# Bridge estimators and the adaptive Lasso under heteroscedasticity


Jens Wagener
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
e-mail: jens.wagener@rub.de

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
e-mail: holger.dette@rub.de


March 4, 2011


**Abstract**

In this paper we investigate penalized least squares methods in linear regression models with heteroscedastic error structure. It is demonstrated that the basic properties with respect to model selection and parameter estimation of bridge estimators, Lasso and adaptive Lasso do not change if the assumption of homoscedasticity is violated. However, these estimators do not have oracle properties in the sense of Fan and Li (2001). In order to address this problem we introduce weighted penalized least squares methods and demonstrate their advantages by asymptotic theory and by means of a simulation study.



Keywords and Phrases: Lasso, adaptive Lasso, bridge estimators, heteroscedasticity, asymptotic normality, conservative model selection, oracle property.


## 1 Introduction

Penalized least squares and penalized likelihood estimators have received much interest by many authors over the last 15 years because they provide an attractive methodology to select variables and estimate parameters in sparse linear models of the form

$$Y_i = x_i^T \beta_0 + \varepsilon_i \ , \ i = 1, \ldots, n \tag{1.1}$$

where $Y_i \in \mathbb{R}$, $x_i$ is a $p$-dimensional covariate, $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^T$ the unknown (sparse) vector of parameters and the $\varepsilon_i$ are i.i.d. random variables. Frank and Friedman (1993) introduced the

so called 'bridge regression', which shrinks the estimates of the parameters in the (1.1) towards 0 using an objective function penalized by the $L^q$-norm ($q > 0$), that is

$$L(\beta) = \sum_{i=1}^{n}(Y_i - x_i^T\beta)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|^q \tag{1.2}$$

The celebrated Lasso (Tibshirani (1996)) corresponds to a bridge estimator with $q = 1$. Knight and Fu (2000) investigated the asymptotic behavior of bridge estimators in linear regression models. They established asymptotic normality of the estimators of the non-zero components of the parameter vector and showed that the bridge estimators set some parameters exactly to 0 with positive probability for $0 < q \leq 1$. This means that the estimators perform model selection and parameter estimation in a single step. In recent years many procedures with the latter property have been proposed in addition to bridge regression: the non-negative Garotte (Breiman (1995)), the SCAD (Fan and Li (2001)), least angle regression (Efron et al. (2004)), the elastic net (Zou and Hastie (2005)), the adaptive Lasso (Zou (2006)) or the Dantzig selector (Candes and Tao (2007)), which has similar properties as Lasso (James et al. (2009)). All aforementioned procedures have the attractive feature that model selection and parameter estimation can be achieved by a single minimization problem with computational cost growing polynomially with the sample size, while classical subset selection via an information criterion like AIC (Akaike (1973)), BIC (Schwarz (1978)) or FIC (Claeskens and Hjort (2003)) has exponentially growing computational cost. Moreover, there exist efficient algorithms to solve these minimization problems like LARS (Efron et al. (2004)) or DASSO (James et al. (2009)).

Fan and Li (2001) argued that any reasonable estimator should be unbiased, continuous in the data, should estimate zero parameters as exactly zero with probability converging to one (consistency for model selection) and should have the same asymptotic variance as the ideal estimator in the correct model. They called this the 'oracle property' of an estimator, because such an estimator is asymptotically (point-wise) as efficient as an estimator which is assisted by a model selection oracle. In particular they proved the oracle property for the SCAD penalty. Knight and Fu (2000) showed that for $0 < q < 1$ the bridge estimator has the oracle property using a particular tuned parameter $\lambda_n$, while Zou (2006) demonstrated that the Lasso can not have it. This author showed the oracle property for the adaptive Lasso, which determines the estimator minimizing the objective function

$$L(\beta) = \sum_{i=1}^{n}(Y_i - x_i^T\beta)^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma} \tag{1.3}$$

(here $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ denotes a preliminary estimate of $\beta_0$). Fan and Peng (2004), Kim et al. (2008), Huang et al. (2008a) and Huang et al. (2008b) showed generalizations of the aforementioned results in the case where the number of parameters is increasing with the sample size.

The purpose of the present paper is to consider penalized least squares regression under some 'non standard' conditions. To our knowledge most of the literature concentrates on models of the form (1.1) with independent identically distributed errors and in this note we consider the corresponding problem in the case of heteroscedastic errors. We concentrate our analysis on the case of a fixed parameter dimension and generalize the asymptotic results of Knight and Fu (2000) and Zou (2006) for bridge estimators and the adaptive Lasso (we do not analyze bridge estimators for $q > 1$ because these do not perform model selection). In the next section we present the model, the estimators and introduce some notation. In Section 3 we derive the asymptotic properties of bridge estimators and the adaptive Lasso under heteroscedasticity, first if the estimators are tuned to conservative model selection and second if they are tuned to consistent model selection. However these estimators do not have the oracle property. Therefore, in Section 4 we introduce weighted penalized least squares methods, derive the asymptotic properties of the corresponding estimators and establish oracle properties for the bridge estimator with $0 < q < 1$ and the adaptive Lasso. In Section 5 we illustrate the differences between procedures which do not take heteroscedasticity into account and the methods proposed in this paper by means of a simulation study and a data example. Finally, some technical details are given in an appendix.

# 2 Preliminaries

We consider the following (heteroscedastic) linear regression model

$$Y = X\beta_0 + \Sigma(\beta_0)\varepsilon, \tag{2.1}$$

where $Y = (Y_1, \ldots, Y_n)^T$ is an $n-$dimensional vector of observed random variables, $X$ is a $n \times p$-matrix of covariates, $\beta_0$ is a $p$-dimensional vector of unknown parameters, $\Sigma(\beta_0) = diag(\sigma(x_1, \beta_0), \ldots, \sigma(x_n, \beta_0))$ is a positive definite matrix, $x_1^T, \ldots, x_n^T$ denote the rows of the matrix $X$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is a vector of independent random variables with $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = 1$ for $i = 1, \ldots, n$. We assume that the model is sparse in the sense that that $\beta_0 = (\beta_0(1)^T, \beta_0(2)^T)^T$, where $\beta_0(1) \in \mathbb{R}^k$ and $\beta_0(2) = 0 \in \mathbb{R}^{p-k}$, but it is not known which components of the vector $\beta_0$ vanish. Without loss of generality it is assumed that the $k$ nonzero components are given by the vector $\beta_0(1)^T$. The matrix of covariates is partitioned according to $\beta_0$, that is $X = (X(1), X(2))$, where $X(1) \in \mathbb{R}^{n \times k}$ and $X(2) \in \mathbb{R}^{n \times (p-k)}$. The rows of $X(j)$ are denoted by $x_1(j)^T, \ldots, x_n(j)^T$ for $j = 1, 2$. We assume that the matrix $X$ is not random but note that for random covariates all results presented in this paper hold conditionally on $X$.

