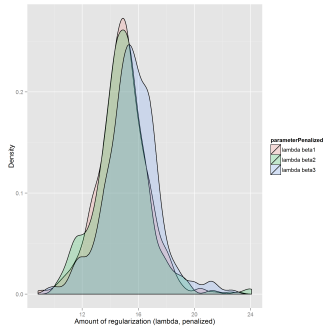
Figure 14: Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=500, p=50)



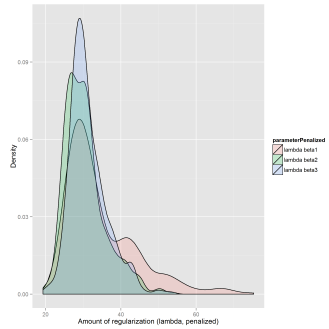
ϵ=1, ρ=0.1, σ_u^2 = 0.01



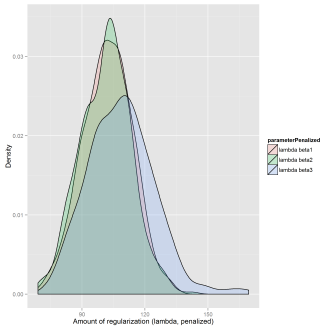
ϵ=1, ρ=0.1, σ_u^2 = 1



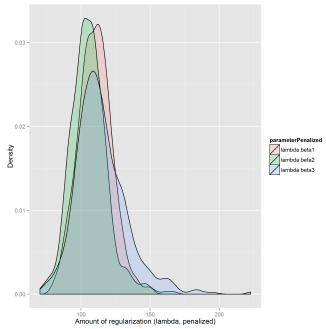
ϵ=1, ρ=0.9, σ_u^2 =0.01



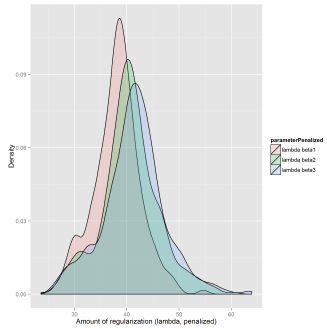
ϵ=1, ρ=0.9, σ_u^2 = 1



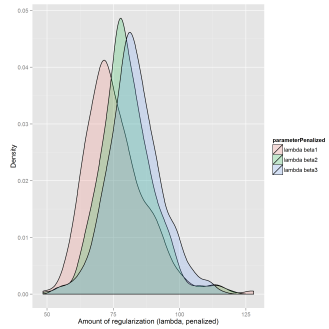
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



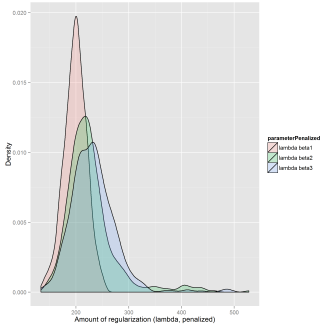
ϵ=2.5, ρ=0.1, σ_u^2 = 1



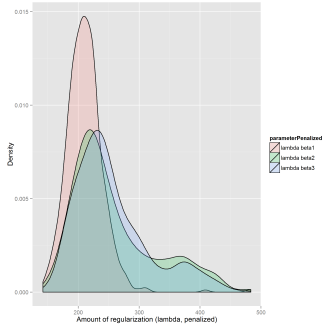
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



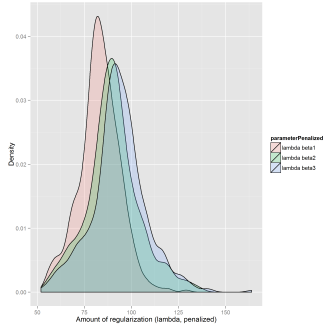
ϵ=2.5, ρ=0.9, σ_u^2 = 1



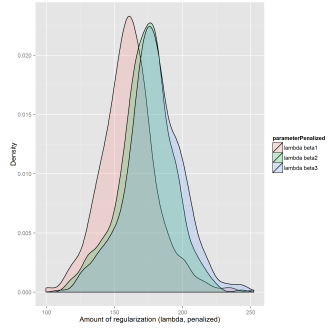
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1

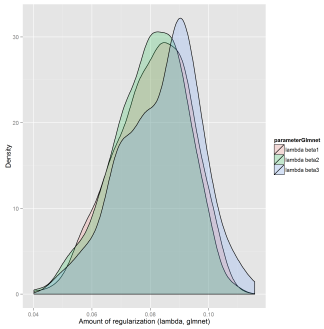


ϵ=5, ρ=0.9, σ_u^2 = 0.01

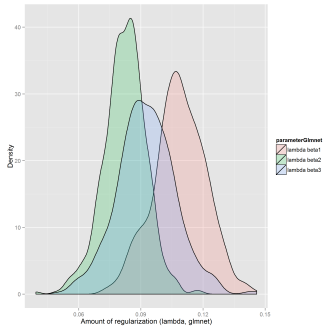


ϵ=5, ρ=0.9, σ_u^2 = 1

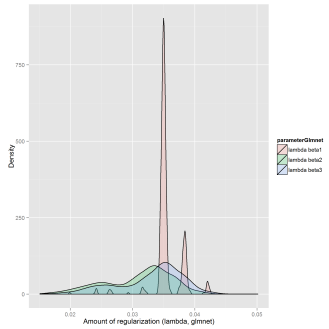
Figure 15: Density plots for the tuning parameter $\lambda_{\texttt{OptL1}}$(n=500, p=250)



