ISSN 1066-5307, Mathematical Methods of Statistics, 2013, Vol. 22, No. 2, pp. 137–154. © Allerton Press, Inc., 2013.

The Adaptive Lasso in High-Dimensional Sparse Heteroscedastic Models

J. Wagener[1*] **and H. Dette**[1**]

[1] *Ruhr-Universität Bochum, Fakultät für Math., Bochum, Germany*

Received January 8, 2013; in final form, April 17, 2013

Abstract—In this paper we study the asymptotic properties of the adaptive Lasso estimate in high-dimensional sparse linear regression models with heteroscedastic errors. It is demonstrated that model selection properties and asymptotic normality of the selected parameters remain valid but with a suboptimal asymptotic variance. A weighted adaptive Lasso estimate is introduced and investigated. In particular, it is shown that the new estimate performs consistent model selection and that linear combinations of the estimates corresponding to the non-vanishing components are asymptotically normally distributed with a smaller variance than those obtained by the "classical" adaptive Lasso. The results are illustrated in a data example and by means of a small simulation study.

Keywords: high-dimensional data, heteroscedasticity, penalized regression, variable selection.

2000 Mathematics Subject Classification: primary 62J05.

DOI: 10.3103/S106653071302004X

1. INTRODUCTION

In recent years the use of penalized likelihood or penalized least squares methods has become very popular in analyzing parametric regression models. An important advantage of some of these methods is that they can be applied in very high-dimensional settings, that is models, where the number of parameters p is larger than the sample size n. Under sparseness assumptions on the true data generating process the use of these estimators can be theoretically justified, in particular, consistency and asymptotic normality can be established. Here "sparseness" means that only a small fraction of the predictors (say $k < n$, where n is the sample size) in the model influences the true data generating process. Some penalized estimators are able to correctly identify the corresponding k non-vanishing coefficients in a linear model and give a reasonable estimate of these, which means that they perform model selection and estimation in a single step. For obtaining asymptotic considerations the high-dimensionality is modeled by a $p = p_n$ dimensional parameter depending on the sample size, which converges to infinity with n.

In recent years substantial progress has been made in analyzing the theoretical and practical properties of these methods. Penalized estimators include bridge estimators (Frank and Friedman (1993)) with the special cases of Lasso (Tibshirani (1996)) and ridge regression (Hoerl and Kennard (1970)), the SCAD (Fan and Li (2001)) or the adaptive Lasso (Zou (2006)). Knight and Fu (2000) established asymptotic properties of bridge estimators (that is least squares estimators with L^q penalty $[0 < q < \infty]$) in the case where the dimension p of the model is fixed. Fan and Li (2001) argued that a reasonable estimator should correctly identify the k important parameters which are influential with probability converging to one and the estimators of these should have the same asymptotic distribution as an estimator which would be used if the k important parameters were known in advance. So the estimator should consistently select a model and the estimators of the parameters of the true model should be asymptotically efficient. They called this the "oracle property". Fan and Li (2001) established

[*] E-mail: jens.wagener@rub.de

[**] E-mail: holger.dette@rub.de

this property for the SCAD in the context of likelihood models and Zou (2006) proved it for the adaptive Lasso in the context of linear models.

The results for the SCAD were generalized to the case, where the dimension of the parameter $p = p_n$ is increasing with the sample size so that $p_n = o(n)$ (see Fan and Peng (2004)), while Kim et al. (2008) showed the oracle property for the SCAD also in the case $p_n > n$. Asymptotic results for bridge estimators with $0 < q \leq 1$ were established in Huang et al. (2008), where oracle properties were shown for $p_n = o(n)$ and $0 < q < 1$. For the case $p_n > n$ a two-stage approach is suggested using marginal bridge estimators, which were shown to consistently select the true model. Although the Lasso does not satisfy the oracle property in the case of fixed p (see Zou (2006)) it can identify the correct model and consistently estimate the important variables in high-dimensional settings (see, e.g., Zhao and Yu (2006) and Wainwright (2009)). Huang et al. (2008) showed that the adaptive Lasso satisfies the oracle property also in high-dimensional linear models under some assumptions (we will sometimes also cite the technical report Huang et al. (2006) foregoing the last mentioned article, because some assumptions are formulated in a more transparent way there). For a broader overview of penalized estimators in high-dimensional models and further references we refer the reader to the recent article of Fan and Lv (2010).

Much of the aforementioned literature concentrates on the case of linear models with independent identically distributed errors. To our best knowledge there has been no attempt to investigate bridge estimators and the adaptive Lasso in high-dimensional linear models with heteroscedastic errors. In the case of fixed p such results were established in Wagener and Dette (2012) who analyzed both bridge estimators with $0 < q \leq 1$ and the adaptive Lasso in the case of heteroscedasticity. Generally speaking the model selection properties of the analyzed estimators still persist under heteroscedasticity. The bridge estimators with $0 < q < 1$ and the adaptive Lasso estimators of the k important parameters are asymptotically normally distributed, but with a suboptimal variance. As a consequence, these authors introduced weighted versions of the bridge and adaptive Lasso estimators, which were shown to have the optimal asymptotic variance.

The present article is devoted to an investigation of problems of this type in the case, where the number of parameters in the model varies with the sample size. It turns out that the analysis differs substantially from the case of fixed p and we concentrate our investigations on the adaptive Lasso estimator, which satisfies the oracle property in homoscedastic linear models and has the advantage of being a solution of a convex minimization problem in contrast to bridge estimators. We will analyze both the "ordinary" adaptive Lasso under heteroscedasticity and a weighted version taking scale information into account. Model selection consistency and asymptotic normality will be established for both estimators and the weighted adaptive Lasso will be shown to satisfy the oracle property.

The remaining part of this paper is organized as follows. In the next section we will introduce some basic notation and define weighted Lasso estimators. In Section 3 we will prove that the weighted adaptive Lasso satisfies the oracle property. The weighted adaptive Lasso requires a preliminary estimator for the determination of the "optimal" weights. Therefore the fourth section is devoted to an investigation of the asymptotic behavior of the "classical" (i.e., unweighted) adaptive Lasso. In particular, we show that under general heteroscedasticity the adaptive unweighted Lasso is still sign consistent and estimates the non-vanishing parameters with an optimal rate, so that it can be used in the weighted procedure as initial estimator. In the last section we present some simulation results and an application of both estimators to a real dataset.

2. PRELIMINARIES

We consider the 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 matrix of covariates, β_0 is a vector of unknown parameters and $\Sigma(\beta_0) = \mathrm{diag}(\sigma(x_1, \beta_0), \ldots, \sigma(x_n, \beta_0))$ is a diagonal matrix with positive entries. We denote by $x_1^T, \ldots, x_n^T$ the rows of the matrix X and assume that $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is a vector of independent identically distributed random variables with $\mathrm{E}[\varepsilon_1] = 0$ and $\mathrm{Var}(\varepsilon_1) = 1$. We further assume that the model is sparse, that is, $\beta_0 = (\beta_0(1)^T, \beta_0(2)^T)^T$, where $\beta_0(1) \in \mathbb{R}^{k_n}$ and $\beta_0(2) = 0 \in \mathbb{R}^{p_n - k_n}$, but we do not know which components of β_0 are 0 (naming the nonzero components $\beta_0(1)^T$ and assuming them to be the first k_n components of β_0 is only for notational convenience). The dimension p_n of the vector β_0 is permitted to grow with the sample size n. Note that Huang et al. (2008) and Huang et al. (2008) considered this model with $\Sigma = I_n$ (the n-dimensional identity matrix), that is under homoscedasticity. Throughout this paper we will use the following notation. We partition the matrix $X = (X(1), X(2))$, where $X(1) \in \mathbb{R}^{n\times k_n}$ and $X(2) \in \mathbb{R}^{n\times(p_n-k_n)}$. The columns of X are denoted by $x(1), \dots, x(p_n)$ and the k_n-dimensional rows of $X(1)$ by $x_1(1)^T, \dots, x_n(1)^T$. We assume X to be nonrandom but with random X all results presented in this paper hold conditionally on the covariates. Let x_{ij} denote the (i, j)th entry of the matrix X and let $\beta_{0,j}$ denote the jth coordinate of the vector β_0. Define the $k_n \times k_n$ matrices

$$C_{11}^{(n)} = \frac{1}{n} X(1)^T X(1) \quad \text{and} \quad D_{11}^{(n)}(\beta) = \frac{1}{n} X(1)^T \Sigma(\beta)^{-2} X(1)$$

and let $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote the maximal and the minimal eigenvalues of the matrix M respectively. In the following discussion we will investigate the estimators

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

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

for the parameter β_0 in model (2.1), where $\tilde{\beta}$ and $\overline{\beta}$ are preliminary estimators for β_0. Here β_j denotes the jth component of the p_n-vector β ($j = 1, \dots, p_n$). Note that $\widehat{\beta}_{\text{lse}}$ is the (unweighted) adaptive Lasso estimator proposed by Zou (2006) and $\widehat{\beta}_{\text{wlse}}$ is a weighted version of it, which addresses the heteroscedastic structure in the data. The parameter λ_n is a tuning parameter which has to be prespecified by the data analyst. It can also be determined by using a data dependent method like cross-validation (Craven and Wahba (1979)).

Following Zhao and Yu (2006) an estimator $\widehat{\beta}$ for β_0 is called *sign consistent* if

$$\lim_{n\to\infty} P(\widehat{\beta} =_s \beta_0) = 1,$$

where $\widehat{\beta} =_s \beta_0$ means that each component of $\widehat{\beta}$ has the same sign as the corresponding component of β_0. Because the sign of 0 is defined as 0, a sign consistent estimator for β_0 estimates all zero components of β_0 as exactly 0 with probability converging to 1 and thus performs consistent model selection. In the following we will use the notation $\operatorname{sign}(x)$ for the sign of $x \in \mathbb{R}$ and $\|\cdot\|_2$ for the l_2-norm in $\mathbb{R}^{k_n}$. For a vector $v \in \mathbb{R}^{p_n}$ and a function $f\colon \mathbb{R} \to \mathbb{R}$ we use the shorthand notation $f(v) = (f(v_1), \dots, f(v_{p_n}))^T$ and inequalities between vectors are understood componentwise. Similarly, a multiplication of column vectors of the same length is also understood componentwise.