In the following we will investigate the estimators of the form

$$\widehat{\beta}_{lse} = \text{argmin}_\beta \Big[ \sum_{i=1}^{n} (Y_i - x_i^T \beta)^2 + \lambda_n P(\beta, \tilde{\beta}) \Big]$$

$$\widehat{\beta}_{wlse} = \text{argmin}_\beta \Big[ \sum_{i=1}^{n} \left( \frac{Y_i - x_i^T \beta}{\sigma(x_i, \overline{\beta})} \right)^2 + \lambda_n P(\beta, \tilde{\beta}) \Big] \tag{2.2}$$

where $\overline{\beta}$ and $\tilde{\beta}$ denote preliminary estimates of the parameter $\beta_0$ and $P(\beta, \tilde{\beta})$ is a penalty function. We are particularly interested in the cases

$$P(\beta, \tilde{\beta}) = P(\beta) = \|\beta\|_q^q \quad (0 < q \leq 1)$$

$$P(\beta, \tilde{\beta}) = \sum_{j=1}^{p} |\beta_j| |\tilde{\beta}_j|^{-\gamma} \quad (\gamma > 0)$$

corresponding to bridge regression (with the special case of Lasso for $q = 1$) and the the adaptive Lasso , respectively. The subscripts 'lse' and 'wlse' correspond to 'ordinary' and 'weighted' least squares regression, respectively. Note that for bridge regression with $q < 1$ the functions minimized above are not convex in $\beta$ and there may exist multiple minimizing values. In that case the argmin is understood as an arbitrary minimizing value and all results stated here are valid for any such value.

Throughout this paper an estimator $\widehat{\beta}$ of the parameter $\beta_0$ in model (2.1) is called consistent for model selection, if

$$\lim_{n \to \infty} P(\widehat{\beta}_j = 0) = 1 \quad \text{for all } j > k \tag{2.3}$$

and $\widehat{\beta}$ performs conservative model selection, if

$$\lim_{n \to \infty} P(\widehat{\beta}_j = 0) = c \quad \text{for all } j > k \tag{2.4}$$

for some constant $0 < c < 1$ (see e.g. Leeb and Pötscher (2005)). If an estimator performs consistent or conservative model selection, respectively, depends on the choice of the tuning parameter $\lambda_n$. A 'larger' value of $\lambda_n$ usually yields consistent model selection while a 'smaller' value yields conservative model selection. In the following sections we will present results for both cases of tuning separately.

**Remark 2.1** In practice the parameter $\lambda_n$ has to be chosen by a data driven procedure. If the main purpose of the data analysis is the estimation of the parameters, $\lambda_n$ should be chosen to perform conservative model selection. This can be achieved by using cross-validation or generalized

cross-validation (Craven and Wahba (1979)) which is asymptotically equivalent to AIC (see e.g. Shao (1997) or Wang et al. (2007)). If the main purpose of the data analysis is the identification of the relevant covariates, the regularizing parameter $\lambda_n$ should be chosen to perform consistent model selection, which can be achieved minimizing a BIC-like criterion (compare Wang et al. (2007)).

# 3 Penalized least squares estimation

In this section we study the asymptotic behavior of the un-weighted estimators $\widehat{\beta}_{lse}$ in the linear regression model with heteroscedastic errors (2.1). In particular we extend the results obtained by Knight and Fu (2000) and Zou (2006) to the case of heteroscedasticity. Throughout this paper we will use the notation $\operatorname{sgn}(x)$ for the sign of $x \in \mathbb{R}$ with the convention $\operatorname{sgn}(0) = 0$. For a vector $v \in \mathbb{R}^p$ and a function $f : \mathbb{R} \to \mathbb{R}$ we write $f(v) = (f(v_1), \ldots, f(v_p))^T$ and all inequalities between vectors are understood componentwise. By $\mathbb{1}_p$ we denote a $p$-dimensional vector with all elements equal to 1. Our basic assumptions for the asymptotic analysis in this section are the following.

(i) The design matrix satisfies

$$\frac{1}{n} X^T X \to C > 0,$$

where the limit

$$C = \begin{pmatrix} C_{11} & C_{21}^T \\ C_{21} & C_{22} \end{pmatrix}$$

is partitioned according to $X$ (that is $C_{11} \in \mathbb{R}^{k \times k}, C_{22} \in \mathbb{R}^{(p-k)\times(p-k)}$).

(ii)

$$\frac{1}{n} X^T \Sigma(\beta_0)^2 X \to B > 0,$$

where the matrix $B$ is partitioned in the same way as the matrix $C$.

(iii)

$$\frac{1}{n} \max_{1 \le i \le n} x_i^T \sigma(x_i, \beta_0)^2 x_i \to 0.$$

The first two assumptions are posed in order to obtain positive definite limiting covariance matrices of the estimators. The third is needed for the Lindeberg condition to hold.

## 3.1 Conservative model selection

Leeb and Pötscher (2008) showed that an estimator which performs consistent model selection must have an unbounded (scaled) risk function, while the optimal estimator in the true model has

a bounded risk. Pötscher and Leeb (2009) showed that the asymptotic distribution of the Lasso and the SCAD can not be consistently estimated uniformly over the parameter space. The problems arising from this phenomenon are more pronounced for estimators tuned to consistent model selection. Therefore the (global) asymptotic behaviour of a penalized least squares estimator is different from that of an estimator in the true model although it satisfies an 'oracle property' in the sense of Fan and Li (2001). Estimators which do perform conservative (but not consistent) model selection in the sense of (2.4) do not suffer from the drawback of an unbounded risk and estimators of the corresponding asymptotic distribution function are better than those for the asymptotic distribution of estimators tuned to consistent model selection. For these reasons we first study the behavior of the Lasso, the bridge regression and the adaptive Lasso estimator tuned to conservative model selection. The following result will be proved in the Appendix.

**Lemma 3.1** *Let the basic assumptions (i)-(iii) be satisfied. If* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$*, then the Lasso estimator* $\widehat{\beta}_{lse}$ *satisfies*

$$\sqrt{n}(\widehat{\beta}_{lse} - \beta_0) \xrightarrow{\mathcal{D}} \operatorname{argmin}(V), \tag{3.1}$$

*where the function* $V$ *is given by*

$$V(u) = -2u^T W + u^T C u + \lambda_0 \sum_{j=1}^{k} u_j \operatorname{sgn}(\beta_{0,j}) + \lambda_0 \sum_{j=k+1}^{p} |u_j| \tag{3.2}$$

*and* $W \sim \mathcal{N}(0, B)$.

**Remark 3.2** Let $u = (u(1)^T, u(2)^T)^T$ with $u(1) \in \mathbb{R}^k$ and $u(2) \in \mathbb{R}^{p-k}$ and introduce a similar decomposition for the random variable $W = (W(1)^T, W(2)^T)^T$ defined in Lemma 3.1. Then by the Karush-Kuhn-Tucker (KKT) conditions the function $V$ defined in (3.2) is minimized at $\widehat{u} = (\widehat{u}(1)^T, 0)^T$ if and only if

$$\widehat{u}(1) = C_{11}^{-1}(W(1) - \lambda_0 \operatorname{sgn}(\beta_0(1))/2) \sim \mathcal{N}(-C_{11}^{-1}\lambda_0 \operatorname{sgn}(\beta_0(1))/2, C_{11}^{-1} B_{11} C_{11}^{-1})$$

and

$$-\lambda_0/2\mathbb{1}_{p-k} < C_{21}\widehat{u}(1) - W(2) < \lambda_0/2\mathbb{1}_{p-k}.$$

This yields that there is a positive probability that the Lasso estimates zero components of the parameter vector as exactly zero if $\lambda_0 \neq 0$, but this probability is usually strictly less than 1. Consequently, under heteroscedasticity the Lasso performs conservative model selection in the same way as in the homoscedastic case. The asymptotic covariance matrix of the Lasso estimator of the non-zero parameters is given by $C_{11}^{-1} B_{11} C_{11}^{-1}$ which is the same as for the ordinary least squares (OLS) estimator in heteroscedastic linear models. This covariance is not the best one achievable, because under heteroscedasticity the OLS estimator is dominated by a generalized LS estimator. Additionally the estimator is biased if $\lambda_0 \neq 0$.

