High-dimensional, robust, heteroscedastic variable selection with the adaptive LASSO, and applications to random coefficient regression

Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Philipps-Universität Marburg

vorgelegt von

Philipp Hermann
Master of Science
aus Dresden

Marburg, 2021

Erstgutachter und Betreuer: Prof. Dr. Hajo Holzmann
Zweitgutachter: Prof. Dr. Martin Spindler
Drittgutachter: Prof. Dr. Christoph Breunig

Tag der Einreichung: 8. Januar 2021
Tag der Disputation: 19. April 2021

Erscheinungsort und -jahr: Marburg, 2021
Hochschulkennziffer: 1180

In loving memory of my grandfather Hans.

Acknowledgements

First and foremost, I want to thank my doctoral supervisor Hajo Holzmann for proposing the problem and suggesting methods, many fruitful discussions and helping me to improve my understanding of the field significantly.

Moreover, I thank Professor Martin Spindler for making the second assessment of this thesis.

I would like to thank my colleagues at the Department of Mathematics, especially Pavel Tafo, and my former colleagues Markus Bibinger, Heiko Werner and Viktor Bengs, for the always pleasant working atmosphere and making the last few years very enjoyable.

Great thanks to my family for their support during my studies.

Contents

Notations

Asymptotic growth rates

$a_n \simeq b_n$	The sequences a_n and b_n are of the same order.
$a_n \lesssim b_n$	The sequence a_n is smaller than the sequence b_n up to multiplicative constants.
$a_n = \mathcal{O}\left(b_n\right)$	Big O notation. The sequence a_n is smaller than the sequence b_n up to multiplicative constants.

Functions

$\mathbb{1}\{\,\cdot\,\}$	Indicator function.
$\Gamma(\,\cdot\,)$	Gamma function.
$\lvert A \rvert$	Cardinality of the set A.
$\operatorname{supp}(v)$	Support of the vector v. That is $\{k \in \{1,\dots,d\} \,\vert\, v_k \neq 0\}$ if $v \in \mathbb{R}^d$.
$\operatorname{sign}(v)$	Vector which contains the signs, encoded by $\{-1, 0, 1\}$, of the vector v. It is $\operatorname{sign}(v_k) = \mathbb{1}\{v_k > 0\} - \mathbb{1}\{v_k < 0\}$.
$v_1 \odot v_2$	Hadamard product (component-wise product) of the vectors v_1 and v_2.
$\langle v_1, v_2 \rangle$	Euclidean inner product of the vectors v_1 and v_2.
$\operatorname{vec}(M)$	Half-vectorization of the matrix M.

Norms

$\lVert v \rVert_1$	ℓ_1 norm, Manhattan norm of the vector v.
$\lVert v \rVert_2$	ℓ_2 norm, Euclidean norm of the vector v.
$\lVert v \rVert_\infty$	ℓ_∞ norm, maximum norm of the vector v.
$\lVert M \rVert_{\mathrm{M},2}$	ℓ_2 operator norm, spectral norm of the matrix M.

$\|M\|_{\mathrm{M},\infty}$	ℓ_∞ operator norm, row sum norm of the matrix M.

Operators

∇f	Gradient of the smooth function f.
∂f	Subdifferential of the convex function f.

Probability

$\mathbb{P}$	Probability.
$\mathbb{E}$	Expected value.
$\mathbb{V}\mathrm{ar}, \mathbb{C}\mathrm{ov}$	Variance, covariance.
$\mathbb{C}\mathrm{or}$	Correlation.
$\xrightarrow{\mathbb{P}}$	Convergence in probability.
$\xrightarrow{a.s.}$	Almost sure convergence.
$\xrightarrow{d}$	Convergence in distribution.
$\mathcal{U}[a, b]$	Uniform distribution on the interval $[a, b]$.
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2.
$\mathcal{N}_d(\mu, \Sigma)$	Multivariate normal distribution with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$.
$X_n = \mathrm{o}_{\mathbb{P}}(a_n)$	Small O in probability notation. The sequence (X_n/a_n) converges in probability to zero. For random vectors with respect to the ℓ_∞ norm (or any other vector norm).
$X_n = \mathcal{O}_{\mathbb{P}}(a_n)$	Big O in probability notation. The sequence (X_n/a_n) is tight/ stochastically bounded. For random vectors with respect to the ℓ_∞ norm (or any other vector norm).

Sets

$\mathbb{N}$	The set of natural numbers.
$\mathbb{R}$	The set of real numbers.
A^c	Relative complement of the set A.

Vectors/Matrices

e_k	The k^{th} unit vector, with k^{th} coordinate equal to 1, and zero entries otherwise. The dimension of e_k will depend on and be clear from the context.
$\mathbf{0}_d$	Zero vector of dimension d.
v_A	v_A denotes either the subvector $(v_k)_{k \in A} \in \mathbb{R}^{\lvert A \rvert}$ or the vector $v_A \in \mathbb{R}^d$ with $(v_A)_k = v_k$, $k \in A$, and $(v_A)_k = 0$, $k \in \{1, \ldots, d\} \setminus A$ if $v \in \mathbb{R}^d$ and $A \subseteq \{1, \ldots, d\}$.
$<, \leq$	Inequality signs are understood component-wise for vectors.
$\lvert v \rvert$	The absolute value is understood component-wise for vectors. That is $(\lvert v_1 \rvert, \ldots, \lvert v_d \rvert)^\top$ if $v \in \mathbb{R}^d$.
$1/v$	The reciprocal is understood component-wise for vectors. That is $(1/v_1, \ldots, 1/v_d)^\top$ if $v \in \mathbb{R}^d$.
I_d	Identity matrix of dimension $d \times d$.
$\mathbf{0}_{d_1 \times d_2}$	Zero matrix of dimension $d_1 \times d_2$.
$\operatorname{diag}(v_1, \ldots, v_d)$	Diagonal matrix with entries $v_1, \ldots, v_d$.
$M^\top$	Transpose of the matrix M.
M^{-1}	Inverse of the matrix M.
M_A	$M_A \in \mathbb{R}^{d_1 \times \lvert A \rvert}$ has entries of M according to column indices in A if $M \in \mathbb{R}^{d_1 \times d_2}$ and $A \subseteq \{1, \ldots, d_2\}$.
M_{AB}	$M_{AB} \in \mathbb{R}^{\lvert A \rvert \times \lvert B \rvert}$ has entries of M according to row indices in A and column indices in B if $M \in \mathbb{R}^{d_1 \times d_2}$ and $A \subseteq \{1, \ldots, d_1\}, B \subseteq \{1, \ldots, d_2\}$.