3. WEIGHTED ADAPTIVE LASSO

In this section we investigate the asymptotic properties of the weighted adaptive Lasso estimator $\widehat{\beta}_{\text{wlse}}$. Throughout this section we assume that the following conditions hold:

(i) The covariates are scaled so that

$$\frac{1}{n} \sum_{i=1}^n x_{ij}^2 = 1 \quad \text{for } j = 1, \dots, p_n.$$

(ii) There exists a constant $b > 0$ such that the preliminary estimator $\tilde{\beta}$ satisfies

$$\lim_{n\to\infty} P(b \min_{1\le j\le k_n} |\tilde{\beta}_j| < b_n) = 0,$$

where

$$b_n = \min\{|\beta_{0,j}| \mid j \leq k_n\} \tag{3.1}$$

is the minimum of the absolute values of the non-vanishing components of the parameter β_0.

(iii) There exists a sequence $r_n \to \infty$ such that the preliminary estimator $\tilde{\beta}$ satisfies

$$\lim_{n\to\infty} P\Big(\max_{k_n+1\leq j\leq p_n} |\tilde{\beta}_j| \geq \frac{1}{r_n}\Big) = 0.$$

(iv) There exist positive constants K and $\tilde{K}$ and a constant d with $1 \leq d \leq 2$ such that the errors in model (2.1) satisfy

$$P(|\varepsilon_1| > x) \leq K \exp(-\tilde{K} x^d).$$

(v) The sequences λ_n, k_n, p_n, b_n and r_n satisfy

(a) $\dfrac{k_n(\log n)^{I\{d=1\}}}{\sqrt{n} b_n} \to 0,$ (b) $\dfrac{\lambda_n}{\sqrt{n k_n} b_n} \to k \in \mathbb{R},$

(c) $\dfrac{(\log(p_n - k_n))^{1/d}(\log n)^{I\{d=1\}}\sqrt{n}}{\lambda_n r_n} \to 0,$ (d) $\dfrac{k_n\sqrt{n}}{\lambda_n r_n} \to 0.$

(vi) There exist constants λ_1, λ_2 such that the inequalities

$$0 < \lambda_{\min}(C_{11}^{(n)}) \leq \lambda_{\max}(C_{11}^{(n)}) \leq \lambda_1 < \infty$$

and

$$0 < \lambda_2 \leq \lambda_{\min}(D_{11}^{(n)}(\beta_0)) \leq \lambda_{\max}(D_{11}^{(n)}(\beta_0)) < \infty$$

hold.

(vii) There exist constants $\underline{\sigma}$ and $\overline{\sigma}$ such that the variance function satisfies

$$0 < \underline{\sigma} \leq \sigma(x, \beta) \leq \overline{\sigma} < \infty$$

for all x in the range of x_i and for all β in a neighborhood of β_0.

(viii) The mapping $\beta \mapsto \sigma(x, \beta)$ is two times differentiable in a neighborhood of β_0 and the first and second partial derivatives with respect to the first k_n coordinates of β are bounded uniformly with respect to x.

(ix) The preliminary estimator $\overline{\beta}$ is sign consistent for β_0 and its first k_n coordinates $\overline{\beta}(1)$ satisfy

$$\|\overline{\beta}(1) - \beta_0(1)\|_2 = O_p\bigg(\sqrt{\frac{k_n}{n}}\bigg).$$

Conditions (i)–(iv) are the same as in Huang et al. (2006). The properties (ii) and (iii) together were called *zero consistency with rate* r_n and mean that the preliminary estimator $\tilde{\beta}$ can distinguish between zero and non-zero components of the parameter vector well. If $p_n = o(n^{1/2})$, the least squares estimate is consistent, which implies zero consistency with rate r_n, see Huang et al. (2006). These authors also proposed a *marginal regression estimate* and proved its zero consistency in the case $p_n > n$ under a partial orthogonality condition. For the Lasso estimate condition (iii) can easily be derived from estimates for the probability of sign consistency [see, for example, Theorems 3 and 4 in Zhao and Yu (2006)], while condition (ii) follows from sharp thresholds for the Lasso estimate [see, for example, Theorem 1 in Wainwright (2009)]. In both cases some additional assumptions are required, which are not elaborated here for the sake of brevity. Condition (iv) excludes heavy-tailed errors. It can be relaxed if we modify condition (v) appropriately (see Remark 3.1 for details). In order to better understand condition (v) assume b_n to be fixed and $d > 1$. Then condition (v)(a) permits $k_n \sim \sqrt{n}a_n$ for a sequence a_n converging to 0. With such a sequence k_n we can choose $\lambda_n \sim n^{3/4}\sqrt{a_n}$ by condition (v)(b) and this choice requires $r_n \sim n^{1/4}a_n^{1/2-\delta}$ for an arbitrary small $\delta > 0$. Note that this is not a strong assumption, because under some conditions on the covariates we can obtain $r_n \sim n^{1/2-\delta}$ (compare Huang et al. (2006)). With these choices p_n can grow with every polynomial order and even of order $\exp(n^{d/2}a_n^{d(1-\delta+\varepsilon)})$, where $\varepsilon > 0$. The first part of condition (vi) is standard in high-dimensional regression models (see, for example, Fan and Pen (2004), where it is imposed on the Fisher information matrix instead of $C_{11}^{(n)}$). The second part of condition (vi) is needed to address heteroscedasticity and reduces to a standard condition on $C_{11}^{(n)}$ in the case of homoscedasticity. Condition (vi) can be relaxed in that way that the rates of growth of $\lambda_{\max}(C_{11}^{(n)})$ and of decay of $\lambda_{\min}(D_{11}^{(n)}(\beta_0))$ are not too fast provided that condition (v) is modified appropriately. Conditions (vii) and (viii) are standard in heteroscedastic regression. Condition (ix) is a critical one and it is, for example, satisfied for the estimator $\widehat{\beta}_{\text{lse}}$ as shown in Theorems 4.1 and 4.2 in the following section.

Theorem 3.1. *If assumptions* (i)–(ix) *are satisfied then the weighted adaptive Lasso estimator* $\widehat{\beta}_{wlse}$ *is sign consistent for* β_0.

Remark 3.1. Theorem 3.1 also holds without assumption (iv) of light-tailed errors if condition (v)(c) is replaced by the stronger assumption

$$\frac{(p_n - k_n)n}{\lambda_n^2 r_n^2} \to 0 \tag{3.2}$$

on the number of covariates. If we assume b_n to be fixed, $k_n = \sqrt{n}a_n$ for some sequence a_n converging to 0 and $\lambda_n \sim n^{3/4}\sqrt{a_n}$, we require

$$\frac{(p_n - k_n)}{a_n\sqrt{n}r_n^2} \to 0. \tag{3.3}$$

Thus even if r_n is almost "optimal", that is $r_n \sim n^{1/2-\delta}$ for some small $\delta > 0$, the dimension of the model p_n cannot grow polynomially in this case. Nevertheless the case $p_n > n$ growing faster than linearly with n is still covered here.

To obtain the validity of Theorem 3.1 under these different assumptions we recall that condition (iv) was only used in the proof of Theorem 3.1 to obtain estimates for the probabilities $P(A_1)$ and $P(A_3)$ in (6.8). If (3.2) holds we use the inequality

$$P(A_1) \le \sum_{j=1}^{k_n} P\left(\frac{1}{n}|\chi_j(\beta_0)| \ge \frac{b_n}{4}\right) + P\left(\frac{1}{n}\max_{1\le j\le k_n}|\chi_j(\beta_0) - \chi_j(\overline{\beta})| \ge \frac{b_n}{4}\right).$$

The second term on the right-hand side of this equation converges to zero by the same arguments as in the proof of Theorem 3.1. For the first one we use the Chebychev inequality and obtain

$$\sum_{j=1}^{k_n} P\left(\frac{1}{n}|\chi_j(\beta_0)| \ge \frac{b_n}{4}\right) \le \frac{16}{n^2 b_n^2}\sum_{j=1}^{k_n} \mathrm{E}[\chi_j(\beta_0)^2] = O\left(\frac{k_n}{nb_n^2}\right).$$

Thus condition (v)(a) yields $P(A_1) \to 0$. For the probability $P(A_3)$ we use a similar argument and $\mathrm{E}[\eta_j(\beta_0)^2] = O(n)$ to obtain

$$P(A_3) = O\left(\frac{(p_n - k_n)n}{\lambda_n^2 r_n^2}\right) + o(1).$$

Therefore the sign consistency of $\widehat{\beta}_{\text{wlse}}$ under these different assumptions follows.

Theorem 3.2. *Let conditions* (i)–(ix) *or condition* (3.2) *instead of* (iv) *and* (v)(c) *be satisfied and additionally let*

$$\text{(x)} \quad \frac{\lambda_n \sqrt{k_n}}{\sqrt{n b_n}} \to 0, \qquad \frac{k_n^5}{n} \to 0,$$

$$\text{(xi)} \quad \frac{1}{n} \max_{1 \le i \le n} \|x_i(1)\|_2^2 \to 0$$

hold. Then for all $\alpha_n \in \mathbb{R}^{k_n}$ *with* $\|\alpha_n\|_2 = 1$ *the following weak convergence holds*

$$\frac{\sqrt{n}}{s_n} \alpha_n^T \big(\widehat{\beta}_{wlse}(1) - \beta_0(1)\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{3.4}$$

where $s_n^2 = \alpha_n^T \big(D_{11}^{(n)}(\beta_0)\big)^{-1} \alpha_n$.

Assumption (x) is a stronger condition than conditions (v)(a) and (v)(b). In the case, where b_n is constant and $d > 1$, it requires $k_n = n^{1/5} a_n$ for some sequence a_n converging to 0. With this maximal choice of k_n it is satisfied for $\lambda_n \sim n^{2/5} a_n$. Thus conditions (v)(c) and (v)(d) yield stronger assumptions on p_n and r_n than in the case, where only (v)(b) gives a condition on the rate of growth of λ_n. Nevertheless, p_n still can grow exponentially fast. Condition (xi) is needed for the application of the Lindeberg central limit theorem. In view of condition (i) and the second part of (x) it is a very weak assumption because the dimension of the vectors $x_i(1)$ is $k_n = o(n^{-1/5})$.

Remark 3.2. Theorems 3.1 and 3.2 indicate that the weighted adaptive Lasso estimator $\widehat{\beta}_{\text{wlse}}$ is able to perform consistent model selection and consistent estimation of the non-null parameters simultaneously. Moreover the estimators of the non-null parameters are unbiased and asymptotically normal with the same asymptotic variance as the generalized least squares estimator, which would be used if the true model and $\Sigma(\beta_0)$ were known in advance. Thus $\widehat{\beta}_{\text{wlse}}$ satisfies the oracle property in the sense of Fan and Li (2001).