**Lemma 3.3** *Let the basic assumptions (i)-(iii) be satisfied and assume that* $q \in (0,1)$.

*(1) If* $\lambda_n/n^{q/2} \to \lambda_0 \geq 0$, *then the bridge estimator* $\widehat{\beta}_{lse}$ *satisfies (3.1) where the function* $V$ *is given by*

$$V(u) = -2u^T W + u^T C u + \lambda_0 \sum_{j=k+1}^{p} |u_j|^q$$

*and* $W = (W(1)^T, W(2)^T)^T \sim \mathcal{N}(0, B)$.

*(2) Assume additionally that there exists a sequence* $0 < a_n \to \infty$, *such that the preliminary estimate* $\tilde{\beta}$ *is a continuous function of all data points and satisfies*

$$a_n(\tilde{\beta} - \beta_0) \xrightarrow{\mathcal{D}} Z,$$

*where* $Z$ *denotes a random vector with components having no pointmass in 0. If*

$$\lambda_n a_n^{\gamma}/\sqrt{n} \to \lambda_0 \geq 0 \tag{3.3}$$

*then the adaptive Lasso estimator* $\widehat{\beta}_{lse}$ *satisfies (3.1) where the function* $V$ *is given by*

$$V(u) = -2u^T W + u^T C u + \lambda_0 \sum_{j=k+1}^{p} |Z_j|^{-\gamma} |u_j|$$

*and* $W \sim \mathcal{N}(0, B)$.

**Remark 3.4**

(1) With the same notation as in Remark 3.2 we obtain from the KKT conditions that the function $V$ in part (1) of Lemma 3.3 is minimized at $(\widehat{u}(1)^T, 0)^T$ if and only if

$$\widehat{u}(1) = C_{11}^{-1} W(1) \sim \mathcal{N}(0, C_{11}^{-1} B_{11} C_{11}^{-1})$$

and for each small $\delta > 0$

$$-q\lambda_0 \delta^{q-1}/2\mathbb{1}_{p-k} < C_{21}\widehat{u}(1) - W(2) < \lambda_0 \delta^{q-1}/2\mathbb{1}_{p-k}.$$

Thus the bridge estimator also performs conservative model selection, whenever $\lambda_0 > 0$. Again the asymptotic covariance matrix is given by $C_{11}^{-1} B_{11} C_{11}^{-1}$ and is suboptimal, but in contrast to the Lasso estimator the estimator is unbiased.

(2) The canonical choice of $\tilde{\beta}$ in the second part of Lemma 3.3 is the OLS estimator if $p < n$. In this case we have $a_n = \sqrt{n}$ (the best rate which is achievable) and $Z \sim \mathcal{N}(0, C^{-1}BC^{-1})$. In addition we have $Z = W$, because both random vectors are obtained as limits of the same quantity. Consequently, if $\gamma = 1$ the condition (3.3) becomes $\lambda_n \to \lambda_0$.
The function $V$ defined in the second part of Lemma 3.3 is minimized at $(\widehat{u}(1)^T, 0)^T$ if and only if

$$\widehat{u}(1) = C_{11}^{-1} W_1 \sim \mathcal{N}(0, C_{11}^{-1} B_{11} C_{11}^{-1})$$

and

$$-\lambda_0/2|Z(2)|^{-\gamma} < C_{21}\widehat{u}(1) - W(2) < \lambda_0/2|Z(2)|^{-\gamma}.$$

If $\lambda_0$ is not too small and $\gamma$ not too large this event has positive probability strictly less than one. So the adaptive Lasso performs conservative model selection but again the asymptotic covariance matrix is not the optimal one. Similar to bridge estimators with $0 < q < 1$ the adaptive Lasso estimator is asymptotically unbiased.

In finite samples there may be more parameters than observations, that is $p \geq n$. In this case the OLS estimator is no longer available but the estimator $\widehat{\beta}_{lse}$ in the second part of Lemma 3.3 can still be calculated. For this purpose one could use a penalized least squares estimator like a bridge estimator with $q \leq 1$ tuned to perform conservative model selection as preliminary estimate $\tilde{\beta}$. For the final calculation of $\hat{\beta}_{lse}$ we use the convention that $\beta_j$ is set to 0 and is removed from the penalty function of the adaptive Lasso if $\tilde{\beta}_j = 0$.

## 3.2 Consistent model selection

In this section we use a different tuning parameter $\lambda_n$ in order to obtain consistency in model selection of the considered estimators. Again our first result concerns the Lasso estimator. This result is a generalization of Lemma 3 of Zou (2006). The proof follows along the same lines as in the homoscedastic case and is therefore omitted.

**Lemma 3.5** *Let the basic assumptions (i)-(iii) be satisfied and additionally $\lambda_n/n \to 0$, $\lambda_n/\sqrt{n} \to \infty$. Then the Lasso estimator $\widehat{\beta}_{lse}$ satisfies*

$$\frac{n}{\lambda_n}(\widehat{\beta}_{lse} - \beta_0) \xrightarrow{P} \operatorname{argmin}(V),$$

*where the function $V$ is given by*

$$V(u) = u^T C u + \sum_{j=1}^{k} u_j \operatorname{sgn}(\beta_{0,j}) + \sum_{j=k+1}^{p} |u_j|.$$

**Remark 3.6** The function $V$ defined in Lemma 3.5 is minimized in $(\widehat{u}(1), 0)$ if and only if

$$\widehat{u}(1) = -C_{11}^{-1}\text{sgn}(\beta_0(1))/2$$

and

$$-\mathbb{1}_{p-k} < 2C_{21}\widehat{u}(1) < \mathbb{1}_{p-k}.$$

The second condition is equivalent to

$$|C_{21}C_{11}^{-1}\text{sgn}(\beta_0(1))| < \mathbb{1}_{p-k}, \tag{3.4}$$

which is the so called strong irrepresentable condition (compare e.g. Zhao and Yu (2006)). So Lemma 3.5 directly yields that in the case of heteroscedasticity the Lasso estimator is consistent for model selection if the strong irrepresentable condition (3.4) is satisfied. Moreover, the Lasso estimator is still consistent for parameter estimation but not with the optimal rate $\sqrt{n}$. This means that the results from the homoscedastic case can be extended in a straightforward manner.

The following lemma presents the asymptotic properties of bridge estimators for $0 < q < 1$ and the adaptive Lasso. The proof follows by similar arguments as given in Knight and Fu (2000) and Zou (2006) and is omitted for the sake of brevity.

**Lemma 3.7** *Let the basic assumptions (i)-(iii) be satisfied.*

*(1) If $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n/n^{q/2} \to \infty$, then the bridge estimator $\widehat{\beta}_{lse}$ satisfies (3.1), where the function $V$ is given by*

$$V(u) = V(u(1), u(2)) = \begin{cases} -2u(1)^T W(1) + u(1)^T C_{11} u(1) & \text{if } u(2) = 0, \\ \infty & \text{otherwise}, \end{cases} \tag{3.5}$$

*and $W(1) \sim \mathcal{N}(0, B_{11})$.*

*(2) If $\tilde{\beta}$ is a preliminary estimator of $\beta_0$ such that there exists a sequence $0 < a_n \to \infty$ with $a_n(\tilde{\beta} - \beta_0) = O_p(1)$, and $\lambda_n/\sqrt{n} \to 0$, $\lambda_n/\sqrt{n}a_n^\gamma \to \infty$, then the adaptive Lasso estimator $\widehat{\beta}_{lse}$ satisfies (3.1), where the function $V$ is given by (3.5).*

**Remark 3.8** Lemma 3.7 shows that bridge and adaptive Lasso estimators are able to perform consistent model selection and estimation of the non-zero parameters with the optimal rate simultaneously. The asymptotic distribution is again normal with covariance matrix $C_{11}^{-1}B_{11}C_{11}^{-1}$. Both estimators are unbiased due to the assumption $\lambda_n/\sqrt{n} \to 0$.