1. Introduction

Modeling the linear relationship between a response variable and some explanatory variables has been of mayor interest in statistics over the last decades. Practical applications cover a wide range of areas such as behavioral and social sciences, finance and economics. However, especially in consumer data analysis and medicine the marginal effects can vary across the individuals. Regression models with random coefficients are very useful for analyzing and modeling this unobserved heterogeneity. Hildreth and Houck (1968) and Swami (1970) considered appropriate linear models from a parametric point of view and proposed some consistent estimators for the means and variances of the coefficients, assuming that the covariances vanish. In addition much research in the area of nonparametric identification and estimation of the joint distribution of the coefficients has been done over the past decades, e.g. Beran and Hall (1992), Beran and Millar (1994), Beran et al. (1996), Hoderlein et al. (2010), Dunker et al. (2019) and Holzmann and Meister (2020). Furthermore, Lewbel and Pendakur (2017) considered nonlinear and additive models, Ichimura and Thompson (1998) and Gautier and Kitamura (2013) binary choice models, and Gautier and Hoderlein (2011) and Hoderlein et al. (2017) triangular models with random coefficients.

In this thesis we consider the linear random coefficient regression model, and especially the means, variances and covariances of the coefficients, which are arguably of most interest in many applications, in a high-dimensional framework with focus on variable selection. This means that the number of regressors can exceed the number of observations, however, only a few of them have influence and/or heterogeneous effects. High-dimensional statistics in general gained a lot of attention over the last years since larger data sets with huge numbers of features are collected in many industrial and scientific fields, e.g. in microarray data analysis, functional magnetic resonance imaging or consumer data analysis. A broad overview about theory and methods in this topic can be found in Bühlmann and van de Geer (2011), Giraud (2014), Hastie et al. (2015), Vershynin (2018) and Wainwright (2019).

A very common and effective tool for variable selection in high-dimensional, sparse regression models are estimators with penalization functions. An important and well-studied one is the LASSO, which was proposed by Tibshirani (1996) and combines the empirical quadratic loss function with the ℓ_1-penalization. Oracle inequalities in linear regression models with independent and normally distributed errors for the LASSO are provided by Bickel et al. (2009) and Meinshausen and Yu (2009). In order to do this an assumption on the design matrix is always necessary, van de Geer and Bühlmann (2009) and Foucart and Rauhut (2013) discuss several of these assumptions and their relationship to each other. Moreover, Zhao and Yu (2006) showed that for sign-consistency of the LASSO an additional assumption, commonly called irrepresentable or mutual incoherence condition, is required, and Wainwright (2009b) introduced the primal-dual witness characterization of the LASSO and gives sufficient and necessary conditions for sparsity recovery under independent sub-Gaussian errors. A line of research, followed e.g. in Wainwright (2009a), then investigated minimal conditions under which for certain design matrices, consisting for example of independent and identically distributed Gaussian entries, support recovery is possible for distinct constellations of the numbers of observations and regressors, the order of sparsity and the minimal non-zero entry of the coefficient vector in absolute value. Comprehensive results in this direction, which even include non-Gaussian, heavy-tailed errors, are provided in Ndaoud and Tsybakov (2020). Another line of investigation tries to get rid of the mutual incoherence condition for variable selection. In this context Zou (2006) proposed the adaptive LASSO which enjoys additionally the oracle properties under homoscedasticity and a fixed number of coefficients. For a growing dimension Huang et al. (2008) provide asymptotics, Wagener and Dette (2012) and Wagener and Dette (2013) extend these asymptotic results for heteroscedastic errors. In addition Zhou et al. (2009) and van de Geer et al. (2011) considered the adaptive LASSO in the high-dimensional framework under independent and identically normally distributed errors. Moreover, Loh and Wainwright (2017) provide a high-dimensional analysis of nonconvex penalizations, such as smoothly clipped absolute deviation (Fan and Li, 2001, SCAD) or minimax concave penalty (Zhang, 2010, MCP), to drop the mutual incoherence condition.

If the number of regressors exceeds the number of observations, most of the aforementioned theory is based on a sub-Gaussian tail inequality for the independent errors in the linear regression model. Dropping this light tail assumption may lead to sub-optimal rates for the ℓ_1, ℓ_2 and ℓ_∞ norm of the estimation error. However, by the results in Lederer and Vogt (2020) it follows that the ordinary LASSO retains the optimal rates from the light-tail case if the covariates are uniformly bounded and the errors have slightly more than a finite fourth moment. Especially in the linear random coefficient regression model the estimation of the variances and covariances leads to a heteroscedastic mean regression model where the errors contain the centered squares and pairwise products of the coefficients. Hence, if we assume a sub-Gaussian distribution for the coefficients, the aforementioned errors are potentially sub-Exponential.