Leeb and Pötscher (2008) showed that an estimator performing consistent model selection must have an unbounded (scaled) risk function although the risk function of the optimal estimator in the "true" model is bounded. As a consequence, they criticized the "oracle concept" of Fan and Li (2001) because it identifies estimators as optimal although their properties do not hold uniformly over the parameter space. We expect that these problems do not disappear in a high-dimensional setting and some results in this direction can be found in Pötscher and Schneider (2011).

4. UNWEIGHTED ADAPTIVE LASSO

In the previous section the asymptotic properties of the estimator $\widehat{\beta}_{\text{wlse}}$ were derived. A critical assumption in the asymptotic theory in Theorems 3.1 and 3.2 is the existence of a preliminary estimator $\overline{\beta}$ for β_0 which is sign consistent and estimates the non-null parameters with the optimal rate. In this section we will establish that the unweighted adaptive Lasso estimator $\widehat{\beta}_{\text{lse}}$ satisfies these requirements. Moreover we will derive the asymptotic distribution of its non-null components and show that it is asymptotically dominated by $\widehat{\beta}_{\text{wlse}}$. For this purpose we use the same notation as in the previous section and assume that assumptions (i)–(v) hold. Moreover, we replace conditions (vi) and (vii) by the following ones.

(vi′) There exist constants λ_1, λ_2 such that the inequality

$$0 < \lambda_1 \le \lambda_{\min}(C_{11}^{(n)}) \le \lambda_{\max}(C_{11}^{(n)}) \le \lambda_2 < \infty$$

holds.

(vii′) There exists a constant $\overline{\sigma}$ such that the inequality

$$0 < \sigma(x,\beta) \le \overline{\sigma} < \infty$$

holds for all x in the range of x_i and for all β in a neighborhood of β_0.

Note that condition (vi) is slightly modified and condition (vii) is relaxed. Our first result establishes the sign consistency of the unweighted adaptive Lasso estimate in the heteroscedastic model (2.1).

Theorem 4.1. *Under conditions* (i)–(v), (vi′) *and* (vii′) *the unweighted adaptive Lasso estimator $\widehat{\beta}_{lse}$ is sign consistent for β_0.*

As a consequence of Theorem 4.1 we obtain that $\widehat{\beta}_{\mathrm{lse}}$ is a candidate for a preliminary estimate in the weighted adaptive Lasso because it satisfies the first part of condition (ix). As explained in Remark 3.1 one can drop condition (iv) at the cost of requiring (3.2) instead of (v)(c). The sign consistency of $\widehat{\beta}_{\mathrm{lse}}$ also holds under these different assumptions, which directly follows from the proof of Theorem 4.1. Moreover, it also satisfies the second part of condition (ix) as shown in the next theorem. Thus $\widehat{\beta}_{\mathrm{lse}}$ can be used instead of $\overline{\beta}$ for the calculation of $\widehat{\beta}_{\mathrm{wlse}}$. The proof of Theorem 4.2 is obtained from Theorem 4.1 and the representation (6.23) analogously to the proof of Theorem 3.2 and is therefore omitted.

Theorem 4.2. *Let conditions* (i)–(v), (vi′) *and* (vii′) *or condition* (3.2) *instead of* (iv) *and* (v)(c) *be satisfied and additionally let* (x) *and* (xi) *hold. Then for all $\alpha_n \in \mathbb{R}^{k_n}$ with $\|\alpha_n\|_2 = 1$ the following weak convergence holds*

$$\frac{\sqrt{n}}{\tilde{s}_n}\alpha_n^T\big(\widehat{\beta}_{lse}(1) - \beta_0(1)\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{4.1}$$

where $\tilde{s}_n^2 = n^{-1}\alpha_n^T\big(C_{11}^{(n)}\big)^{-1}X(1)\Sigma(\beta_0)X(1)^T\big(C_{11}^{(n)}\big)^{-1}\alpha_n$.

Theorem 4.2 also shows that $\widehat{\beta}_{\mathrm{wlse}}$ dominates $\widehat{\beta}_{\mathrm{lse}}$ in terms of asymptotic variance because $\widehat{\beta}_{\mathrm{lse}}$ has to be scaled by $\tilde{s}_n$, which is the same scaling needed for the ordinary least squares estimator if the true model were known. But in a heteroscedastic model the ordinary least squares estimator is dominated by a generalized one, which has the same scaling as $\widehat{\beta}_{\mathrm{wlse}}$. Thus $\widehat{\beta}_{\mathrm{lse}}$ consistently selects a model and has the optimal rate for estimating the non-null parameters but it yields a suboptimal variance.

5. FINITE-SAMPLE PROPERTIES

5.1. Simulation Study

In order to investigate the small sample performance of the adaptive Lasso estimators $\widehat{\beta}_{\mathrm{wlse}}$ and $\widehat{\beta}_{\mathrm{lse}}$ in models with heteroscedastic errors we present the results of a small simulation study. All calculations were performed using the package "penalized" available for R on http://www.R-project.org (R Development Core Team (2008)). The data were generated using a linear model of the form (2.1). We followed Huang et al. (2008) and considered a design matrix X with $n = 100$ rows and $p = 200$ columns in the following way: the n rows of X are independent normally distributed random vectors. The first 15 covariates $(x_{i,1},\dots,x_{i,15})$ are independent of the remaining 185 covariates. The pairwise correlation between $x_{i,k}$ and $x_{i,l}$ is $0.5^{|k-l|}$ both if $k,l \in \{1,\dots,15\}$ or if $k,l \in \{16,\dots,200\}$. The first five coordinates of β_0 were set to 2.5, the coordinates 6–10 were set to 1.5 and the coordinates 11–15 to 0.5; all remaining coordinates of β_0 were 0. The entries of the diagonal matrix Σ were chosen as

(a) $\sigma(x_i,\beta_0) = \frac{1}{2}\sqrt{x_i^T\beta_0}$, (b) $\sigma(x_i,\beta_0) = \frac{1}{4}|x_i^T\beta_0|$,

(c) $\sigma(x_i,\beta_0) = \frac{1}{25}\exp|x_i^T\beta_0|$, (d) $\sigma(x_i,\beta_0) = \frac{1}{50}\exp(x_i^T\beta_0)^2$.

The preliminary estimator $\tilde{\beta}$ was a Lasso estimator in our simulation study. We also investigated the marginal regression estimator proposed in Huang et al. (2008) but all results based on the last mentioned method were inferior to the ones using a Lasso estimator and are therefore not depicted. The estimator $\overline{\beta}$ needed for the calculation of $\widehat{\beta}_{\mathrm{wlse}}$ was the adaptive Lasso estimator $\widehat{\beta}_{\mathrm{lse}}$, which was shown to satisfy the requirements of Theorems 3.1 and 3.2. The tuning parameters λ_n for $\widehat{\beta}_{\mathrm{lse}}$ and $\widehat{\beta}_{\mathrm{wlse}}$ were chosen as 0.95 times the values obtained by cross-validation. In general we observed that the performance of the procedures was not very sensitive with respect to the choice of λ_n. All reported results are based on 100 simulation runs. We expect that results based on a larger number of simulation runs would look very similar. Because such simulation studies are very time consuming we restrict ourselves to the case of 100 simulation runs.

Table 1. Mean number of correctly zero and correctly non-zero estimated parameters in model (2.1) (the ideal values are 185 and 15, respectively)

		σ			
		(a)	(b)	(c)	(d)
$\widehat{\beta}_{\mathrm{lse}}$	$= 0$	159.66	161.09	142.45	145.79
	$\neq 0$	14.1	12.73	12.27	13.65
$\widehat{\beta}_{\mathrm{wlse}}$	$= 0$	166.67	172.66	164.66	163
	$\neq 0$	14.34	13.67	12.23	13.70

The model selection properties of the investigated estimators are reported in Table 1. We observe that both estimators perform quite good model selection. The weighted estimator $\widehat{\beta}_{\mathrm{wlse}}$ always excludes more variables correctly from the model than the "classical" adaptive Lasso estimator $\widehat{\beta}_{\mathrm{lse}}$. In all cases except of example (c) it also includes slightly more variables correctly in the model. Thus the estimator $\widehat{\beta}_{\mathrm{wlse}}$ was superior to $\widehat{\beta}_{\mathrm{lse}}$ in terms of model selection in our simulations.

Table 2. Averaged mean squared error of the estimators of the non-zero coefficients in model (2.1) with $\beta_{0,1} = \cdots = \beta_{0,5} = 2.5$, $\beta_{0,6} = \cdots = \beta_{0,10} = 1.5$, $\beta_{0,11} = \cdots = \beta_{0,15} = 0.5$

		σ			
		(a)	(b)	(c)	(d)
$\widehat{\beta}_{\mathrm{lse}}$	$\beta_1, \ldots, \beta_5$	0.0641	0.2374	0.4827	0.1346
	$\beta_6, \ldots, \beta_{10}$	0.0742	0.2741	0.5583	0.1569
	$\beta_{11}, \ldots, \beta_{15}$	0.1016	0.2085	0.2812	0.1394
$\widehat{\beta}_{\mathrm{wlse}}$	$\beta_1, \ldots, \beta_5$	0.0410	0.2514	0.4614	0.1173
	$\beta_6, \ldots, \beta_{10}$	0.0458	0.1456	0.4858	0.1345
	$\beta_{11}, \ldots, \beta_{15}$	0.0760	0.1101	0.2150	0.1396

In Table 2 we present the mean squared error (MSE) for the estimators of the non-zero components of β_0. The displayed values are MSEs averaged over the first five (big) components of β_0, over the sixth to tenth (moderately sized) components of β_0 and over the eleventh to fifteenth (small) components of β_0, respectively. For most cases we observe that the weighted Lasso estimator $\widehat{\beta}_{\mathrm{wlse}}$ yields more precise estimates of the non-zero components of β_0 than $\widehat{\beta}_{\mathrm{lse}}$ in terms of mean squared error. In several cases the improvement is substantial (see, for example, model (a) and model (b) for the parameters $\beta_6, \ldots, \beta_{10}$ and $\beta_{11}, \ldots, \beta_{15}$). Only in model (b) the estimators for the large components $\beta_1, \ldots, \beta_5$ of the parameter β_0 have a slightly smaller mean squared error if no scaling is used, while the estimators for the small components in model (d) perform nearly identically. Thus the simulations in these examples support our theoretical findings.

5.2. Data Example

In this section we illustrate the different properties of the estimators $\widehat{\beta}_{\text{lse}}$ and $\widehat{\beta}_{\text{wlse}}$ in a real data example. We use the diabetes data considered in Efron et al. (2004). 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). Further, we consider the squares of all covariates and their interactions. This finally results in $p = 65$ covariates (including an intercept), while there are $n = 442$ observations.