# 4 Weighted penalized least squares estimators

We have seen in the last section that the asymptotic variance of the estimator $\widehat{\beta}_{lse}$ is suboptimal, because with the additional knowledge of the non-vanishing components and variances $\sigma^2(x_i^T, \beta_0)$ the best linear unbiased estimator in model (2.1) would be

$$\widehat{\beta}_{gls} = \text{argmin}_{\beta(1)} \Big[ \sum_{i=1}^{n} \Big( \frac{Y_i - x_i(1)^T \beta(1)}{\sigma(x_i, \beta_0)} \Big)^2 \Big].$$

This estimator has asymptotic variance $D_{11}^{-1}$ with $D_{11} = \lim_{n\to\infty} X(1)^T \Sigma(\beta_0)^{-2} X(1)/n$ (provided that the limit exists). In order to construct an oracle estimator we use a preliminary estimator $\overline{\beta}$ to estimate $\sigma(x_i^T, \beta_0)$ and apply a penalized least squares regression to identify the non zero components of $\beta_0$. The resulting procedure is defined in (2.2). The goal of this section is to establish its model selection properties and the asymptotic normality of the estimator of the non-zero components with covariance matrix given by $D_{11}^{-1}$.

In order to derive these results we make the following basic assumptions.

(i)' The design matrix $X$ satisfies

$$\frac{1}{n} X^T \Sigma(\beta_0)^{-2} X \to D > 0$$

where the matrix

$$D = \begin{pmatrix} D_{11} & D_{21}^T \\ D_{21} & D_{22} \end{pmatrix}$$

is partitioned according to $X(1)$ and $X(2)$ (that is $D_{11} \in \mathbb{R}^{k\times k}, D_{22} \in \mathbb{R}^{(p-k)\times(p-k)}$).

(ii)'

$$\frac{1}{n} \max_{1\le i\le n} x_i^T x_i \to 0.$$

(iii)' The variance function $\sigma(\cdot, \beta)$ is bounded away from zero for all $\beta$ in a neighborhood of $\beta_0$. Additionally $\sigma(x, \beta)$ is two times differentiable with respect to $\beta$ in a neighborhood of $\beta_0$ for all $x$ and all second partial derivatives are bounded.

(iv)' There exists a sequence $0 < b_n \to \infty$ such that $b_n/n^{1/4} \to \infty$ and such that the preliminary estimator of $\beta_0$ satisfies $b_n(\overline{\beta} - \beta_0) = O_p(1)$.

Assumptions (i)' and (ii)' are analogs of (i)-(iii) of Section 3 and are posed in order to obtain non-degenerate normal limit distributions of the estimators. (iii)' imposes sufficient smoothness of the function $\sigma$ such that $\sigma(x, \beta_0)$ can be well approximated by $\sigma(x, \overline{\beta})$. (iv)' asserts that $\overline{\beta}$ is consistent for $\beta_0$ with a reasonable rate. If $p < n$ one could use the OLS estimator for $\overline{\beta}$ and in this case we have $b_n = \sqrt{n}$. If $p \ge n$ (which may happen in finite samples) one could use the estimator $\widehat{\beta}_{lse}$ tuned to conservative model selection as shown in Lemma 3.1 and 3.3.

## 4.1 Conservative model selection

Again we first state our results for the estimator $\widehat{\beta}_{wlse}$ in the case where the tuning parameter $\lambda_n$ is chosen such that $\widehat{\beta}_{wlse}$ performs conservative model selection. We begin with a result for the scaled Lasso estimator which is proved in the Appendix.

**Theorem 4.1** *Let the basic assumptions (i)'-(iv)' be satisfied and additionally* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$. *Then the weighted Lasso estimator* $\widehat{\beta}_{wlse}$ *converges weakly, i.e.*

$$\sqrt{n}(\widehat{\beta}_{wlse} - \beta_0) \xrightarrow{\mathcal{D}} \operatorname{argmin}(V), \tag{4.1}$$

*where the function* $V$ *is given by*

$$V(u) = -2u^T W + u^T D u + \lambda_0 \sum_{j=1}^{k} u_j \operatorname{sgn}(\beta_{0,j}) + \lambda_0 \sum_{j=k+1}^{p} |u_j|$$

*and* $W \sim \mathcal{N}(0, D)$.

**Remark 4.2** By similar argument as in Remark 3.2 one obtains from Theorem 4.1 that the scaled Lasso estimator $\widehat{\beta}_{wlse}$ performs conservative model selection whenever $\lambda_0 \neq 0$. The estimators of the non zero components are asymptotically normal distributed with expectation $-D_{11}^{-1}\lambda_0 \operatorname{sgn}(\beta_0(1))/2$ and covariance matrix $D_{11}^{-1}$. So this estimator has the optimal asymptotic variance but is biased.

We now state corresponding results for scaled bridge estimators with $0 < q < 1$ and the scaled adaptive Lasso estimator.

**Theorem 4.3** *Let the basic assumptions (i)'-(iv)' be satisfied and assume that* $q \in (0, 1)$.

*(1) If* $\lambda_n/n^{q/2} \to \lambda_0 \geq 0$, *then the scaled bridge estimator* $\widehat{\beta}_{wlse}$ *satisfies (4.1) where the function* $V$ *is given by*

$$V(u) = -2u^T W + u^T D u + \lambda_0 \sum_{j=k+1}^{p} |u_j|^q$$

*and* $W \sim \mathcal{N}(0, D)$.

*(2) Let* $\tilde{\beta}$ *denote an estimator of* $\beta_0$ *that is a continuous function of all data points such that*

$$a_n(\tilde{\beta} - \beta_0) \xrightarrow{\mathcal{D}} Z.$$

*for some positive sequence $a_n \to \infty$ . If the distribution of the random vector $Z$ has no point mass in 0 and $\lambda_n a_n^\gamma/\sqrt{n} \to \lambda_0 \geq 0$, then the scaled adaptive Lasso estimator $\widehat{\beta}_{wlse}$ satisfies (4.1), where the function $V$ is given by*

$$V(u) = -2u^T W + u^T D u + \lambda_0 \sum_{j=k+1}^{p} |Z_j|^{-\gamma} |u_j|$$

*and $W \sim \mathcal{N}(0, D)$.*

Theorem 4.3 shows that the scaled bridge estimator and scaled adaptive Lasso estimator both can be tuned to perform conservative model selection. In both cases the estimators of the non-zero parameters are unbiased and asymptotically normal distributed with optimal asymptotic variance $D_{11}^{-1}$. The proof of Theorem 4.3 follows along the same lines as the one of Theorem 4.1 and is therefore omitted.

## 4.2 Consistent model selection

Finally, we provide results for weighted Lasso, bridge and adaptive Lasso estimators when tuned to perform consistent model selection. In particular, we demonstrate that in this case one obtains the optimal asymptotic covariance matrix $D_{11}^{-1}$ for the estimators of the non-zero parameters. Therefore the weighted bridge estimators and the weighted adaptive Lasso satisfy an oracle property in the sense of Fan and Li (2001). The proofs of the results are omitted, because they follow like that one of Theorem 4.1.