Nowadays a current strand of research deals with robustifying the available methodology in the high-dimensional framework with respect to deviations from light-tail assumptions on the errors, and sometimes also on the predictors. One common approach is to replace the squared loss function by some other, fixed, robust loss function such as the check function from quantile regression and in particular absolute deviation for the median (Li and Zhu, 2008; Zou and Yuan, 2008; Belloni and Chernozhukov, 2011; Wang, 2013; Fan et al., 2014). However, doing so generally changes the target parameter away from the mean, particularly in the random design regression models with potentially heteroscedastic, asymmetric errors. Furthermore, Loh (2017) considered homoscedastic models with independent covariates and errors in scenarios where a fixed, robust loss function gives the desired mean parameter.

Another approach, proposed by Lambert-Lacroix and Zwald (2011), Fan et al. (2017)

and Sun et al. (2020), is to use the Huber loss function (Huber, 1964) with an additional tuning parameter. The Huber loss combines squared loss for small values and absolute loss for larger ones. The tuning parameter, we call it robustification parameter, is necessary to control the bias since the estimation error contains in general also an approximation error. Lambert-Lacroix and Zwald (2011) provide asymptotic results for the adaptive LASSO with a fixed choice of the robustification parameter under symmetric errors. If the tuning parameter converges with an appropriate rate, depending on the sample size and dimension of the coefficient vector, the LASSO with Huber loss achieves in high-dimensional, heteroscedastic linear regression models with sub-Gaussian regressors and errors with finite second moments the same rates in ℓ_1 and ℓ_2 norm as the ordinary LASSO under homoscedastic, light-tailed errors. Non-asymptotic upper and lower bounds are provided in Fan et al. (2017) and Sun et al. (2020). However, variable selection and the ℓ_∞ norm of the estimation error have not been studied in this framework to the best of our knowledge yet. We consider for that purpose a strictly convex, smooth variant of the Huber loss function and the adaptive LASSO penalty for computational efficiency. For the resulting estimator we show in the first part of this thesis sign-consistency and also optimal rates of convergence in the ℓ_∞ norm as in the homoscedastic, light-tailed setting.

The thesis is structured as follows. At the beginning we introduced the basic notation needed throughout the work. In Chapter 2 we give a brief overview about sign-consistency in high-dimensional, homoscedastic linear regression models, and motivate the necessity of similar results for heteroscedastic errors to perform variable selection for the first and second moments in linear random coefficient regression models. In Chapter 3 we introduce the pseudo Huber loss function, and show sign-consistency and optimal rates in ℓ_∞ norm for the adaptive LASSO in heteroscedastic linear mean regression models with sub-Gaussian regressors and errors with slightly more than a finite second moment. Simulations illustrate the favorable numerical performance of the proposed methodology in comparison to the ordinary adaptive LASSO. The results of Chapter 3 are also provided in Hermann and Holzmann (2020). In the second part of the thesis we consider the linear random coefficient regression model and, in particular, the means, variances and covariances of the coefficients. Firstly, we give in Chapter 4 sufficient conditions for the identifiability of the first and second moments. In doing so we focus on situations of regressors having potentially bounded or even finite support, which is in contrast to the large support required for nonparametric identification of the joint distribution of the coefficients. In Chapter 5 we establish at first the sparse, heteroscedastic linear regression models of the first and second moments of the random coefficients. Later on, we proceed with asymptotic results for the appropriate adaptive LASSO estimators if the number of coefficients is fixed. Support estimation is our main goal again. Finally, in Chapter 6 we apply the methods of Chapter 3 to the high-dimensional regression models of the moments of the coefficients and discuss remaining issues.

2. Preliminaries

In this chapter we give a brief overview about variable selection in high-dimensional linear regression models. For that purpose we repeat some suitable estimators and results of the recent literature, where a sub-Gaussian distribution assumption on the errors plays an crucial role. We omit an introduction of this class of distributions here, and refer to Wainwright (2019, Chapter 2) and Vershynin (2018, Chapter 2) for a broad overview. Furthermore, we outline a current strand of research which deals with robustifying the available methodology, especially with respect to deviations from the light tail assumption on the errors. Finally, we introduce the linear regression model with random coefficients and emphasize the importance of sign-consistency results in heteroscedastic linear regression models if we are interested in the means, variances and covariances of the coefficients.

This chapter is structured as follows. In Section 2.1 we introduce the well-known linear regression model and in Section 2.2 we display the idea of variable selection in high-dimensional sparse models. Moreover, we discuss recent results in this framework and state the ordinary LASSO and adaptive LASSO. Later on, we proceed with a discussion about robust regularization methods in Section 2.3. In the end, in Section 2.4 we introduce briefly the linear regression model with random coefficients.

2.1. Linear regression models

Assume one observes the pairs $(Y_1, \mathbf{x_1}^\top)^\top, \dots, (Y_n, \mathbf{x_n}^\top)^\top$ of data according to the linear regression model

$$Y_i = \mathbf{x_i}^\top \beta^* + \varepsilon_i \,, \qquad i = 1, \dots, n \,, \tag{2.1}$$

where $Y_i \in \mathbb{R}$ are the response variables, $\mathbf{x_i} \in \mathbb{R}^p$ the regressors, $\beta^* \in \mathbb{R}^p$ the unknown and fixed coefficient vector, and $\varepsilon_i \in \mathbb{R}$ additive noise modeled through random variables with $\mathbb{E}[\varepsilon_i] = 0$. We call the model homoscedastic if $\mathbb{V}\mathrm{ar}(\varepsilon_i) = \omega^2$ for $\omega > 0$ and all $i \in \{1, \dots, n\}$, otherwise the errors are called heteroscedastic. Let $\mathbb{Y}_n = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$, $\mathbb{X}_n = [\mathbf{x_1}, \dots, \mathbf{x_n}]^\top \in \mathbb{R}^{n \times p}$ the design matrix and $\vec{\varepsilon}_n = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$, then the linear regression model (2.1) can be written in matrix notation as