First we calculated the unweighted adaptive Lasso estimate $\widehat{\beta}_{\text{lse}}$ using a cross-validated (conservative) tuning parameter λ_n. We used an unweighted Lasso estimator in the place of $\tilde{\beta}$ to calculate the weights of the adaptive Lasso estimator. This solution included eight variables in the model, namely, an intercept, the body mass index, the blood pressure, the blood serums HDL, LTG, and the square of GLU and the interactions between age and sex and body mass index and blood pressure. At the next step we calculated the resulting residuals

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

which are plotted in the left panel of Figure 1.

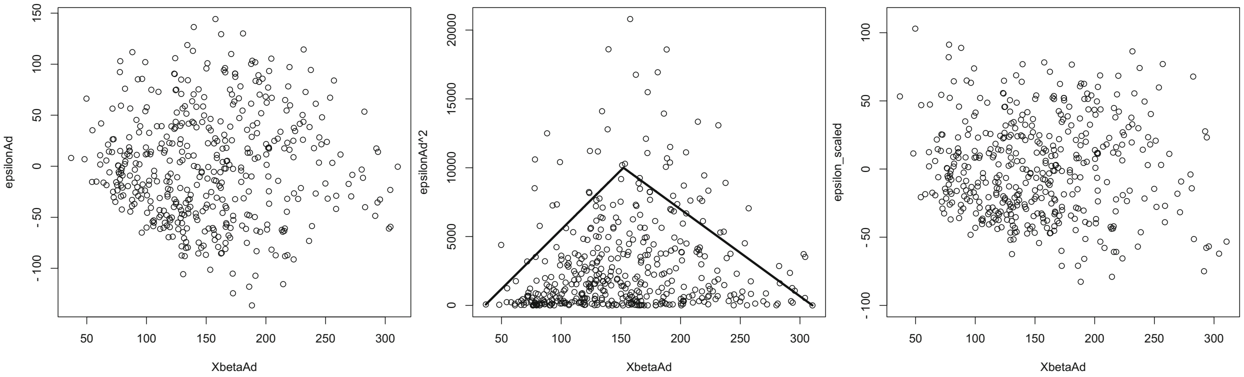

Fig. 1. Left panel: residuals obtained by Lasso; Center: Squared residuals together with a piecewise linear fit, Right panel: rescaled residuals

This picture suggests a heteroscedastic nature of the residuals. In fact the hypothesis of homoscedasticity was rejected at level 5% by the test of Dette and Munk (1998) (p-value 0.033). Next we computed an estimator of the conditional variance $\sigma(x_i^T\beta)$ of the residuals. We used the ad-hoc chosen piecewise linear function

$$\sigma^2(y) = (86.1y - 3110.4)I\{y \le \overline{X\widehat{\beta}_{\text{lse}}}\} + (-63.1y + 19606.9)I\{y > \overline{X\widehat{\beta}_{\text{lse}}}\}$$

(see the middle panel in Figure 1 which shows the squared residuals plotted against the values $x_i^T\widehat{\beta}_{\text{lse}}$ together with the estimator). In the right panel of Figure 1 we present the rescaled residuals $\tilde{\varepsilon}_i = (Y_i - x_i^T\widehat{\beta}_{\text{lse}})/\sigma(|x_i^T\widehat{\beta}_{\text{lse}}|)sd(\varepsilon)$. These look "more homoscedastic" than the unscaled residuals and the test of Dette and Munk (1998) yields a p-value of 0.173, thus not rejecting the hypothesis of homoscedasticity. The weighted adaptive Lasso estimator $\widehat{\beta}_{\text{wlse}}$ was calculated by (2.2) on the basis of the weights $\sigma(x_i^T\widehat{\beta}_{\text{lse}})$. This estimator included the same variables as $\widehat{\beta}_{\text{lse}}$ and additionally the interactions between age and blood pressure, between BMI and GLU and between HDL and LTG.

In Figure 2 we present the data plotted against the fitted values $x_i^T\widehat{\beta}_{\text{lse}}$ and $x_i^T\widehat{\beta}_{\text{wlse}}$ and the residuals in the weighted model. The final residuals look very homoscedastic and both fits are of comparable (moderate) quality.

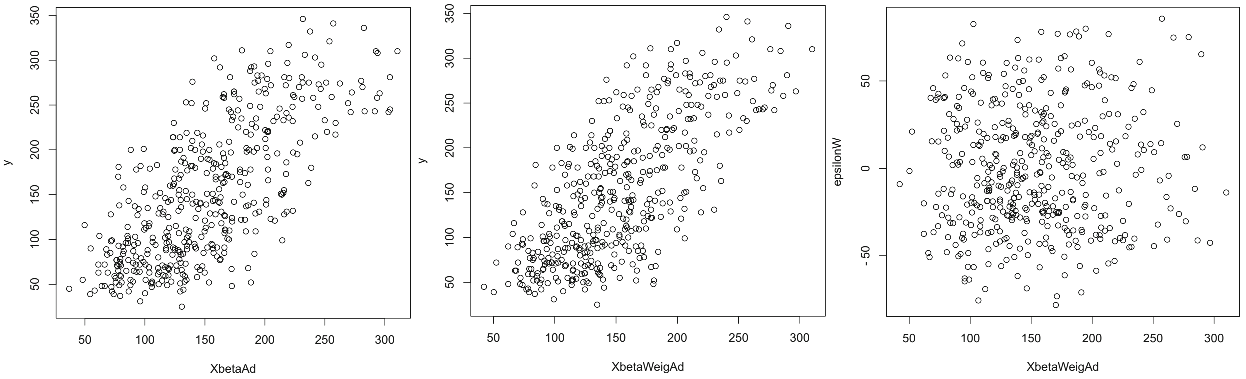

Fig. 2. Left panel: scatterplot of Y and $X\widehat{\beta}_{\text{lse}}$; Center: scatterplot Y and $X\widehat{\beta}_{\text{wlse}}$; Right panel: residuals in the weighted model

6. APPENDIX: PROOF OF THE MAIN RESULTS

6.1. Proof of Theorem 3.1

Throughout this paper let $\|M\|_{\text{op}} = \max\{\|Mx\|_2 \mid \|x\|_2 = 1\}$ and $\|M\|_2 = \sqrt{\operatorname{tr}(M^T M)}$ denote the operator and the Frobenius norm of the matrix M respectively. Further, for a random variable X let

$$\|X\|_{\psi_d} = \inf\{C > 0 \mid \mathrm{E}[\psi_d(|X|/C)] \le 1\} \tag{6.1}$$

denote its Orlicz norm with respect to the function $\psi_d(x) = \exp(x^d) - 1$ $(1 \le d \le 2)$. In the following we make frequent use of the inequalities

$$\|AB\|_{\text{op}} \le \|A\|_{\text{op}} \|B\|_{\text{op}} \tag{6.2}$$

for arbitrary matrices A, B and

$$\|Av\|_2 \le \|A\|_{\text{op}} \|v\|_2 \tag{6.3}$$

for a vector v. Define $\widehat{w}_j = |\tilde{\beta}_j|^{-1}$, then the Karush–Kuhn–Tucker (KKT) conditions directly imply that the vector $\beta = (\beta(1)^T, 0^T_{p_n - k_n})$ minimizes

$$\sum_{i=1}^{n} \left(\frac{Y_i - x_i^T \beta}{\sigma(x_i, \overline{\beta})} \right)^2 + \lambda_n \sum_{j=1}^{p_n} |\beta_j| |\tilde{\beta}_j|^{-1} = (Y - X\beta)^T \Sigma(\overline{\beta})^{-2} (Y - X\beta) + \lambda_n \sum_{j=1}^{p_n} |\beta_j| \widehat{w}_j$$

if and only if the conditions

$$x(j)^T \Sigma(\overline{\beta})^{-2} (Y - X\beta) = \frac{\lambda_n}{2} \widehat{w}_j \operatorname{sign}(\beta_j) \qquad \text{if} \quad \beta_j \neq 0, \tag{6.4}$$

$$|x(j)^T \Sigma(\overline{\beta})^{-2} (Y - X\beta)| < \frac{\lambda_n}{2} \widehat{w}_j \qquad \text{if} \quad \beta_j = 0, \tag{6.5}$$

are satisfied. We define

$$\widehat{\beta}(1) = \beta_0(1) + \frac{1}{n} \big(D_{11}^{(n)}(\overline{\beta})\big)^{-1} X(1)^T \Sigma(\overline{\beta})^{-2} \Sigma(\beta_0) \varepsilon - \frac{1}{n} \big(D_{11}^{(n)}(\overline{\beta})\big)^{-1} \frac{\lambda_n}{2} \widehat{s}(1), \tag{6.6}$$

where $\widehat{s}(1) = (\operatorname{sign}(\beta_{0,1})\widehat{w}_1, \ldots, \operatorname{sign}(\beta_{0,k_n})\widehat{w}_{k_n})^T$ and $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,p_n})$. If $\widehat{\beta}(1) =_s \beta_0(1)$ one easily obtains from the representation (6.6) that the vector $\widehat{\beta} = (\widehat{\beta}(1)^T, 0^T_{p_n - k_n})^T$ satisfies (6.4). Note that $\widehat{\beta}(1) =_s \beta_0(1)$ if $\operatorname{sign}(\beta_{0,j})(\beta_{0,j} - \widehat{\beta}_j) < |\beta_{0,j}|$ for $j = 1, \ldots, k_n$. It also follows by similar arguments as in Huang et al. (2008) that the KKT condition (6.5) is satisfied if

$$|\eta_j(\overline{\beta}) + \zeta_j(\overline{\beta})| < \frac{\lambda_n}{2} \widehat{w}_j \qquad \text{for all} \quad j > k_n, \tag{6.7}$$

where the quantities η_j are defined by

$$\eta_j(\beta) = x(j)^T\Sigma(\beta)^{-2}\Big(I_n - \frac{1}{n}X(1)\big(D_{11}^{(n)}(\beta)\big)^{-1}X(1)^T\Sigma(\beta)^{-2}\Big)\Sigma(\beta_0)\varepsilon$$

and

$$\zeta_j(\beta) = \frac{\lambda_n}{2n}x(j)^T\Sigma(\beta)^{-2}X(1)\big(D_{11}^{(n)}(\beta)\big)^{-1}\widehat{s}(1).$$