**Theorem 4.4** *Let the basic assumptions (i)'-(iv)' be satisfied.*

*(1) If $\lambda_n/n \to 0$, $\lambda_n/\sqrt{n} \to \infty$, then the weighted Lasso estimator $\widehat{\beta}_{wlse}$ satisfies*

$$\frac{n}{\lambda_n}(\widehat{\beta}_{wlse} - \beta_0) \xrightarrow{P} \operatorname{argmin}(V),$$

*where the function $V$ is given by*

$$V(u) = u^T D u + \lambda_0 \sum_{j=1}^{k} u_j \operatorname{sgn}(\beta_{0,j}) + \lambda_0 \sum_{j=k+1}^{p} |u_j|.$$

*(2) If $q \in (0,1)$, $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n/n^{q/2} \to \infty$, then the weighted bridge estimator $\widehat{\beta}_{wlse}$ satisfies (4.1) where the function $V$ is given by*

$$V(u) = V(u(1), u(2)) = \begin{cases} -2u(1)^T W(1) + u(1)^T D_{11} u(1) & \text{if } u(2) = 0, \\ \infty & \text{otherwise}, \end{cases} \tag{4.2}$$

*and $W(1) \sim \mathcal{N}(0, D_{11})$.*

*(3) Let $\tilde{\beta}$ be an estimator of $\beta_0$ such that there exists a sequence $0 < a_n \to \infty$ so that $a_n(\tilde{\beta}-\beta_0) = O_p(1)$. If $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n/\sqrt{n}a_n^\gamma \to \infty$, then the weighted adaptive Lasso estimator $\widehat{\beta}_{wlse}$ satisfies (4.1) where the function $V$ is given by (4.2) and $W(1) \sim \mathcal{N}(0, D_{11})$.*

**Remark 4.5** As in Remark 3.4 one obtains that the weighted Lasso estimator is consistent for model selection if and only if

$$|D_{21}D_{11}^{-1}\mathrm{sgn}(\beta_0(1))| < \mathbb{1}_{p-k},$$

which corresponds to the strong irrepresentable condition (3.4) for the "classical" Lasso estimator. In particular, the weighted Lasso does not have the optimal rate for parameter estimation.

**Remark 4.6** The last theorem shows that weighted bridge estimators and the weighted adaptive Lasso estimator both can be tuned to perform consistent model selection and estimation of the non zero parameters with the optimal rate simultaneously. Moreover, the corresponding standardized estimator of the non-vanishing components is asymptotically unbiased and normal distributed with optimal covariance matrix $D_{11}^{-1}$.

# 5 Examples

In this section we compare the "classical" penalized estimates (which do not take scale information into account) with the procedures proposed in this paper by means of a small simulation study and a data example.

## 5.1 Simulation study

For the sake of brevity we concentrated on the Lasso and adaptive Lasso. These estimators can be calculated by convex optimization and we used the package "penalized" available for $R$ on http://www.R-project.org (R Development Core Team (2008)) to perform all computations.

In all examples the data were generated using a linear model (2.1). The errors $\varepsilon$ were iid standard normal and the matrix $\Sigma$ was a diagonal matrix with entries $\sigma(x_i, \beta_0)$ on the diagonal where $\sigma$ was given by one of the following functions:

$$\begin{aligned}
(a) \quad & \sigma(x_i, \beta_0) = \frac{1}{2}\sqrt{x_i^T\beta_0}, \\
(b) \quad & \sigma(x_i, \beta_0) = \frac{1}{4}|x_i^T\beta_0|, \\
(c) \quad & \sigma(x_i, \beta_0) = \frac{1}{20}\exp|x_i^T\beta_0|, \\
(d) \quad & \sigma(x_i, \beta_0) = \frac{1}{50}\exp(x_i^T\beta_0)^2.
\end{aligned}$$

**Table 1:** *Mean number of correctly zero and correctly non-zero estimated parameters in model (2.1) with* $\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)$

| | | $\sigma$ | | | |
|---|---|---|---|---|---|
| | | (a) | (b) | (c) | (d) |
| Lasso | $= 0$ | 1.67 | 3.33 | 1.58 | 2.64 |
| | $\neq 0$ | 3 | 3 | 2.99 | 3 |
| adaptive Lasso | $= 0$ | 4.51 | 4.32 | 2.95 | 4.48 |
| | $\neq 0$ | 3 | 3 | 2.95 | 3 |
| weighted Lasso | $= 0$ | 0.97 | 1.53 | 0.67 | 0.43 |
| | $\neq 0$ | 3 | 3 | 3 | 3 |
| weighted adaptive Lasso | $= 0$ | 3.97 | 4.09 | 3.29 | 3.91 |
| | $\neq 0$ | 3 | 3 | 3 | 3 |

The different factors were chosen in order to generate data with comparable variance in each of the four models. The tuning parameter $\lambda_n$ was chosen by fivefold generalized cross validation performed on a training data set. For the preliminary estimator $\tilde{\beta}$ we used the OLS estimator and for $\overline{\beta}$ the un-weighted Lasso estimator. All reported results are based on 100 simulation runs.

We considered the same scenario as investigated by Zou (2006). More precisely the design matrix was generated having independent normally distributed rows and the covariance between the $i$-th and $j$-th entry in each row was $0.5^{|i-j|}$. The sample size was given by $n = 60$

At first we considered the parameter $\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)$. The average variance in the examples (a), (b) and (d) was given by about 55 and by about 60 in example (c) in this setup (note that Zou (2006) considered variances of similar size). The model selection performance of the estimators is presented in Table 1, where we show the mean of the correctly zero and correctly non-zero estimated parameters. In the ideal case these should be 5 and 3, respectively. It can be seen from Table 1 that the adaptive Lasso always performs better model selection than the Lasso, in accordance with the asymptotic theory. The weighted Lasso performs very poor model selection in all models and the un-weighted Lasso does a better job. The model selection performance of the weighted and un-weighted adaptive Lasso are comparable. In Table 2 we present the mean squared error of the estimates for the non-vanishing components $\beta_1, \beta_2, \beta_3$. In terms of this error the weighted versions of the estimators nearly always do a (in some cases substantially) better job than their un-weighted counterparts. This is in good accordance with the theory while the poor model selection performance of the weighted Lasso is a surprising fact.

As a second example we considered the vector $\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$ in model (2.1) (the sample size is again $n = 60$). In this case the average variance in the examples (a), (b) and (d) was given by about 38 and by about 45 in example (c). The corresponding

**Table 2:** *Mean squared error of the estimators of the non-zero coefficients in model* (2.1) *with* $\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)$

| | | $\sigma$ | | | |
|---|---|---|---|---|---|
| | | (a) | (b) | (c) | (d) |
| Lasso | $\beta_1$ | 0.0308 | 0.0682 | 0.3480 | 0.0692 |
| | $\beta_2$ | 0.0306 | 0.0374 | 0.2461 | 0.0784 |
| | $\beta_3$ | 0.0322 | 0.0484 | 0.3483 | 0.1141 |
| adaptive Lasso | $\beta_1$ | 0.0293 | 0.0593 | 0.3514 | 0.0668 |
| | $\beta_2$ | 0.0330 | 0.0393 | 0.3241 | 0.1027 |
| | $\beta_3$ | 0.0285 | 0.0416 | 0.3871 | 0.1126 |
| weighted Lasso | $\beta_1$ | 0.0215 | 0.0424 | 0.1431 | 0.2004 |
| | $\beta_2$ | 0.0171 | 0.0133 | 0.0458 | 0.0174 |
| | $\beta_3$ | 0.0191 | 0.0202 | 0.1086 | 0.0780 |
| weighted adaptive Lasso | $\beta_1$ | 0.0193 | 0.0152 | 0.0944 | 0.1953 |
| | $\beta_2$ | 0.0168 | 0.0069 | 0.0293 | 0.0134 |
| | $\beta_3$ | 0.0165 | 0.0080 | 0.0864 | 0.0763 |

results for the correctly non-zero estimated parameters are presented in Table 3 (in the ideal case this should be 8) while Table 4 contains the average of the mean squared errors of the eight components of $\beta$. We see that in this example (which was also taken from Zou (2006)) model selection is not a very challenging task. In fact with variance functions (a) and (b) all estimators perform perfect model selection. In model (c) only the weighted Lasso is perfect with respect to the criterion of model selection, which is in good agreement with its conservative behaviour in the first example. The weighted adaptive Lasso is nearly perfect in this model. In the last model both versions of Lasso perfectly select the model while the adaptive versions make a few mistakes. In terms of estimation error adaptive and non-adaptive Lasso perform comparable. Again the weighted versions yield substantially better estimation errors in all scenarios under consideration.