$$\mathbb{Y}_n = \mathbb{X}_n \, \beta^* + \vec{\varepsilon}_n \,.$$

Common estimators of the coefficients β^* are the least squares estimator and the generalized least squares estimator for heteroscedastic and correlated errors. Let

$$\mathcal{L}_n^{\mathrm{LS}}(\beta) := \frac{1}{n} \, \|\mathbb{Y}_n - \mathbb{X}_n \, \beta\|_2^2 \,, \qquad \beta \in \mathbb{R}^p \,, \tag{2.2}$$

be the empirical quadratic loss function, then the least squares estimator is given by

$$\widehat{\beta}_n^{\mathrm{LS}} \in \rho_n^{\mathrm{LS}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \mathcal{L}_n^{\mathrm{LS}}\big(\beta\big) .$$

Furthermore, if the covariance matrix $\Omega_n = \mathbb{C}\mathrm{ov}(\vec{\varepsilon}_n)$ of the errors is known and positive definite, we can define the generalized least squares estimator by

$$\widehat{\beta}_n^{\mathrm{GLS}} \in \rho_n^{\mathrm{GLS}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{n} \left\| \Omega_n^{-1/2}\big(\mathbb{Y}_n - \mathbb{X}_n\,\beta\big)\right\|_2^2 .$$

A definition of the expression $\Omega_n^{-1/2}$ is given subsequently.

Definition 2.1. *Let $M \in \mathbb{R}^{d\times d}$ be a symmetric and positive definite matrix, then the matrix is diagonalizable with $M = T\,D\,T^\top$ where $D = \operatorname{diag}(\tau_1, \dots, \tau_d) \in \mathbb{R}^{d\times d}$ is a diagonal matrix and $T \in \mathbb{R}^{d\times d}$ a orthogonal matrix. We define*

$$M^{-1/2} := T\,D^{-1/2}\,T^\top$$

with $D^{-1/2} = \operatorname{diag}\big(\tau_1^{-1/2}, \dots, \tau_d^{-1/2}\big)$.

Note that for uncorrelated and homoscedastic errors with $\mathbb{V}\mathrm{ar}(\varepsilon_i) = \omega^2$ for $\omega > 0$ the generalized least squares estimator is equal to the least squares estimator since $\Omega_n = \omega^2\,\mathrm{I}_n$, and hence

$$\operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{n} \left\| \Omega_n^{-1/2}\big(\mathbb{Y}_n - \mathbb{X}_n\,\beta\big)\right\|_2^2 = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{\omega^2\,n} \|\mathbb{Y}_n - \mathbb{X}_n\,\beta\|_2^2 = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \mathcal{L}_n^{\mathrm{LS}}\big(\beta\big) .$$

As a result the estimator is independent of the knowledge of the error level ω^2.

Remark 2.2. Consider the linear regression model (2.1) and suppose that the number p of coefficients is smaller or equal to the number n of observations, and that the design matrix $\mathbb{X}_n$ has full rank p. Then the least squares estimator is unique and can be expressed by

$$\widehat{\beta}_n^{\mathrm{LS}} = \big(\mathbb{X}_n^\top \mathbb{X}_n\big)^{-1} \mathbb{X}_n^\top \mathbb{Y}_n .$$

In addition the generalized least squares estimator is also unique and has the analogous form

$$\widehat{\beta}_n^{\mathrm{GLS}} = \big(\mathbb{X}_n^\top \Omega_n^{-1} \mathbb{X}_n\big)^{-1} \mathbb{X}_n^\top \Omega_n^{-1} \mathbb{Y}_n .$$

2.2. Variable selection in the high-dimensional framework

Now we consider the linear regression model (2.1) in a high-dimensional framework. This means that the number n of observations can be smaller than the dimension p of the regressors, however, only a few of them have influence. This is formalized by sparsity of the coefficient vector β^*. For that purpose let

$$S := \operatorname{supp}\big(\beta^*\big) = \Big\{ k \in \{1, \ldots, p\} \,\Big|\, \beta^*_k \neq 0 \Big\}$$

be the support of β^* and $s := |S|$ the number of coefficients unequal to zero. Furthermore, let $S^c := \{1, \ldots, p\} \setminus S$ the relative complement of S. Sparsity then means that $s < p$ is satisfied, and hence the linear regression model (2.1) can also be expressed by

$$\mathbb{Y}_n = \mathbb{X}_{n,S}\, \beta^*_S + \vec{\varepsilon}_n \,.$$

One major goal in this framework is variable selection, meaning to find estimators $\widehat{\beta}_n$ and appropriate conditions so that $\widehat{\beta}_{n,S^c} = \mathbf{0}_{p-s}$ holds with high probability. A further extended property is the so-called sign-consistency of an estimator, which includes variable selection and in addition the estimation of the true signs on the support. Such results depend on the smallest absolute value of the entries of β^* on its support S, hence let

$$\beta^*_{\min} := \min_{k \in S} \big|\beta^*_k\big| \,.$$

Common and well-studied estimators to perform variable selection are the ordinary LASSO (Tibshirani, 1996) and the adaptive LASSO (Zou, 2006) which are based on the empirical quadratic loss function and a (adaptive) ℓ_1-penalization. The LASSO is given in the Lagrangian form by