Thus we obtain by the representation (6.6) and by (6.7)

$$P(\widehat{\beta} \neq_s \beta_0) \leq P(A_1) + P(A_2) + P(A_3) + P(A_4), \tag{6.8}$$

where the events $A_1, \ldots, A_4$ are given by

$$A_1 = \Big\{\frac{1}{n}|\chi_j(\overline{\beta})| \geq \frac{|\beta_{0,j}|}{2} \text{ for some } j \leq k_n\Big\}, \qquad A_2 = \Big\{\frac{\lambda_n}{n}|\phi_j(\overline{\beta})| \geq |\beta_{0,j}| \text{ for some } j \leq k_n\Big\},$$

$$A_3 = \Big\{|\eta_j(\overline{\beta})| \geq \frac{\lambda_n}{4}\widehat{w}_j \text{ for some } j > k_n\Big\}, \qquad A_4 = \Big\{|\zeta_j(\overline{\beta})| \geq \frac{\lambda_n}{4}\widehat{w}_j \text{ for some } j > k_n\Big\}$$

and we use the notation

$$\chi_j(\beta) = e_j^T\big(D_{11}^{(n)}(\beta)\big)^{-1}X(1)^T\Sigma(\beta)^{-2}\Sigma(\beta_0)\varepsilon, \qquad \phi_j(\beta) = e_j^T\big(D_{11}^{(n)}(\beta)\big)^{-1}\widehat{s}(1)$$

(here e_j denotes the jth unit vector in $\mathbb{R}^{k_n}$). In the following we show

$$P(A_j) \to 0 \qquad \text{for} \quad j = 1, \ldots, 4, \tag{6.9}$$

which implies the assertion of the theorem. By the definition of b_n in (3.1) we obtain

$$\begin{aligned} P(A_1) &\leq P\Big(\frac{1}{n}\max_{1\leq j\leq k_n}|\chi_j(\overline{\beta})| \geq \frac{b_n}{2}\Big) \\ &\leq P\Big(\frac{1}{n}\max_{1\leq j\leq k_n}|\chi_j(\beta_0)| \geq \frac{b_n}{4}\Big) + P\Big(\frac{1}{n}\max_{1\leq j\leq k_n}|\chi_j(\beta_0) - \chi_j(\overline{\beta})| \geq \frac{b_n}{4}\Big). \end{aligned}$$

The definition of the operator norm and (6.2) yield

$$\begin{aligned} \Big\|\frac{1}{\sqrt{n}}\Sigma(\beta_0)^{-1}X(1)\big(D_{11}^{(n)}(\beta_0)\big)^{-1}e_j\Big\|_2 &\leq \Big\|\frac{1}{\sqrt{n}}\Sigma(\beta_0)^{-1}X(1)\Big\|_{\text{op}}\Big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\Big\|_{\text{op}} \\ = \big\|D_{11}^{(n)}(\beta_0)\big\|_{\text{op}}^{1/2}\big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\big\|_{\text{op}} &\leq \big\|C_{11}^{(n)}\big\|_{\text{op}}^{1/2}\big\|\Sigma(\beta_0)^{-2}\big\|_{\text{op}}^{1/2}\big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\big\|_{\text{op}} \\ &\leq \lambda_2^{-1}\sqrt{\lambda_1}\underline{\sigma}^{-1}, \end{aligned}$$

where the last inequality follows from assumptions (vi) and (vii). Thus condition (iv) and Lemma 1 of Huang et al. (2006) (which is a slight generalization of Lemma 1 of Huang et al. (2008)) yield

$$\Big\|\frac{1}{\sqrt{n}}\chi_j(\beta_0)\Big\|_{\psi_d} \leq c(\log n)^{I\{d=1\}}$$

for some constant c independent of n and j. Now an application of the results in Section 2.2 of Vaart and Wellner (1996) gives

$$P\Big(\frac{1}{n}\max_{1\leq j\leq k_n}|\chi_j(\beta_0)| \geq \frac{b_n}{4}\Big) \leq \Bigg(\exp\Bigg(\frac{\sqrt{n}^d b_n^d}{4^d C^d(\log n)^{I\{d=1\}}\log(1+k_n)}\Bigg) - 1\Bigg)^{-1}$$

for some constant $C > 0$ and we obtain by assumption (v)(a)

$$P\Big(\frac{1}{n}\max_{1\leq j\leq k_n}|\chi_j(\beta_0)| \geq \frac{b_n}{4}\Big) \to 0.$$

Using the definition of $\chi_j(\beta)$ it follows from the Cauchy–Schwarz inequality for each $j \leq k_n$

$$|\chi_j(\beta_0) - \chi_j(\overline{\beta})| = \Big|e_j^T\Big[\big(D_{11}^{(n)}(\beta_0)\big)^{-1}X(1)^T\big(\Sigma(\beta_0)^{-2} - \Sigma(\overline{\beta})^{-2}\big)$$

$$
\begin{aligned}
&\quad + \Big(\big(D_{11}^{(n)}(\beta_0)\big)^{-1} - \big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\Big) X(1)^T \Sigma(\overline{\beta})^{-2}\Big]\Sigma(\beta_0)\varepsilon\Big| \\
&\le \Big\|\big(\Sigma(\beta_0)^{-2} - \Sigma(\overline{\beta})^{-2}\big) X(1) \big(D_{11}^{(n)}(\beta_0)\big)^{-1} e_j\Big\|_2 \|\Sigma(\beta_0)\varepsilon\|_2 \\
&\quad + \Big\|\Sigma(\overline{\beta})^{-2} X(1) \Big(\big(D_{11}^{(n)}(\beta_0)\big)^{-1} - \big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\Big) e_j\Big\|_2 \|\Sigma(\beta_0)\varepsilon\|_2 \\
&\le \big\|\Sigma(\beta_0)^{-2} - \Sigma(\overline{\beta})^{-2}\big\|_{\mathrm{op}} \big\|nC_{11}^{(n)}\big\|_{\mathrm{op}}^{1/2} \big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\big\|_{\mathrm{op}} \|\Sigma(\beta_0)\varepsilon\|_2 \\
&\quad + \big\|\Sigma(\overline{\beta})^{-2}\big\|_{\mathrm{op}} \big\|nC_{11}^{(n)}\big\|_{\mathrm{op}}^{1/2} \big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1} - \big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\big\|_{\mathrm{op}} \|\Sigma(\beta_0)\varepsilon\|_2. \qquad (6.10)
\end{aligned}
$$

Next we use assumption (viii) to obtain the Taylor expansion

$$
\begin{aligned}
\frac{1}{\sigma(x_i,\overline{\beta})^2} &= \frac{1}{\sigma(x_i,\beta_0)^2} - 2\frac{(\partial\sigma/\partial\beta)(x_i,\beta_0)}{\sigma(x_i,\beta_0)^3}(\overline{\beta}-\beta_0) \\
&\quad + (\overline{\beta}-\beta_0)^T \frac{3\big[(\partial\sigma/\partial\beta)(x_i,\xi)\big]^T(\partial\sigma/\partial\beta)(x_i,\xi) - \sigma(x_i,\xi)(\partial^2\sigma/\partial^2\beta)(x_i,\xi)}{\sigma(x_i,\xi)^4}(\overline{\beta}-\beta_0) \\
&= \frac{1}{\sigma(x_i,\beta_0)^2} - 2\frac{(\partial\sigma/\partial\beta)(x_i,\beta_0)}{\sigma(x_i,\beta_0)^3}(\overline{\beta}-\beta_0) + (\overline{\beta}-\beta_0)^T M(x_i,\xi)(\overline{\beta}-\beta_0),
\end{aligned}
$$

where the vector ξ satisfies $\|\xi-\beta_0\|_2 \le \|\overline{\beta}-\beta_0\|_2$ and the last line defines $M(x_i,\xi)$ in an obvious way. Consequently, we have

$$
\big\|\Sigma(\beta_0)^{-2} - \Sigma(\overline{\beta})^{-2}\big\|_{\mathrm{op}} \le \max_{1\le i\le n}\left|\frac{2}{\sigma(x_i,\beta_0)^3}\frac{\partial\sigma}{\partial\beta}(x_i,\beta_0)(\overline{\beta}-\beta_0)\right| + \max_{1\le i\le n}\big|(\overline{\beta}-\beta_0)^T M(x_i,\xi)(\overline{\beta}-\beta_0)\big|.
$$

On the event $\{\overline{\beta} =_s \beta_0\}$ conditions (vii), (viii) and the Cauchy–Schwarz inequality yield

$$
\max_{1\le i\le n}\left|\frac{2}{\sigma(x_i,\beta_0)^3}\frac{\partial\sigma}{\partial\beta}(x_i,\beta_0)(\overline{\beta}-\beta_0)\right| \le c\sqrt{k_n}\|(\overline{\beta}(1)-\beta_0(1))\|_2
$$

for some constant c. By condition (ix) we have $P(\overline{\beta} =_s \beta_0) \to 1$ and

$$
\max_{1\le i\le n}\left|\frac{2}{\sigma(x_i,\beta_0)^3}\frac{\partial\sigma}{\partial\beta}(x_i,\beta_0)(\overline{\beta}-\beta_0)\right| = O_P\left(\frac{k_n}{\sqrt{n}}\right).
$$

Let $M_{11}(x_i,\xi)$ denote the upper left $k_n \times k_n$ block of the matrix $M(x_i,\xi)$. Because of assumption (viii) we obtain

$$
\|M_{11}(x_i,\xi)\|_{\mathrm{op}} \le \|M_{11}(x_i,\xi)\|_2 \le Ck_n
$$

for some constant C independent of x_i and ξ. Thus on the event $\{\overline{\beta} =_s \beta_0\}$ it follows

$$
\max_{1\le i\le n}\big|(\overline{\beta}-\beta_0)^T M(x_i,\xi)(\overline{\beta}-\beta_0)\big| \le Ck_n\|(\overline{\beta}(1)-\beta_0(1))\|_2^2 = O_P\left(\frac{k_n^2}{n}\right),
$$

where the last estimate follows again from condition (ix). This gives

$$
\big\|\Sigma(\beta_0)^{-2} - \Sigma(\overline{\beta})^{-2}\big\|_{\mathrm{op}} = O_p\left(\frac{k_n}{\sqrt{n}} + \frac{k_n^2}{n}\right). \qquad (6.11)
$$

By assumption (vi) we have

$$
\big\|nC_{11}^{(n)}\big\|_{\mathrm{op}}^{1/2} = O(\sqrt{n}), \qquad (6.12)
$$

$$
\big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\big\|_{\mathrm{op}} = O(1). \qquad (6.13)
$$