In some circumstances it might be difficult to specify a form of the variance function. In such cases we propose to estimate the function $\sigma$ from the data $(x_i^T\bar{\beta}, Y_i)_{i=1,\dots,n}$ nonparametrically and use the weighted penalized least squares methodology on the basis of weights $\hat{\sigma}(x_i^T\bar{\beta})$ in (2.2). In order to investigate the performance of the corresponding semi-parametric estimate in such a situation we calculated $\widehat{\beta}_{wlse}$ in the two scenarios considered in the previous paragraph using a local polynomial estimate (compare e.g. Wand and Jones (1995)) of $\sigma$ instead of the true function. The bandwidth for this estimator was chosen by the direct plug-in method of Ruppert et al. (1995) and a Gaussian kernel was used. The results are reported in Tables 5 and 6.

Comparing Tables 1 and 5 we see that both weighted Lasso and weighted adaptive Lasso do a

**Table 3:** *Mean number of correctly non-zero estimated parameters in model* (2.1) *with* $\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$

| | | $\sigma$ | | | |
|---|---|---|---|---|---|
| | | (a) | (b) | (c) | (d) |
| Lasso | $\neq 0$ | 8 | 8 | 7.68 | 8 |
| adaptive Lasso | $\neq 0$ | 8 | 8 | 6.88 | 7.95 |
| weigthed Lasso | $\neq 0$ | 8 | 8 | 8 | 8 |
| weighted adaptive Lasso | $\neq 0$ | 8 | 8 | 7.97 | 7.99 |

**Table 4:** *Mean squared estimation error of the estimators of the non-zero coefficients in model* (2.1) *with* $\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$

| | $\sigma$ | | | |
|---|---|---|---|---|
| | (a) | (b) | (c) | (d) |
| Lasso | 0.0235 | 0.0293 | 0.2108 | 0.0717 |
| adaptive Lasso | 0.0246 | 0.0352 | 0.3078 | 0.0757 |
| weighted Lasso | 0.0115 | 0.0044 | 0.0410 | 0.0240 |
| weighted adaptive Lasso | 0.0125 | 0.0044 | 0.0427 | 0.0246 |

**Table 5:** *Mean number of correctly non-zero estimated parameters in model* (2.1) *with estimated variance function*

| | | $\sigma$ | | | |
|---|---|---|---|---|---|
| | | (a) | (b) | (c) | (d) |
| $\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)$ | | | | | |
| weighted Lasso | $= 0$ | 1.79 | 3.43 | 2.08 | 2.50 |
| | $\neq 0$ | 3 | 3 | 2.99 | 3 |
| weighted adaptive Lasso | $= 0$ | 4.37 | 4.57 | 3.08 | 4.29 |
| | $\neq 0$ | 3 | 3 | 2.99 | 3 |
| $\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$ | | | | | |
| weighted Lasso | $\neq 0$ | 8 | 8 | 7.96 | 7.99 |
| weighted adaptive Lasso | $\neq 0$ | 7.99 | 8 | 7.77 | 7.83 |

**Table 6:** *Mean squared estimation error of the estimators of the non-zero coefficients in model* (2.1) *with estimated variance function*

| | | $\sigma$ | | | |
|---|---|---|---|---|---|
| | | (a) | (b) | (c) | (d) |
| $\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)$ | | | | | |
| weighted Lasso | $\beta_1$ | 0.0218 | 0.0857 | 0.3206 | 0.1476 |
| | $\beta_2$ | 0.0329 | 0.0423 | 0.1719 | 0.0824 |
| | $\beta_3$ | 0.0346 | 0.0542 | 0.2674 | 0.1102 |
| weighted adaptive Lasso | $\beta_1$ | 0.0227 | 0.0530 | 0.2839 | 0.1302 |
| | $\beta_2$ | 0.0382 | 0.0419 | 0.1950 | 0.1067 |
| | $\beta_3$ | 0.0329 | 0.0356 | 0.2607 | 0.1018 |
| $\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)$ | | | | | |
| weighted Lasso | | 0.0237 | 0.0166 | 0.0955 | 0.0775 |
| weighted adaptive Lasso | | 0.0295 | 0.0166 | 0.1296 | 0.1093 |

better job in identifying the zero components of the parameter vector when an estimated variance function is used. Especially the weighted Lasso is much better in this case. Both estimators are less conservative in model selection and do more often exclude important parameters from the model. In terms of estimation error (compare Table 6 with Tables 2 and 4) the weighted Lasso and weighted adaptive Lasso with estimated variance function perform in some cases better and in some cases worse than their non weighted counterparts, thus not identifying a clear winner. However, in most cases the differences are not substantial. Obviously the weighted penalized least squares procedures with a correctly specified variance function yield smaller mean squared errors than the procedures where this function is estimated nonparametrically.

## 5.2 Data example

In this section we investigate the different properties of the estimators $\widehat{\beta}_{lse}$ and $\widehat{\beta}_{wlse}$ in a real data example. We use the diabetes data also considered in Efron et al. (2004) and analyzed with the unweighted Lasso. The data consist of a response variable $Y$ which is a quantitative measure of diabetes progression one year after baseline and of ten covariates (age, sex, body mass index, average blood pressure and six blood serum measurements). It includes $n = 442$ observations.
First we calculated the unweighted Lasso estimate $\widehat{\beta}_{lse}$ using a cross-validated (conservative) tuning parameter $\lambda_n$. This solution excluded three covariates from the model (age and the blood serum measurements LDL and TCH). In a next step we calculated the resulting residuals

$$\varepsilon = Y - X\widehat{\beta}_{lse}$$

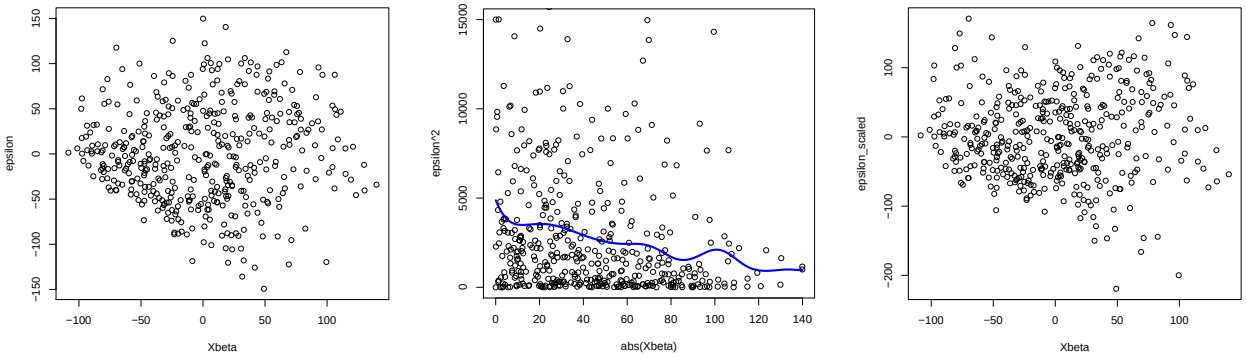


**Figure 1: Left: Lasso residuals, Center: Squared residuals together with local polynomial estimate, Right: rescaled residuals**