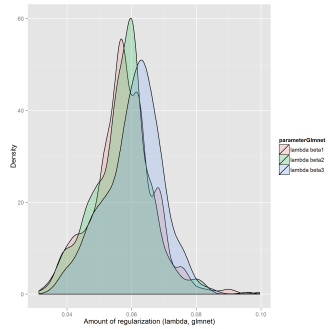
ϵ=1, ρ=0.1, σ_u^2 = 0.01



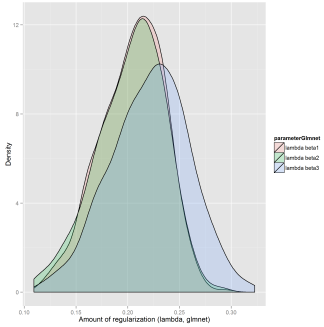
ϵ=1, ρ=0.1, σ_u^2 = 1



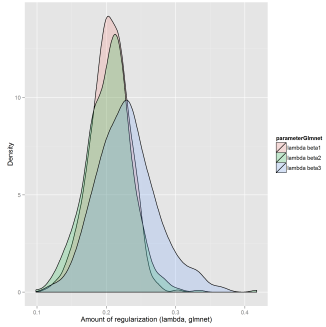
ϵ=1, ρ=0.9, σ_u^2 =0.01



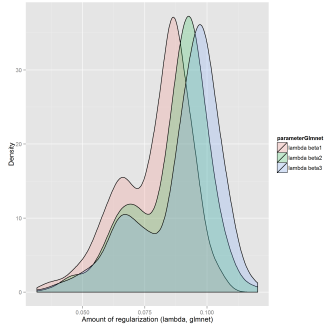
ϵ=1, ρ=0.9, σ_u^2 = 1



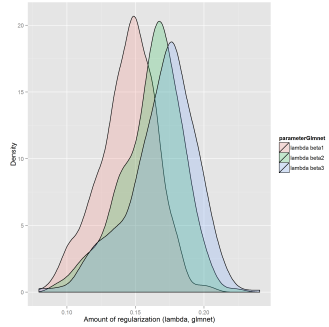
ϵ=2.5, ρ=0.1, σ_u^2 =0.01



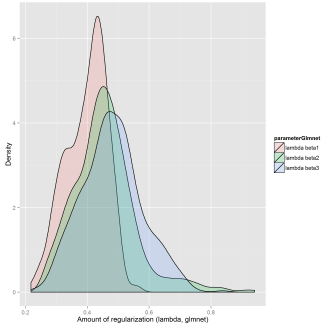
ϵ=2.5, ρ=0.1, σ_u^2 = 1



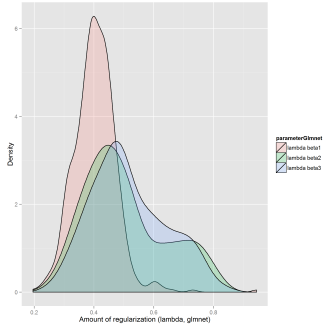
ϵ=2.5, ρ=0.9, σ_u^2 = 0.01



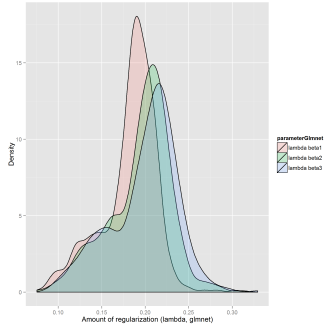
ϵ=2.5, ρ=0.9, σ_u^2 = 1



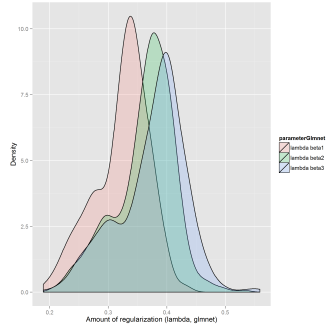
ϵ=5, ρ=0.1, σ_u^2 = 0.01



ϵ=5, ρ=0.1, σ_u^2 = 1



ϵ=5, ρ=0.9, σ_u^2 = 0.01



ϵ=5, ρ=0.9, σ_u^2 = 1

Figure 16: Density plots for the tuning parameter $\lambda_{\texttt{cv.glmnet}}$ (n=500, p=250)