Condition (vii) and the law of large numbers yield

$$
\|\Sigma(\beta_0)\varepsilon\|_2 = O_p(\sqrt{n}).
$$

From these estimates and (6.11) we obtain for the first term in (6.10)

$$\big\|\Sigma(\beta_0)^{-2}-\Sigma(\overline{\beta})^{-2}\big\|_{\text{op}}\big\|nC_{11}^{(n)}\big\|_{\text{op}}^{1/2}\big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\big\|_{\text{op}}\big\|\Sigma(\beta_0)\varepsilon\big\|_2=O_P(k_n\sqrt{n}+k_n^2).$$

Next the estimates (6.11), (6.12) and condition (vi) yield

$$\begin{aligned}\|D_{11}^{(n)}(\overline{\beta})-D_{11}^{(n)}(\beta_0)\|_{\text{op}}&=\Big\|\frac{1}{n}X(1)^T\big(\Sigma(\beta_0)^{-2}-\Sigma(\overline{\beta})^{-2}\big)X(1)\Big\|_{\text{op}}\\&\le\|C_{11}^{(n)}\|_{\text{op}}\|\Sigma(\beta_0)^{-2}-\Sigma(\overline{\beta})^{-2}\|_{\text{op}}=O_p\Big(\frac{k_n}{\sqrt{n}}+\frac{k_n^2}{n}\Big).\end{aligned}\tag{6.14}$$

For each invertible matrix A the mapping $A\mapsto A^{-1}$ is Fréchet differentiable and its derivative at A evaluated at the matrix B is given by $-A^{-1}BA^{-1}$ (compare, e.g., Example X.4.2 of Bhatia (1997)). With the notation $A=D_{11}^{(n)}(\beta_0)$ and $B=D_{11}^{(n)}(\overline{\beta})-D_{11}^{(n)}(\beta_0)$ this directly implies

$$\begin{aligned}\big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}-\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\big\|_{\text{op}}&=\big\|A^{-1}-(A+B)^{-1}\big\|_{\text{op}}\\&\le\big\|A^{-1}-(A+B)^{-1}+A^{-1}BA^{-1}\big\|_{\text{op}}+\big\|A^{-1}BA^{-1}\big\|_{\text{op}}\\&\le O(\|B\|_{\text{op}})+\|A^{-1}\|_{\text{op}}^2\|B\|_{\text{op}}=O_p\Big(\frac{k_n}{\sqrt{n}}+\frac{k_n^2}{n}\Big),\end{aligned}\tag{6.15}$$

where the last line follows from (6.14) and assumptions (vi) and (vii). By condition (vii) we obtain the estimate $\|\Sigma(\overline{\beta})^{-2}\|_{\text{op}}=O_p(1)$, which gives for the second term in (6.10) the estimate

$$\big\|\Sigma(\overline{\beta})^{-2}\big\|_{\text{op}}\big\|nC_{11}^{(n)}\big\|_{\text{op}}^{1/2}\big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}-\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\big\|_{\text{op}}\big\|\Sigma(\beta_0)\varepsilon\big\|_2=O_P(k_n\sqrt{n}+k_n^2).$$

Combining these arguments finally yields

$$\frac{1}{n}\max_{1\le j\le k_n}|\chi_j(\beta_0)-\chi_j(\overline{\beta})|=O_P\Big(\frac{k_n}{\sqrt{n}}+\frac{k_n^2}{n}\Big)$$

and by condition (v)(a) it follows

$$P\Big(\frac{1}{n}\max_{1\le j\le k_n}|\chi_j(\beta_0)-\chi_j(\overline{\beta})|\ge\frac{b_n}{4}\Big)\to 0,$$

which implies (6.9) for the case $j=1$.

Next we consider the probability $P(A_2)$ and observe

$$P(A_2)\le P\Big(\frac{\lambda_n}{n}\max_{1\le j\le k_n}|\phi_j(\overline{\beta})|\ge b_n\Big).$$

For each $j\le k_n$ we have

$$|\phi_j(\overline{\beta})|\le\big\|\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\big\|_{\text{op}}\|\widehat{s}(1)\|_2.$$

Let

$$\lambda_1(\beta_0),\ldots,\lambda_{k_n}(\beta_0)\qquad\text{and}\qquad\lambda_1(\overline{\beta}),\ldots,\lambda_{k_n}(\overline{\beta})$$

denote the ordered eigenvalues of the matrices $\big(D_{11}^{(n)}(\beta_0)\big)^{-1}$ and $\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}$ respectively. Weyl's perturbation theorem (see, e.g., Corollary III.2.6 of Bhatia (1997)) shows

$$\max_{1\le j\le k_n}|\lambda_j(\beta_0)-\lambda_j(\overline{\beta})|\le\big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}-\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\big\|_{\text{op}}\xrightarrow{P}0$$

and thus condition (vi) implies that for each $\varepsilon>0$ and $\delta>0$ there exists an $n_0\in\mathbb{N}$ such that for all $n\ge n_0$

$$\big\|\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\big\|_{\text{op}}\le\lambda_2^{-1}+\delta\tag{6.16}$$

with probability at least $(1-\varepsilon)$. Condition (ii) implies that with probability at least $(1-\varepsilon)$ and for n sufficiently large the inequality

$$\|\widehat{s}(1)\|_2 \le \sqrt{k_n}\sqrt{\max_{1\le j\le k_n} |\tilde{\beta}_j|^{-1}} \le \frac{\sqrt{k_n b}}{\sqrt{b_n}} \tag{6.17}$$

is satisfied. The last two estimates and assumptions (v)(a) and (b) directly yield $P(A_2)\to 0$.

Now we consider the term $P(A_3)$. We obtain analogously to $P(A_1)$ the inequality

$$\begin{aligned} P(A_3) \le P\biggl(\max_{k_n+1\le j\le p_n} |\eta_j(\beta_0)| \ge \frac{\lambda_n r_n}{8}\biggr) + P\biggl(\max_{k_n+1\le j\le p_n} |\eta_j(\beta_0)-\eta_j(\overline{\beta})| \ge \frac{\lambda_n r_n}{8}\biggr) \\ + P\biggl(\max_{k_n+1\le j\le p_n} |\tilde{\beta}| > \frac{1}{r_n}\biggr). \end{aligned} \tag{6.18}$$

Define

$$H(\beta_0) = \Sigma(\beta_0)^{-1}x(j) - \Sigma(\beta_0)^{-1}\frac{1}{n}X(1)\bigl(D_{11}^{(n)}(\beta_0)\bigr)^{-1}X(1)^T\Sigma(\beta_0)^{-2}x(j).$$

By condition (i) we have $\|x(j)\|_2 = \sqrt{n}$ and thus

$$\begin{aligned} \|H(\beta_0)\|_2 &\le \|\Sigma(\beta_0)^{-1}\|_{\text{op}}\Bigl(1+\Bigl\|\frac{1}{n}X(1)\bigl(D_{11}^{(n)}(\beta_0)\bigr)^{-1}X(1)^T\Bigr\|_{\text{op}}\|\Sigma(\beta_0)^{-2}\|_{\text{op}}\Bigr)\sqrt{n} \\ &\le \underline{\sigma}^{-1}\bigl(1+\lambda_2^{-1}\lambda_1\underline{\sigma}^{-2}\bigr)\sqrt{n}, \end{aligned}$$

where the last inequality follows from conditions (vi) and (vii). So Lemma 1 of Huang et al. (2006) is applicable to $n^{-1/2}\eta_j$ and we obtain

$$\Bigl\|\frac{1}{\sqrt{n}}\eta_j(\beta_0)\Bigr\|_{\psi_d} = \Bigl\|\frac{1}{\sqrt{n}}H(\beta_0)^T\varepsilon\Bigr\|_{\psi_d} \le c(\log n)^{I\{d=1\}}$$

with some constant c independent of n and j. Again the arguments given in Section 2.2 of Vaart and Wellner (1996) yield

$$P\biggl(\max_{k_n+1\le j\le p_n} |\eta_j(\beta_0)| \ge \frac{\lambda_n r_n}{8}\biggr) \le \biggl(\exp\biggl(\frac{(\lambda_n r_n)^d}{8^d C^d \sqrt{n}^d(\log n)^{I\{d=1\}}\log(1+p_n-k_n)}\biggr)-1\biggr)^{-1}$$

for some constant $C>0$. By assumption (v)(c) the right-hand side of the last inequality converges to zero. Next we consider the second term in (6.18). Using condition (i) and the Cauchy–Schwarz inequality it follows

$$\begin{aligned} |\eta_j(\beta_0)-\eta_j(\overline{\beta})| \le \sqrt{n}\bigl\|\Sigma(\overline{\beta})^{-2}-\Sigma(\beta_0)^{-2}\bigr\|_{\text{op}}\bigl\|\Sigma(\beta_0)\varepsilon\bigr\|_2 \\ + \sqrt{n}\Bigl\|\frac{1}{n}\Bigl(\Sigma(\overline{\beta})^{-2}X(1)\bigl(D_{11}^{(n)}(\overline{\beta})\bigr)^{-1}X(1)^T\Sigma(\overline{\beta})^{-2} \\ - \Sigma(\beta_0)^{-2}X(1)\bigl(D_{11}^{(n)}(\beta_0)\bigr)^{-1}X(1)^T\Sigma(\beta_0)^{-2}\Bigr)\Bigr\|_{\text{op}}\|\Sigma(\beta_0)\varepsilon\|_2. \end{aligned}$$