**Table 7:** *The Lasso estimators $\widehat{\beta}_{lse}$ and $\widehat{\beta}_{wlse}$ for the tuning parameter $\lambda_n$ selected by cross-validation*

| | Intercept | Age | Sex | BMI | BP | TC | LDL | HDL | TCH | LTG | GLU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\widehat{\beta}_{lse}$ | 152.1 | 0.0 | -186.5 | 520.9 | 291.3 | -90.3 | 0.0 | -220.2 | 0.0 | 506.6 | 49.2 |
| $\widehat{\beta}_{wlse}$ | 183.8 | -110.3 | -271.3 | 673.3 | 408.3 | 84.1 | -547.6 | 0.0 | 449.4 | 213.7 | 138.5 |

which are plotted in the left panel of Figure 1. This picture suggests a heteroscedastic nature of the residuals. In fact the hypothesis of homoscedastic residuals was rejected by the test of Dette and Munk (1998) which had a $p$-value of 0.006. Next we computed a local linear fit of the squared residuals in order to estimate the conditional variance $\sigma(x_i^T\beta)$ of the residuals. The middle panel of Figure 1 presents the squared residuals plotted against its absolute values $|x_i^T\widehat{\beta}_{lse}|$ together with the local linear smoother, say $\hat{\sigma}^2$. In the right panel of Figure 1 we present the rescaled residuals $\tilde{\varepsilon}_i = (Y_i - x_i^T\widehat{\beta}_{lse})/\widehat{\sigma}(|x_i^T\widehat{\beta}_{lse}|)$. These look "more homoscedastic" than the unscaled residuals and the test of Dette and Munk (1998) has a $p$-value of 0.514, thus not rejecting the hypothesis of homoscedasticity. The weighted Lasso estimator $\widehat{\beta}_{wlse}$ was calculated by (2.2) on the basis of the "nonparametric" weights $\hat{\sigma}(x_i^T\widehat{\beta}_{lse})$ and the results are depicted in Table 7. In contrast to $\widehat{\beta}_{lse}$, the weighted Lasso only excludes one variable from the model, namely the blood serum HDL if $\lambda_n$ is chosen by cross-validation.

# 6 Conclusions

We have shown that the attractive properties of bridge estimators and the adaptive Lasso estimator regarding model selection and parameter estimation in linear models with iid errors persist if the errors are heteroscedastic. Nevertheless the asymptotic variance is suboptimal if one does not use a weighted penalized least squares criterion in contrast to the homoscedastic case. Therefore we proposed weighted penalized least squares estimators, where the parameters in the variance function are obtained from preliminary estimators. The resulting estimators have the same model selection properties as the classical estimators but additionally yield optimal asymptotic variances of the estimators corresponding to the non vanishing components. The asymptotic results are supported by a finite sample study. In the examples under consideration the weighted versions of the estimates usually yield a substantially smaller mean squared error. Moreover in most cases the new estimators - in particular the weighted adaptive Lasso - have model selection properties comparable with the estimators, which do not take scale information into account. All results are formulated for the case of a finite dimensional explanatory variable. The case where the dimension of the explanatory $p = p_n$ increases with $n$ requires a completely different asymptotic analysis and is devoted to future research.

**Acknowledgements** This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C1) of the German Research Foundation (DFG). The authors would also like to thank Torsten Hothorn for pointing out some important references on the subject.

# 7 Appendix: Proofs

**Proof of Lemma 3.1:** The proof follows mainly along the same lines as the one of Theorem 2 in Knight and Fu (2000) and we only mention the main differences here. The quantity $\sqrt{n}(\widehat{\beta}_{lse} - \beta_0)$ minimizes the function $V_n$ defined by

$$V_n(u) = \sum_{i=1}^{n} \Big[ \Big( \sigma(x_i, \beta_0)\varepsilon_i - \frac{1}{\sqrt{n}} u^T x_i \Big)^2 - \sigma(x_i, \beta_0)^2 \varepsilon_i^2 \Big] + \lambda_n \Big( \Big\| \beta_0 + \frac{1}{\sqrt{n}} u \Big\|_1 - \|\beta_0\|_1 \Big).$$

By assumptions (i)-(iii), the Lindeberg CLT and the lemma of Slutsky we obtain

$$\sum_{i=1}^{n} \Big[ \Big( \sigma(x_i, \beta_0)\varepsilon_i - \frac{1}{\sqrt{n}} u^T x_i \Big)^2 - \sigma(x_i, \beta_0)^2 \varepsilon_i^2 \Big] \xrightarrow{\mathcal{D}} -2u^T W + u^T C u$$

for every $u \in \mathbb{R}^p$, where $W \sim \mathcal{N}(0, B)$. Now the assertion of the Lemma follows exactly as the one of Theorem 2 in Knight and Fu (2000). □

**Proof of Lemma 3.3:** The proof of the first part follows the same way as the one of Theorem 3 in Knight and Fu (2000) and is therefore omitted. For a proof of the second part we define $u = \sqrt{n}(\beta - \beta_0)$ and obtain (by adding constant terms) that $\sqrt{n}(\widehat{\beta}_{lse} - \beta_0)$ minimizes the function $V_n$ which is defined by

$$V_n(u) = \sum_{i=1}^{n} \Big[ \Big( \sigma(x_i, \beta_0)\varepsilon_i - \frac{1}{\sqrt{n}} u^T x_i \Big)^2 - \sigma(x_i, \beta_0)^2 \varepsilon_i^2 \Big] + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \widehat{w}_j \sqrt{n} \Big( \Big| \beta_{0,j} + \frac{u_j}{\sqrt{n}} \Big| - |\beta_{0,j}| \Big).$$

Here we use the notation $\widehat{w}_j = |\tilde{\beta}|^{-\gamma}$. If $1 \le j \le k$ we obtain

$$\sqrt{n}(|\beta_{0,j} + u_j/\sqrt{n}| - |\beta_{0,j}|) \to u_j \mathrm{sgn}(\beta_{0,j})$$

and the assumptions of the lemma yield $\widehat{w}_j \xrightarrow{P} |\beta_{0,j}|^{-\gamma}$ and $\lambda_n/\sqrt{n} \to 0$.
If $k+1 \le j \le p$ we have

$$\sqrt{n}(|\beta_{0,j} + u_j/\sqrt{n}| - |\beta_{0,j}|) = |u_j| \quad \text{and} \quad \lambda_n/\sqrt{n}\widehat{w}_j = \lambda_n a_n^{\gamma}/\sqrt{n} |a_n \tilde{\beta}|^{-\gamma} \xrightarrow{\mathcal{D}} \lambda_0 |Z_j|^{-\gamma}.$$

Now by the continuous mapping theorem and the proof of Lemma 3.1 it follows that

$$V_n(u) \xrightarrow{\mathcal{D}} V(u)$$

for each $u \in \mathbb{R}^p$. $V_n$ is convex and $V$ strictly convex which can be proved by calculating the second derivatives. Therefore $V$ has a unique minimizing value and Theorem 3.2 of Geyer (1996) yields the assertion. □