A.2.5 Comparison of $LASSO_{\texttt{cv.glmnet}}$ and $LASSO_{\texttt{OptL1}}$

	$n = 100$			$n = 500$	
	$p = 10$	$p = 50$	$p = 150$	$p = 50$	$p = 250$
$rho = 0.1, \sigma_u^2 = 0.01$					
$Lasso_{naive}$	0.014	0.008	0.005	0.002	0.001
$Lasso_{TMEV}$	0.014	0.009	0.008	0.002	0.001
$Lasso_{EMEV}$	0.091	0.125	0.175	0.051	0.041
$rho = 0.1, \sigma_u^2 = 1$					
$Lasso_{naive}$	0.020	0.015	0.012	0.002	0.001
$Lasso_{TMEV}$	0.261	0.869	0.617	0.087	0.113
$Lasso_{EMEV}$	0.002	0.009	0.029	0.001	0.001
$rho = 0.9, \sigma_u^2 = 0.01$					
$Lasso_{naive}$	0.010	0.025	0.003	0.003	0.001
$Lasso_{TMEV}$	0.011	0.902	0.043	0.004	0.013
$Lasso_{EMEV}$	0.009	0.016	0.022	0.012	0.003
$rho = 0.9, \sigma_u^2 = 1$					
$Lasso_{naive}$	0.004	0.012	0.023	0.001	0.001
$Lasso_{TMEV}$	0.062	0.410	0.859	0.031	0.091
$Lasso_{EMEV}$	0.165	0.006	0.048	0.001	0.001

Table 47: Mean absolute difference between $LASSO_{\texttt{cv.glmnet}}$ and $LASSO_{\texttt{penalized}}$ for β_3 ($\epsilon = 1$)

	$n = 100$			$n = 500$	
	$p = 10$	$p = 50$	$p = 150$	$p = 50$	$p = 250$
$rho = 0.1, \sigma_u^2 = 0.01$					
$Lasso_{naive}$	0.014	0.009	0.005	0.002	0.001
$Lasso_{TMEV}$	0.015	0.009	0.008	0.002	0.001
$Lasso_{EMEV}$	0.073	0.121	0.171	0.022	0.038
$rho = 0.1, \sigma_u^2 = 1$					
$Lasso_{naive}$	0.013	0.010	0.007	0.002	0.001
$Lasso_{TMEV}$	0.236	0.577	0.418	0.045	0.068
$Lasso_{EMEV}$	0.001	0.004	0.019	0.000	0.001
$rho = 0.9, \sigma_u^2 = 0.01$					
$Lasso_{naive}$	0.017	0.016	0.005	0.004	0.001
$Lasso_{TMEV}$	0.019	0.646	0.059	0.004	0.019
$Lasso_{EMEV}$	0.037	0.007	0.030	0.003	0.005
$rho = 0.9, \sigma_u^2 = 1$					
$Lasso_{naive}$	0.003	0.006	0.007	0.001	0.000
$Lasso_{TMEV}$	0.036	0.278	0.219	0.044	0.063
$Lasso_{EMEV}$	0.006	0.003	0.010	0.000	0.001

Table 48: Mean absolute difference between $LASSO_{\texttt{cv.glmnet}}$ and $LASSO_{\texttt{penalized}}$ for $\boldsymbol{\beta}_2$ ($\epsilon = 1$)

	$n = 100$			$n = 500$	
	$p = 10$	$p = 50$	$p = 150$	$p = 50$	$p = 250$
$rho = 0.1, \sigma_u^2 = 0.01$					
$Lasso_{naive}$	0.015	0.006	0.003	0.002	0.001
$Lasso_{TMEV}$	0.015	0.007	0.006	0.002	0.001
$Lasso_{EMEV}$	0.062	0.010	0.154	0.031	0.051
$rho = 0.1, \sigma_u^2 = 1$					
$Lasso_{naive}$	0.014	0.006	0.004	0.002	0.001
$Lasso_{TMEV}$	0.219	0.334	0.409	0.525	0.049
$Lasso_{EMEV}$	0.002	0.001	0.020	0.001	0.001
$rho = 0.9, \sigma_u^2 = 0.01$					
$Lasso_{naive}$	0.015	0.012	0.005	0.003	0.001
$Lasso_{TMEV}$	0.018	0.870	0.059	0.003	0.014
$Lasso_{EMEV}$	0.012	0.004	0.030	0.004	0.004
$rho = 0.9, \sigma_u^2 = 1$					
$Lasso_{naive}$	0.003	0.004	0.004	0.001	0.000
$Lasso_{TMEV}$	0.059	0.184	0.183	0.039	0.010
$Lasso_{EMEV}$	0.018	0.002	0.006	0.000	0.000

Table 49: Mean absolute difference between $LASSO_{\texttt{cv.glmnet}}$ and $LASSO_{\texttt{penalized}}$ for $\boldsymbol{\beta}_3$ ($\epsilon = 1$)

Declaration of Authenticity

The work contained in this thesis is original and has not been previously submitted for examination which has led to the award of a degree. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made.

Place, Date	Signature
München, den 5.11.2015	Kathrin Jahreis