$$\widehat{\beta}^{\mathrm{L}}_n \in \rho^{\mathrm{L}}_{n,\lambda_n} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \Big(\mathcal{L}^{\mathrm{LS}}_n\big(\beta\big) + 2\lambda_n \left\|\beta\right\|_1 \Big) \,,$$

where $\lambda_n > 0$ is the regularization parameter and the loss function $\mathcal{L}^{\mathrm{LS}}_n$ is given in (2.2), and the adaptive LASSO by

$$\widehat{\beta}^{\mathrm{AL}}_n \in \rho^{\mathrm{AL}}_{n,\lambda_n} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \Bigg(\mathcal{L}^{\mathrm{LS}}_n\big(\beta\big) + 2\lambda_n \sum_{k=1}^{p} \frac{|\beta_k|}{\big|\widehat{\beta}^{\mathrm{init}}_{n,k}\big|} \Bigg) \,,$$

where $\widehat{\beta}^{\mathrm{init}}_n = (\widehat{\beta}^{\mathrm{init}}_{n,1}, \ldots, \widehat{\beta}^{\mathrm{init}}_{n,p})^\top \in \mathbb{R}^p$ is an initial estimator of β^*. If $\widehat{\beta}^{\mathrm{init}}_{n,k} = 0$, we require $\beta_k = 0$ in the above definition. These estimators are regularized M-estimators of the coefficients β^* and solutions to convex optimization problems, or more specifically to quadratic programs with convex constraints. Numerically these problems are solved with the so-called coordinate descent algorithm, see for example Friedman et al. (2010) and Hastie et al. (2015, Section 2.4) for more details.

To give results on sign-consistency, Wainwright (2009b) introduced the primal-dual witness characterization of the LASSO and its adaptive version, which we provide in the following lemmas.

Lemma 2.3 (Primal-dual witness characterization of the LASSO)**.** *Assume that* $s \leq n$ *and* $\operatorname{rank}(\mathbb{X}_{n,S}) = s$ *hold. Furthermore, if*

$$\left\| \mathbb{X}_{n,S^c}^\top \mathbb{X}_{n,S} \left(\mathbb{X}_{n,S}^\top \mathbb{X}_{n,S} \right)^{-1} \operatorname{sign}\big(\beta_S^*\big) + \frac{1}{n\,\lambda_n} \mathbb{X}_{n,S^c}^\top \mathrm{P}_{\mathbb{X}_{\mathrm{n,S}}^\perp}\, \varepsilon_n \right\|_\infty < 1 \tag{2.3}$$

with

$$\mathrm{P}_{\mathbb{X}_{\mathrm{n,S}}^\perp} = \mathrm{I}_n - \mathbb{X}_{n,S} \left(\mathbb{X}_{n,S}^\top \mathbb{X}_{n,S} \right)^{-1} \mathbb{X}_{n,S}^\top$$

holds, and

$$\widetilde{\beta}_{n,S} = \beta_S^* + \left(\frac{1}{n} \mathbb{X}_{n,S}^\top \mathbb{X}_{n,S} \right)^{-1} \left(\frac{1}{n} \mathbb{X}_{n,S}^\top \varepsilon_n - \lambda_n \operatorname{sign}\big(\beta_S^*\big) \right)$$

satisfies $\operatorname{sign}\big(\widetilde{\beta}_{n,S}\big) = \operatorname{sign}\big(\beta_S^*\big)$*, then the unique LASSO solution* $\rho_{n,\lambda_n}^{\mathrm{L}} = \big\{\widehat{\beta}_n^{\mathrm{L}}\big\}$ *satisfies*

$$\operatorname{sign}\big(\widehat{\beta}_n^{\mathrm{L}}\big) = \operatorname{sign}\big(\beta^*\big)\,, \quad \widehat{\beta}_{n,S}^{\mathrm{L}} = \widetilde{\beta}_{n,S} \quad \text{and} \quad \widehat{\beta}_{n,S^c}^{\mathrm{L}} = \mathbf{0}_{p-s}\,.$$

Proof. Cf. Wainwright (2009b, Lemma 3). □

Lemma 2.4 (Primal-dual witness characterization of the adaptive LASSO)**.** *Assume that* $s \leq n$ *and* $\operatorname{rank}(\mathbb{X}_{n,S}) = s$ *hold. Furthermore, if*

$$\left| \mathbb{X}_{n,S^c}^\top \mathbb{X}_{n,S} \left(\mathbb{X}_{n,S}^\top \mathbb{X}_{n,S} \right)^{-1} \lambda_n \left(\frac{1}{|\widehat{\beta}_{n,S}^{\,\mathrm{init}}|} \odot \operatorname{sign}\big(\beta_S^*\big) \right) + \frac{1}{n} \mathbb{X}_{n,S^c}^\top \mathrm{P}_{\mathbb{X}_{\mathrm{n,S}}^\perp}\, \varepsilon_n \right| < \frac{\lambda_n}{|\widehat{\beta}_{n,S^c}^{\,\mathrm{init}}|} \tag{2.4}$$

with

$$\mathrm{P}_{\mathbb{X}_{\mathrm{n,S}}^\perp} = \mathrm{I}_n - \mathbb{X}_{n,S} \left(\mathbb{X}_{n,S}^\top \mathbb{X}_{n,S} \right)^{-1} \mathbb{X}_{n,S}^\top$$

holds, and

$$\widetilde{\beta}_{n,S} = \beta_S^* + \left(\frac{1}{n} \mathbb{X}_{n,S}^\top \mathbb{X}_{n,S} \right)^{-1} \left(\frac{1}{n} \mathbb{X}_{n,S}^\top \varepsilon_n - \lambda_n \left(\frac{1}{|\widehat{\beta}_{n,S}^{\,\mathrm{init}}|} \odot \operatorname{sign}\big(\beta_S^*\big) \right) \right)$$

satisfies $\operatorname{sign}\big(\widetilde{\beta}_{n,S}\big) = \operatorname{sign}\big(\beta_S^*\big)$*, then the unique adaptive LASSO solution* $\rho_{n,\lambda_n}^{\mathrm{AL}} = \big\{\widehat{\beta}_n^{\mathrm{AL}}\big\}$ *satisfies*

$$\operatorname{sign}\big(\widehat{\beta}_n^{\mathrm{AL}}\big) = \operatorname{sign}\big(\beta^*\big)\,, \quad \widehat{\beta}_{n,S}^{\mathrm{AL}} = \widetilde{\beta}_{n,S} \quad \text{and} \quad \widehat{\beta}_{n,S^c}^{\mathrm{AL}} = \mathbf{0}_{p-s}\,.$$