By assumption (vii), the law of large numbers and (6.11) we obtain for the first term

$$\sqrt{n}\bigl\|\Sigma(\overline{\beta})^{-2}-\Sigma(\beta_0)^{-2}\bigr\|_{\text{op}}\bigl\|\Sigma(\beta_0)\varepsilon\bigr\|_2 = O_p\bigl(k_n\sqrt{n}+k_n^2\bigr),$$

while the second term can be estimated as follows:

$$\begin{aligned} &\Bigl\|\frac{1}{n}\Bigl(\Sigma(\overline{\beta})^{-2}X(1)\bigl(D_{11}^{(n)}(\overline{\beta})\bigr)^{-1}X(1)^T\Sigma(\overline{\beta})^{-2} - \Sigma(\beta_0)^{-2}X(1)\bigl(D_{11}^{(n)}(\beta_0)\bigr)^{-1}X(1)^T\Sigma(\beta_0)^{-2}\Bigr)\Bigr\|_{\text{op}} \\ &\quad\le \Bigl\|\frac{1}{n}\bigl(\Sigma(\overline{\beta})^{-2}-\Sigma(\beta_0)^{-2}\bigr)X(1)\bigl(D_{11}^{(n)}(\overline{\beta})\bigr)^{-1}X(1)^T\Sigma(\overline{\beta})^{-2}\Bigr\|_{\text{op}} \\ &\qquad + \Bigl\|\frac{1}{n}\Sigma(\beta_0)^{-2}X(1)\Bigl(\bigl(D_{11}^{(n)}(\overline{\beta})\bigr)^{-1}-\bigl(D_{11}^{(n)}(\beta_0)\bigr)^{-1}\Bigr)X(1)^T\Sigma(\overline{\beta})^{-2}\Bigr\|_{\text{op}} \end{aligned}$$

$$
\begin{aligned}
&+ \Big\| \frac{1}{n}\Sigma(\beta_0)^{-2}X(1)\big(D_{11}^{(n)}(\beta_0)\big)^{-1}X(1)^T\big(\Sigma(\overline{\beta})^{-2}-\Sigma(\beta_0)^{-2}\big)\Big\|_{\mathrm{op}} \\
\le{}& \big\|\Sigma(\overline{\beta})^{-2}-\Sigma(\beta_0)^{-2}\big\|_{\mathrm{op}}\big\|C_{11}^{(n)}\big\|_{\mathrm{op}}\big\|\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\big\|_{\mathrm{op}}\big\|\Sigma(\overline{\beta})^{-2}\big\|_{\mathrm{op}} \\
&+ \big\|\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}-\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\big\|_{\mathrm{op}}\big\|C_{11}^{(n)}\big\|_{\mathrm{op}}\big\|\Sigma(\beta_0)^{-2}\big\|_{\mathrm{op}}\big\|\Sigma(\overline{\beta})^{-2}\big\|_{\mathrm{op}} \\
&+ \big\|\Sigma(\overline{\beta})^{-2}-\Sigma(\beta_0)^{-2}\big\|_{\mathrm{op}}\big\|C_{11}^{(n)}\big\|_{\mathrm{op}}\big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\big\|_{\mathrm{op}}\big\|\Sigma(\beta_0)^{-2}\big\|_{\mathrm{op}} = O_p\Big(\frac{k_n}{\sqrt{n}}+\frac{k_n^2}{n}\Big),
\end{aligned}
$$

where the last line is a consequence of (6.11), (6.15), (6.16) and conditions (vi), (vii) and (ix). The last three estimates yield

$$
\max_{k_n+1\le j\le p_n} |\eta_j(\beta_0)-\eta_j(\overline{\beta})| = O_P(k_n\sqrt{n}+k_n^2)
$$

and by assumption (v)(d) we have

$$
P\Big(\max_{k_n+1\le j\le p_n} |\eta_j(\beta_0)-\eta_j(\overline{\beta})| \ge \frac{\lambda_n r_n}{8}\Big) \to 0.
$$

Thus condition (iii) and (6.18) yield $P(A_3)\to 0$.

Finally, we consider $P(A_4)$ and observe

$$
P(A_4) \le P\Big(\max_{k_n+1\le j\le p_n} |\zeta_j(\overline{\beta})| \ge \frac{\lambda_n r_n}{4}\Big) + P\Big(\max_{k_n+1\le j\le p_n} |\tilde{\beta}_j| > \frac{1}{r_n}\Big).
$$

The definition of ζ_j yields

$$
\begin{aligned}
|\zeta_j(\overline{\beta})| &\le \frac{\lambda_n}{2n}\big\|\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}X(1)^T\Sigma(\overline{\beta})^{-2}x(j)\big\|_2\|\widehat{s}(1)\|_2 \\
&\le \frac{\lambda_n}{2}\big\|\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\big\|_{\mathrm{op}}\big\|\Sigma(\overline{\beta})^{-2}\big\|_{\mathrm{op}}\Big\|\frac{1}{\sqrt{n}}X(1)^T\Big\|_{\mathrm{op}}\big\|\widehat{s}(1)\big\|_2 = O_P\Big(\frac{\sqrt{k_n}\lambda_n}{\sqrt{b_n}}\Big),
\end{aligned}
$$

where the last line follows from (6.17) and arguments given above. Thus conditions (v)(b), (v)(d) and (iii) show that $P(A_4)\to 0$ and the sign consistency of $\widehat{\beta}_{\mathrm{wlse}}$ follows from (6.8) and (6.9). □

6.2. Proof of Theorem 3.2

By Theorem 3.1 the probability of the event $\{\widehat{\beta}_{\mathrm{wlse}} =_s \beta_0\}$ converges to one. On that event we have by (6.6) the identity

$$
\widehat{\beta}_{\mathrm{wlse}}(1) = \beta_0(1) + \frac{1}{n}\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}X(1)^T\Sigma(\overline{\beta})^{-2}\Sigma(\beta_0)\varepsilon - \frac{1}{n}\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\frac{\lambda_n}{2}\widehat{s}(1),
$$

where we use the same notation as in the proof of Theorem 3.1. Thus we obtain the representation

$$
\frac{\sqrt{n}}{s_n}\alpha_n^T\big(\widehat{\beta}_{\mathrm{wlse}}(1)-\beta_0(1)\big) = \frac{1}{\sqrt{n}s_n}\alpha_n^T\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}X(1)^T\Sigma(\overline{\beta})^{-2}\Sigma(\beta_0)\varepsilon - \frac{1}{\sqrt{n}s_n}\alpha_n^T\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\frac{\lambda_n}{2}\widehat{s}(1).
$$

Let $\varepsilon > 0$. First the proof of Theorem 3.1 implies that for n sufficiently large and small $\delta > 0$ the inequality

$$
\Big|\frac{1}{\sqrt{n}s_n}\alpha_n^T\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\frac{\lambda_n}{2}\widehat{s}(1)\Big| \le \frac{\lambda_n\sqrt{k_n b}}{2s_n\sqrt{nb_n}}(\lambda_2^{-1}+\delta)
$$

holds with probability at least $(1-\varepsilon)$. Further we have $s_n^2 \ge \lambda_1^{-1}\underline{\sigma}^2$ by conditions (vi) and (vii). Thus

$$
\Big|\frac{1}{\sqrt{n}s_n}\alpha_n^T\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}\frac{\lambda_n}{2}\widehat{s}(1)\Big| = O_p\Big(\frac{\lambda_n\sqrt{k_n}}{\sqrt{nb_n}}\Big) = o_p(1),
$$

where the last equality follows from the first part of condition (x). Next we use the decomposition

$$
\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}X(1)^T\Sigma(\overline{\beta})^{-2}\Sigma(\beta_0) = \big(D_{11}^{(n)}(\beta_0)\big)^{-1}X(1)^T\Sigma(\beta_0)^{-1}
$$

$$+\Big(\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}-\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\Big)X(1)^T\Sigma(\beta_0)^{-1}$$

$$+\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}X(1)^T\big(\Sigma(\overline{\beta})^{-2}-\Sigma(\beta_0)^{-2}\big)\Sigma(\beta_0)=A_n+B_n+C_n, \qquad (6.19)$$

where the last line defines A_n, B_n and C_n in an obvious way. We directly obtain

$$\frac{1}{\sqrt{n}s_n}\alpha_n^T A_n\varepsilon=\sum_{i=1}^n c_i\varepsilon_i,$$

where the numbers c_i $(i=1,\dots,n)$ are given by

$$c_i=\frac{1}{\sqrt{n}s_n\sigma(x_i,\beta_0)}\alpha_n^T\big(D_{11}^{(n)}(\beta_0)\big)^{-1}x_i(1).$$

Direct calculations yield $\mathrm{E}\Big[\sum_{i=1}^n c_i\varepsilon_i\Big]=0$ and

$$\mathrm{E}\Big[\Big(\sum_{i=1}^n c_i\varepsilon_i\Big)^2\Big]=\sum_{i=1}^n c_i^2=\frac{1}{s_n^2}\alpha_n^T\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\frac{1}{n}\sum_{i=1}^n\frac{x_i(1)x_i(1)^T}{\sigma(x_i,\beta_0)^2}\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\alpha_n=1$$

by the definition of s_n^2. Conditions (vi), (vii) and (xi) yield

$$\max_{1\le i\le n}|c_i|\le\frac{1}{\sqrt{n}s_n}\big\|\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\big\|_{\mathrm{op}}\max_{1\le i\le n}\Big\|\frac{x_i(1)}{\sigma(x_i,\beta_0)}\Big\|_2\le\sqrt{\lambda_1}\lambda_2^{-1}\underline{\sigma}^{-2}\frac{1}{\sqrt{n}}\max_{1\le i\le n}\|x_i(1)\|_2\to 0.$$

Thus is follows from the Lindeberg CLT

$$\frac{1}{\sqrt{n}s_n}\alpha_n^T A_n\varepsilon\xrightarrow{\mathcal{D}}\mathcal{N}(0,1). \qquad (6.20)$$

Now we consider the term B_n in (6.19). Its definition yields

$$\Big|\frac{1}{\sqrt{n}s_n}\alpha_n^T B_n\varepsilon\Big|\le\frac{\sqrt{\lambda_1}\underline{\sigma}^{-1}}{\sqrt{n}}\Big\|\big(D_{11}^{(n)}(\overline{\beta})\big)^{-1}-\big(D_{11}^{(n)}(\beta_0)\big)^{-1}\Big\|_{\mathrm{op}}\|X(1)^T\Sigma(\beta_0)^{-1}\varepsilon\|_2.$$

By Markov's inequality we obtain for every $t>0$

$$P\Big(\frac{1}{nk_n}\|X(1)^T\Sigma(\beta_0)^{-1}\varepsilon\|_2^2>t\Big)\le\frac{1}{tk_n}\sum_{j=1}^{k_n}\frac{1}{n}\mathrm{E}\Big[\Big(\sum_{i=1}^n x_{ij}\frac{\varepsilon_i}{\sigma(x_i,\beta_0)}\Big)^2\Big]\le\frac{1}{t\underline{\sigma}^2},$$

where the last line follows from conditions (i) and (vii). Thus equation (6.15) and condition (x) yield

$$\Big|\frac{1}{\sqrt{n}s_n}\alpha_n^T B_n\varepsilon\Big|=O_P\Big(\frac{k_n^{3/2}}{\sqrt{n}}\Big)=o_P(1). \qquad (6.21)$$