**Proof of Theorem 4.1:** As in the proof of Lemma 3.1 we obtain that the quantity $\sqrt{n}(\widehat{\beta}_{wlse} - \beta_0)$ minimizes

$$V_n(u) = \sum_{i=1}^{n} \Big[ \Big( \sigma(x_i, \beta_0)\varepsilon_i - \frac{1}{\sqrt{n}} u^T x_i \Big)^2 \frac{1}{\sigma(x_i, \overline{\beta})^2} - \frac{\sigma(x_i, \beta_0)^2}{\sigma(x_i, \overline{\beta})^2} \varepsilon_i^2 \Big] + \lambda_n \Big( \Big\| \beta_0 + \frac{1}{\sqrt{n}} u \Big\|_1 - \|\beta_0\|_1 \Big).$$

The second term in the equation above converges to $\lambda_0 \sum_{j=1}^{k} u_j \mathrm{sgn}(\beta_{0,j}) + \lambda_0 \sum_{j=k+1}^{p} |u_j|$ by the arguments given in the proof of Theorem 2 in Knight and Fu (2000)). So we only have to show weak convergence of the first term which is given by

$$\tilde{V}_n^{(1)}(u) + \tilde{V}_n^{(2)}(u) = -\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \frac{\sigma(x_i, \beta_0)}{\sigma(x_i, \overline{\beta})^2} \varepsilon_i x_i^T u + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma(x_i, \overline{\beta})^2} u^T x_i x_i^T u, \tag{7.1}$$

where $\tilde{V}_n^{(j)}(u)$ is defined in an obvious manner. Using assumption (iii)' a Taylor expansion yields

$$\frac{1}{\sigma(x_i, \overline{\beta})^2} = \frac{1}{\sigma(x_i, \beta_0)^2} - 2\frac{(\partial\sigma/\partial\beta)(x_i, \beta_0)}{\sigma(x_i, \beta_0)^3}(\overline{\beta} - \beta_0) + (\overline{\beta} - \beta_0)^T M(x_i, \xi)(\overline{\beta} - \beta_0),$$

where $\|\xi - \beta_0\| \leq \|\overline{\beta} - \beta_0\|$ and

$$M(x_i, \xi) = \frac{3\left[(\partial\sigma/\partial\beta)(x_i, \xi)\right]^T (\partial\sigma/\partial\beta)(x_i, \xi) - \sigma(x_i, \xi)(\partial^2\sigma/\partial^2\beta)(x_i, \xi)}{\sigma(x_i, \xi)^4}.$$

Using this Taylor expansion in (7.1) we obtain

$$\begin{aligned}
\tilde{V}_n^{(1)}(u) &= -\frac{2}{\sqrt{n}} \sum_{i=1}^n \frac{\varepsilon_i x_i^T u}{\sigma(x_i, \beta_0)} + \frac{4}{\sqrt{n}} \sum_{i=1}^n \frac{\varepsilon_i x_i^T u}{\sigma(x_i, \beta_0)^2} (\partial\sigma/\partial\beta)(x_i, \beta_0)(\overline{\beta} - \beta_0) \\
&\quad - \frac{2}{\sqrt{n}} (\overline{\beta} - \beta_0)^T \sum_{i=1}^n \sigma(x_i, \beta_0) \varepsilon_i x_i^T u M(x_i, \xi)(\overline{\beta} - \beta_0) \\
&= \tilde{V}_{n,1}^{(1)}(u) + \tilde{V}_{n,2}^{(1)}(u) + \tilde{V}_{n,3}^{(1)}(u).
\end{aligned}$$

The random variable $\tilde{V}_{n,1}^{(1)}(u)$ converges in distribution to $-2u^T W$ with $W \sim \mathcal{N}(0, D)$ by assumptions (i)'-(iii)' and the Lindeberg CLT. $\tilde{V}_{n,2}^{(1)}(u)$ converges to 0 in probability which is shown by an application of the Lindberg CLT, the Cramér-Wold device and the lemma of Slutsky using assumptions (i)'-(iv)'.
By assumption (iii)' and (iv)' and the definition of $\xi$ the maximal absolute value of the eigenvalues of the matrix $M(x_i, \xi)$ is bounded by a constant $c > 0$ independent of $x_i$ and $\xi$ for $n$ sufficiently large with probability $1 - \varepsilon$ (where $\varepsilon > 0$ is arbitrary small). Therefore we obtain

$$|(\overline{\beta} - \beta_0)^T M(x_i, \xi)(\overline{\beta} - \beta_0)| \leq c(\overline{\beta} - \beta_0)^T (\overline{\beta} - \beta_0)$$

for $n$ sufficiently large with probability $1 - \varepsilon$. Now assumption (iii)' and the Cauchy-Schwarz inequality yield

$$\left|\tilde{V}_{n,3}^{(1)}(u)\right| \leq \frac{C\sqrt{n}}{b_n^2} b_n^2 (\overline{\beta} - \beta_0)^T (\overline{\beta} - \beta_0) \frac{1}{n} \left(\sum_{i=1}^n \varepsilon_i^2\right)^{1/2} \left(u^T X^T \Sigma(\beta_0)^{-2} X u\right)^{1/2}$$

for some constant $C > 0$ and for $n$ sufficiently large with probability $1 - \varepsilon$. By assumptions (i)' and (iv)' and the law of large numbers the right hand side of the last inequality converges to 0 in probability. Therefore we obtain

$$\tilde{V}_n^{(1)}(u) \xrightarrow{\mathcal{D}} -2u^T W.$$

By similar arguments one shows

$$\tilde{V}_n^{(2)}(u) \xrightarrow{P} u^T D u$$

which finally yields $V_n(u) \xrightarrow{\mathcal{D}} V(u)$ for each $u \in \mathbb{R}^p$. Because $V_n$ and $V$ are convex functions and $V$ has a unique minimizing value with probability one, Theorem 3.2 of Geyer (1996) yields the assertion of the theorem. □

# References


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *International Symposium on Information Theory, 2nd, Tsahkadsor, Armenian SSR*, pages 267–281.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when $p$ is much larger than $n$. *Annals of Statistics*, 35:2313–2351.

Claeskens, G. and Hjort, N. L. (2003). The focussed information criterion (with discussion). *Journal of the American Statistical Association*, 98:900–916.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline function: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numerische Mathematik*, 31:337–403.

Dette, H. and Munk, A. (1998). Testing heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society, Ser. B*, 60:693–708.

Efron, B., Hastie, T., and Tibshirani, R. (2004). Least angle regression (with discussion). *Annals of Statistics*, 32:407–451.

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

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32:928–961.

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

Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization. *Unpublished Manuscript*.

Huang, J., Horowitz, J. L., and Ma, S. (2008a). Asymptotic properties of bridge estimators in sparse high dimensional regression models. *Annals of Statistics*, 36:587–613.

Huang, J., Ma, S., and Zhang, C. (2008b). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618.

James, G. M., Radchenko, P., and Lv, J. (2009). DASSO: Connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society, Ser. B*, 71:127–142.

Kim, Y., Choi, H., and Oh, H. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103:1665–1673.

Knight, F. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics*, 28:1356–1378.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, 21:21–59.

Leeb, H. and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, 142:201–211.

Pötscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, 100:2065–2082.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90:1257–1270.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264.

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

Wand, M. P. and Jones, M. C. (1995). *Kernel Smooting*. Chapman and Hall, London.

Wang, H., Li, R., and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94:553–568.

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Ser. B*, 67:301–320.