Proof. Cf. Zhou et al. (2009, Lemma 12.1) with $\vec{w} = \big(1/|\widehat{\beta}_{n,1}^{\,\mathrm{init}}|, \ldots, 1/|\widehat{\beta}_{n,p}^{\,\mathrm{init}}|\big)^\top \in \mathbb{R}^p$. □

In the subsequent remark we summarize the non-asymptotic results for variable selection in a high-dimensional framework with $p \geq n$. In doing so we omit the ordinary LASSO since then a further assumption, called irrepresentable or mutual incoherence condition, is needed (Zhao and Yu, 2006).

Remark 2.5. Zhou et al. (2009) considered the adaptive LASSO with the LASSO as initial estimator in a homoscedastic linear regression model with normally distributed errors and a random design where the regressors are also independent and identically normally distributed with mean zero and covariance matrix Π. Under a restricted eigenvalue assumption on Π, and the scaling $n \gtrsim s^2 \log(p)$ and $\beta^*_{\min} \gtrsim s\,(\log(p)/n)^{\frac{1}{2}}$ the adaptive LASSO with an appropriate choice of the regularization parameter is unique and sign-consistent with probability at least $1-3/p^2$ in this framework (Zhou et al., 2009, Theorem 5.3). The corresponding proofs show that also $\|\widehat{\beta}^{\mathrm{AL}}_n - \beta^*\|_\infty \lesssim (s\log(p)/n)^{\frac{1}{2}}$ holds with high probability. Moreover, the authors provide analogous results for fixed designs. In a similar regime with independent and sub-Gaussian regressors and errors Loh and Wainwright (2017) proposed estimators based on the empirical quadratic loss function and a nonconvex regularizer such as smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP). If the minimal eigenvalue of the covariance matrix of the regressors is bounded below and the ℓ_∞ operator norm of the inverse of the sample covariance matrix of the covariates is upper bounded by a positive constant, then these estimators, with an appropriate choice of the regularization parameter, are sign-consistent and the ℓ_∞ norm of the estimation error has an order of $(\log(p)/n)^{\frac{1}{2}}$ with high probability under the scaling $\beta^*_{\min} \gtrsim (\log(p)/n)^{\frac{1}{2}}$ and $n \gtrsim s\log(p)$ (Loh and Wainwright, 2017, Corollary 1). In comparison to the adaptive LASSO variable selection is provided under a weaker beta-min condition and a smaller sample size, but the adaptive LASSO is computationally more efficient. Moreover, by the additional bound on the inverse of the sample covariance matrix of the regressors Loh and Wainwright (2017) achieve a faster rate for the ℓ_∞ norm of the estimation error. In a constructive work Loh (2017) allows additionally for heavy-tailed, symmetric errors, but independent of the regressors as well, and considered robust loss functions such as Huber, Tukey and Cauchy loss combined with nonconvex regularizers. Under the additional scaling $n \gtrsim \max(s^2, s\log(p))$ the resulting estimators are sill sign-consistent (Loh, 2017, Theorem 2).

Variable selection in high-dimensional linear regression models with heteroscedastic and potentially heavy tailed errors has not been studied sufficiently to the best of our knowledge yet, which we want to highlight at this point. We will see in Section 2.4 that especially for the estimation of the first and second moments in the linear random coefficient model heteroscedastic regression models play an important role.

2.3. Robust regularization methods

It is well-known that the LASSO achieves the optimal estimation rates in ℓ_1, ℓ_2 and ℓ_∞ norm, and enjoys sign-consistency in high-dimensional, homoscedastic linear regression models if the errors are normal or at least sub-Gaussian (Meinshausen and Yu, 2009; Bickel et al., 2009; Wainwright, 2009b). These results rely heavily on the light tail assumption, and are very sensitive to violations. Moreover, the quadratic loss function is in general also sensitive to outliers in the response variable, and in the regressors as well. In the recent literature robust regression methods have been developed to preserve variable selection and the optimal estimation rates for a broader class of error distributions with potentially heavy tails in the high-dimensional framework. For that purpose various robust loss functions have been considered, for example one major line is least absolute deviation, or more general quantile regression (Belloni and Chernozhukov, 2011; Wang et al., 2012; Wang, 2013; Fan et al., 2014). Another approach with a broader class of loss functions, including Huber, Tukey and Cauchy loss, is provided by Loh (2017). Note that the target parameter of these methods may differ from the mean parameter β^*, especially for asymmetric and/or heteroscedastic errors, and hence an additional approximation error is generated in these settings. In the following remark we give an intuition for the robust regularization methods proposed by Loh (2017).