Finally, we consider the term C_n in (6.19). For each $\varepsilon>0$, arbitrary small $\delta>0$ and n sufficiently large the inequality

$$\Big|\frac{1}{\sqrt{n}s_n}\alpha_n^T C_n\varepsilon\Big|\le\frac{\sqrt{\lambda_1}}{\sqrt{n}\underline{\sigma}}(\lambda_2^{-1}+\delta)\big\|X(1)^T\big(\Sigma(\overline{\beta})^{-2}-\Sigma(\beta_0)^{-2}\big)\Sigma(\beta_0)\varepsilon\big\|_2$$

holds with probability at least $(1-\varepsilon)$. Using a Taylor expansion of the function $\sigma(x_i,\beta)^{-2}$ in a neighborhood of the point β_0 we obtain

$$\sum_{i=1}^n x_{ij}\Big(\frac{1}{\sigma(x_i,\overline{\beta})^2}-\frac{1}{\sigma(x_i,\beta_0)^2}\Big)\sigma(x_i,\beta_0)\varepsilon_i$$

$$=-2\sum_{i=1}^n\frac{\partial\sigma}{\partial\beta}(x_i,\beta_0)(\overline{\beta}-\beta_0)\frac{x_{ij}\varepsilon_i}{\sigma(x_i,\beta_0)^2}+(\overline{\beta}-\beta_0)^T\sum_{i=1}^n x_{ij}\sigma(x_i,\beta_0)\varepsilon_i M(x_i,\xi)(\overline{\beta}-\beta_0),$$

where the matrix $M(x_i,\xi)$ is defined as in the proof of Theorem 3.1. On the event $\{\overline{\beta} =_s \beta_0\}$ we obtain the estimate

$$\Big|\sum_{i=1}^{n} \frac{\partial\sigma}{\partial\beta}(x_i,\beta_0)(\overline{\beta}-\beta_0)\frac{x_{ij}\varepsilon_i}{\sigma(x_i,\beta_0)^2}\Big| \le \sum_{k=1}^{k_n}|\overline{\beta}_k-\beta_{0,k}|\Big|\sum_{i=1}^{n}\frac{\partial\sigma}{\partial\beta_k}(x_i,\beta_0)\frac{x_{ij}\varepsilon_i}{\sigma(x_i,\beta_0)^2}\Big|$$
$$\le \|\overline{\beta}(1)-\beta_0(1)\|_2\bigg(\sum_{k=1}^{k_n}\Big(\sum_{i=1}^{n}\frac{\partial\sigma}{\partial\beta_k}(x_i,\beta_0)\frac{x_{ij}\varepsilon_i}{\sigma(x_i,\beta_0)^2}\Big)^2\bigg)^{1/2} = O_p\Big(\frac{\sqrt{k_n}}{\sqrt{n}}\Big)O_P(\sqrt{nk_n}) = O_P(k_n),$$

where the last line follows by a similar argument as used for the estimate of B_n, using conditions (i), (viii), and (ix). Further the inequality

$$\Big|(\overline{\beta}-\beta_0)^T\sum_{i=1}^{n} x_{ij}\sigma(x_i,\beta_0)\varepsilon_i M(x_i,\xi)(\overline{\beta}-\beta_0)\Big|$$
$$\le \overline{\sigma}\max_{1\le i\le n}\Big|(\overline{\beta}(1)-\beta_0(1))^T M_{11}(x_i,\xi)(\overline{\beta}(1)-\beta_0(1))\Big|\sum_{i=1}^{n}|x_{ij}\varepsilon_i| = O_P\Big(\frac{k_n^2}{n}\Big)O_P(n) = O_P(k_n^2)$$

holds on the event $\{\overline{\beta} =_s \beta_0\}$, where we used conditions (i), (ix), the Chebychev inequality and the estimate $\|M_{11}(x_i,\xi)\|_{\mathrm{op}} = O_P(k_n)$ in the last line. Thus we obtain

$$\Big\|X(1)^T\big(\Sigma(\overline{\beta})^{-2}-\Sigma(\beta_0)^{-2}\big)\Sigma(\beta_0)\varepsilon\Big\|_2$$
$$= \bigg(\sum_{j=1}^{k_n}\Big(\sum_{i=1}^{n}x_{ij}\Big(\frac{1}{\sigma(x_i,\overline{\beta})^2}-\frac{1}{\sigma(x_i,\beta_0)^2}\Big)\sigma(x_i,\beta_0)\varepsilon_i\Big)^2\bigg)^{1/2} = O_P(k_n^{5/2}),$$

which yields

$$\frac{1}{\sqrt{n}s_n}\alpha_n^T C_n\varepsilon = o_P(1) \tag{6.22}$$

using condition (x). Finally, (6.19)–(6.22) and the Slutsky lemma yield the assertion of the theorem. □

6.3. Proof of Theorem 4.2

As in the proof of Theorem 3.1 we obtain

$$\widehat{\beta}_{\mathrm{lse}} =_s \beta_0 \qquad \text{if} \qquad \begin{cases} \operatorname{sign}(\beta_{0,j})(\beta_{0,j}-\widehat{\beta}_{lse,j}) < |\beta_{0,j}| & \text{for all } j\le k_n,\\ |\eta_j+\zeta_j| < \frac{\lambda_n}{2}\widehat{w}_j, & \text{for all } j > k_n,\end{cases}$$

where η_j and ζ_j are given by

$$\eta_j = x(j)^T\Big(I_n - \frac{1}{n}X(1)\big(C_{11}^{(n)}\big)^{-1}X(1)^T\Big)\Sigma(\beta_0)\varepsilon, \qquad \zeta_j = \frac{\lambda_n}{2n}x(j)^T X(1)\big(C_{11}^{(n)}\big)^{-1}\widehat{s}(1),$$

respectively, and

$$\widehat{\beta}_{\mathrm{lse}}(1) = (\widehat{\beta}_{lse,1},\dots,\widehat{\beta}_{lse,k_n})^T = \beta_0(1) + \frac{1}{n}\big(C_{11}^{(n)}\big)^{-1}X(1)^T\Sigma(\beta_0)\varepsilon - \frac{1}{n}\Big(C_{11}^{(n)}\Big)^{-1}\frac{\lambda_n}{2}\widehat{s}(1). \tag{6.23}$$

This directly yields $P(\widehat{\beta}_{\mathrm{lse}} \ne_s \beta_0) \le P(\tilde{A}_1)+P(\tilde{A}_2)+P(\tilde{A}_3)+P(\tilde{A}_4)$, where the events $\tilde{A}_1, \tilde{A}_2, \tilde{A}_3$ and $\tilde{A}_4$ are defined by

$$\tilde{A}_1 = \Big\{\frac{1}{n}\big|e_j^T\big(C_{11}^{(n)}\big)^{-1}X(1)^T\Sigma(\beta_0)\varepsilon\big| \ge \frac{|\beta_{0,j}|}{2} \text{ for some } j\le k_n\Big\},$$
$$\tilde{A}_2 = \Big\{\frac{\lambda_n}{n}\big|e_j^T\big(C_{11}^{(n)}\big)^{-1}\widehat{s}(1)\big| \ge |\beta_{0,j}| \text{ for some } j\le k_n\Big\},$$

$$\tilde{A}_3 = \Big\{ |\eta_j| \geq \frac{\lambda_n}{4} \widehat{w}_j \text{ for some } j > k_n \Big\}$$

and

$$\tilde{A}_4 = \Big\{ |\zeta_j| \geq \frac{\lambda_n}{4} \widehat{w}_j \text{ for some } j > k_n \Big\}.$$

Now $P(\tilde{A}_j) \to 0$ for $j = 1, \ldots, 4$ follows with less complexity analogously to $P(A_j) \to 0$ in the proof of Theorem 3.1. This proves the assertion of the theorem. □

ACKNOWLEDGMENTS

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 and an unknown referee for constructive comments on an earlier version of this manuscript.

REFERENCES

1. R. Bhatia, *Matrix Analysis* (Springer, New York, 1997).
2. P. Craven and G. Wahba, "Smoothing Noisy Data with Spline Function: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross Validation", Numerische Mathematik **31**, 337–403 (1979).
3. H. Dette and A. Munk, "Testing Heteroscedasticity in Nonparametric Regression", J. Roy. Statist. Soc. Ser. B **60**, 693–708 (1998).
4. B. Efron, T. Hastie, and R. Tibshirani, "Least Angle Regression (with Discussion)", Ann. Statist. **32**, 407–451 (2004).
5. J. Fan and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties", J. Amer. Statist. Assoc. **96**, 1348–1360 (2001).
6. J. Fan and J. Lv, "A Selective Overview of Variable Selection in High-Dimensional Feature Space", Statistica Sinica **20**, 101–148 (2010).
7. J. Fan and H. Peng, "Nonconcave Penalized Likelihood with a Diverging Number of Parameters", Ann. Statist. **32**, 928–961 (2004).
8. I. E. Frank and J. H. Friedman, "A Statistical View of Some Chemometrics Regression Tools (with Discussion)", Technometrics **35**, 109–148 (1993).
9. A. Hörl and R. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics, **12**, 55–67 (1970).
10. J. Huang, J. L. Horowitz, and S. Ma, "Asymptotic Properties of Bridge Estimators in Sparse High Dimensional Regression Models", Ann. Statist. **36**, 587–613 (2008).
11. J. Huang, S. Ma, and C. Zhang, *Adaptive Lasso for Sparse High-Dimensional Regression Models*, Technical Report, Univ. of Iowa (2006).
12. J. Huang, S. Ma, and C. Zhang, "Adaptive Lasso for Sparse High-Dimensional Regression Models", Statistica Sinica **18**, 1603–1618 (2008).
13. Y. Kim, H. Choi, and H.-S. Oh, "Smoothly Clipped Absolute Deviation on High Dimensions", J. Amer. Statist. Assoc. **103**, 1665–1673 (2008).
14. K. Knight and W. Fu, (2000). "Asymptotics for Lasso-type Estimators", Ann. Statist, **28**, 1356–1378 (2000).
15. H. Leeb and B. M. Pötscher, "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator", J. Econometrics **142**, 201–211 (2008).
16. B. M. Pötscher and U. Schneider, "Distributional Results for Thresholding Estimators in High-Dimensional Gaussian Regression Models", Electronic J. Statist. **5**, 1876–1934 (2011).
17. R Development Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2009) ISBN 3-900051-07-0.
18. R. Tibshirani, "Regression Shrinkage and Selection via the Lasso", J. Roy. Statist. Soc. B **58**, 267–288 (1996).
19. A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, in *Springer Series in Statist.* (Springer, New York, 1996).
20. J. Wagener and H. Dette, "Bridge Estimators and the Adaptive Lasso under Heteroscedasticity", Math. Methods Statist. **21**, 109–126 (2012).
21. M. J. Wainwright, "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using l_1-Constrained Quadratic Programming (Lasso)", IEEE Trans. Inform. Theory **55**, 2183–2202 (2009).
22. H. Zou, "The Adaptive Lasso and Its Oracle Properties", J. Amer. Statist. Assoc. **101**, 1418–1429 (2006).
23. P. Zhao and B. Yu, "On Model Selection Consistency of Lasso", J. Machine Learning Research **7**, 2541–2563 (2006).