Remark 2.6. Consider the linear regression model (2.1), and let $l : \mathbb{R} \to \mathbb{R}$ denote a differentiable loss function and

$$\mathcal{L}_n\big(\beta\big) = \frac{1}{n}\sum_{i=1}^{n} l\big(Y_i - \mathbf{x_i}^\top \beta\big)$$

the associated empirical loss function. Then a natural regularized M-estimator is given by

$$\widehat{\beta} \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \Big(\mathcal{L}_n\big(\beta\big) + \rho_{\lambda_n}\big(\beta\big)\Big)\,,$$

where $\rho_{\lambda_n} : \mathbb{R}^p \to \mathbb{R}$ is a penalization function with regularization parameter $\lambda_n > 0$. One crucial task to achieve optimal estimation rates for the parameter β^* is to control the rate of the ℓ_∞ norm of the gradient $\nabla\mathcal{L}_n(\beta^*)$ of the empirical loss function, cf. Loh (2017, Theorem 1) for nonconvex regularizers and Negahban et al. (2012, Theorem 1) for decomposable norms, such as the LASSO penalization, as regularizers. Evidently, the gradient is given by

$$\nabla\mathcal{L}_n\big(\beta^*\big) = -\frac{1}{n}\sum_{i=1}^{n} l'\big(Y_i - \mathbf{x_i}^\top\beta^*\big)\,\mathbf{x_i} = -\frac{1}{n}\sum_{i=1}^{n} l'\big(\varepsilon_i\big)\,\mathbf{x_i}\,.$$

If the loss function $l(x) = x^2$ is the quadratic loss function, the factors $l'(\varepsilon_i)$ in the sum are equal to $2\,\varepsilon_i$, which means that also extreme observations have influence on the above quantity. To get in the high-dimensional framework with $p \geq n$ a required bound of the form

$$\big\|\nabla\mathcal{L}_n\big(\beta^*\big)\big\|_\infty \leq C\bigg(\frac{\log(p)}{n}\bigg)^{\frac{1}{2}}$$

with high probability for a positive constant $C > 0$, a sub-Gaussian distribution for the errors is sufficient if they are homoscedastic (Negahban et al., 2012, Corollary 2). However, if we choose a loss function l with a upper bounded first derivative, the contribution of extreme realizations of the errors ε_i to the gradient is bounded as well. Hence the quantity is more robust in terms of error distribution and potential outliers then. All of the mentioned loss functions at the beginning of this section satisfy this condition. Furthermore, if the derivative of the loss function is equal to zero for very large values, extreme outliers have in fact no influence on the gradient. However, note that such functions have to be nonconvex, which leads to an additional cost in computational time. In the robust regression literature these estimators are called redescending M-estimators (Loh, 2017).

It is obvious that for homoscedastic errors and centered regressors the target parameter of the robust regularization methods in Remark 2.6 is the mean vector β^* since $\mathbb{E}[\nabla \mathcal{L}_n(\beta^*)] = \mathbf{0}_p$ holds. Hence the theory of Loh (2017) applies and we obtain sufficient conditions for high-dimensional variable selection and optimal estimation rates in sparse, homoscedastic linear regression models.

In more general mean regression models Fan et al. (2017) and Sun et al. (2020) considered exclusively the Huber loss function (Huber, 1964), defined by

$$\tilde{l}_\alpha(x) = (2\alpha^{-1}|x| - \alpha^{-2})\, \mathbb{1}\{|x| > \alpha^{-1}\} + x^2\, \mathbb{1}\{|x| \leq \alpha^{-1}\} \tag{2.5}$$

with parameter $\alpha > 0$, and a LASSO penalization. The Huber loss combines a quadratic loss for small values and an absolute loss for large ones. To control the additional approximation error for asymmetric and/or heteroscedastic errors they let the tuning parameter α tend to zero with an appropriate rate depending on the sample size n and the dimension p, and, thus, achieve optimal rates for the estimation error in ℓ_1 and ℓ_2 norm. However, variable selection in this framework is still an open issue.

2.4. Linear regression models with random coefficients

In this section we introduce briefly the linear regression model with random coefficients and random design, which will be studied and discussed in the second part of the thesis intensively. However, we want to emphasize the heteroscedastic error structure of the regression models of the first and second central moments of the coefficients here, which motivates the results in the first part of this thesis. The linear random coefficient regression model can be formalized by

$$Y = B_0 + \mathbf{W}^\top \mathbf{B}\,, \tag{2.6}$$

where $\mathbf{B}, \mathbf{W} \in \mathbb{R}^{p-1}$ are random vectors, B_0 is a random variable and $\mathbf{A} = (B_0, \mathbf{B}^\top)^\top \in \mathbb{R}^p$ and $\mathbf{W}$ are independent. In this context $\mathbf{W} = (W_1, \dots, W_{p-1})^\top$ represents the random regressors and $\mathbf{A} = (A_1, \dots, A_p)^\top$ the random regression coefficients. These are used to model unobserved heterogeneity across the individuals in comparison to the ordinary linear regression model. In this thesis we are mainly interested in the first and second moments of the coefficients $\mathbf{A}$. Hence we assume that they exist and set

$$\mu := \mathbb{E}[\mathbf{A}] \quad \in \mathbb{R}^p$$

and

$$\Sigma := \mathbb{C}\mathrm{ov}(\mathbf{A}) \quad \in \mathbb{R}^{p \times p}\,, \qquad \sigma := \mathrm{vec}(\Sigma) \quad \in \mathbb{R}^{\frac{p(p+1)}{2}}\,, \tag{2.7}$$

where

$$\begin{aligned} \mathrm{vec} : \mathbb{R}^{d\times d} &\to \mathbb{R}^{\frac{d(d+1)}{2}}\,, \\ M &\mapsto \big(M_{11},\dots,M_{dd},M_{12},\dots,M_{1d},M_{23},\dots,M_{2d},\dots,M_{(d-1)d}\big)^{\top} \end{aligned} \tag{2.8}$$

is the half-vectorization of quadratic matrices. Since covariance matrices are symmetric no information about the second central moments of the coefficients is lost by the half-vectorization, more precisely redundant information is deleted. Note that the first p entries of σ are the variances of the coefficients and the remaining entries are the covariances. In a high-dimensional framework sparsity of the first and second central moments means that only a few of the random coefficients have influence and/or heterogeneous effects. Variable selection is used to detect these coefficients.

Let $\mathbf{X} = (1, \mathbf{W}^{\top})^{\top} \in \mathbb{R}^{p}$, then we can rewrite model (2.6) in terms of the means by

$$Y = \mathbf{X}^{\top}\mathbf{A} = \mathbf{X}^{\top}\mu + \mathbf{X}^{\top}\big(\mathbf{A}-\mu\big)\,. \tag{2.9}$$

Evidently, this linear regression model has a heteroscedastic error structure. Analogously we can use the squared residuals, if the means are known, to give a quadratic form in the shape of

$$\big(Y-\mathbf{X}^{\top}\mu\big)^2 = \Big(\mathbf{X}^{\top}\big(\mathbf{A}-\mu\big)\Big)^2 = \mathbf{X}^{\top}\Sigma\,\mathbf{X} + \mathbf{X}^{\top}\Big(\big(\mathbf{A}-\mu\big)\big(\mathbf{A}-\mu\big)^{\top} - \Sigma\Big)\,\mathbf{X}$$

for the variances and covariances of the random coefficients $\mathbf{A}$. With the help of the half-vectorization vec and the corresponding vector transformation we can write the above quadratic form as a heteroscedastic linear regression model as well.

As mentioned in the previous Sections 2.2 and 2.3 variable selection in high-dimensional, heteroscedastic mean regression has not been studied sufficiently in the literature. Hence we want to close this gap with the theory provided in the first part of this thesis.

Part I.

High-dimensional, robust, heteroscedastic linear regression based on the pseudo Huber loss

3. Support estimation with the adaptive LASSO

In this chapter we consider the well-known linear regression model, introduced in Section 2.1, where we allow for heteroscedastic and heavy-tailed, non sub-Gaussian errors. However, we restrict ourselves to light-tailed regressors. Indeed, results in Lederer and Vogt (2020) imply that for uniformly bounded covariates, if the errors have slightly more than a finite fourth moment, the ordinary least squares LASSO estimator retains the rates of convergence in ℓ_1 and ℓ_2 norm known from the sub-Gaussian case. For high-dimensional mean regression under still weaker assumptions, as mentioned in Section 2.3, Fan et al. (2017) and Sun et al. (2020) considered LASSO estimates with the Huber loss function (Huber, 1964), which is given in (2.5). To deal with the resulting bias, they let the tuning parameter α of the Huber loss depend in a suitable way on the sample size n and the dimension p of the coefficient vector. A result from Sun et al. (2020) is that if the errors have a finite second moment and the covariates are sub-Gaussian, then for $\alpha \simeq (\log(p)/n)^{\frac{1}{2}}$ the estimator has the same rates of convergence in ℓ_1 and ℓ_2 norm as in the light-tailed case.

We shall study sign-consistency and rates in ℓ_∞ norm in this framework. Our estimator is based on a smooth and strictly convex variant of the Huber loss function and the adaptive LASSO penalty. In our proofs we combine and extend methods from Zhou et al. (2009), Fan et al. (2017), Loh and Wainwright (2017) and Sun et al. (2020). The results are also provided in Hermann and Holzmann (2020).

This chapter is structured as follows. In Section 3.1 we introduce the exact estimator and Section 3.2 contains the main result in a qualitative form where we focus on the orders and discard exact constants. After reporting on the results of numerical experiments in Section 3.3, we present more precise versions of our results together with the main steps of the proofs in Section 3.4. Section 3.5 concludes, while technical proofs are deferred to Section 3.6.

3.1. Model and estimator

We consider the random design linear regression model

$$Y_i = \mathbf{X_i}^\top \beta^* + \varepsilon_i\,, \qquad i = 1, \dots, n\,, \tag{3.1}$$

in which the real-valued responses Y_i and the p-variate covariates $\mathbf{X_i} \in \mathbb{R}^p$ are observed, and $\beta^* \in \mathbb{R}^p$ is the unknown parameter vector. We allow for a random design with heteroscedastic errors, and assume that $(\mathbf{X_1}^\top, \varepsilon_1)^\top, \dots, (\mathbf{X_n}^\top, \varepsilon_n)^\top$ are independent and identically distributed with $\mathbb{E}\big[\varepsilon_i \,|\, \mathbf{X_i}\big] = 0$, so that $\mathbb{E}\big[Y_i \,|\, \mathbf{X_i}\big] = \mathbf{X_i}^\top \beta^*$ is the identified conditional mean. We shall focus on the high-dimensional case where p is at least of the order n, and consider sub-Gaussian regressors and heavy-tailed errors, where we require only slightly more than second moments.

We use the following variant of the Huber loss function, sometimes called pseudo Huber loss,

$$l_\alpha(x) = 2\alpha^{-2}\Big(\sqrt{1+\alpha^2 x^2} - 1\Big)\,, \tag{3.2}$$

as proposed by Charbonnier et al. (1994). In contrast to the Huber loss in (2.5), l_α is smooth and strictly convex. We consider a computationally feasible estimator based on minimizing the empirical pseudo Huber loss function with a weighted LASSO penalty given by

$$\widehat{\beta}_n^{\mathrm{WLH}} \in \mathop{\arg\min}_{\beta\in\mathbb{R}^p,\, \|\beta\|_2 \le C_\beta} \left(\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta\big) + \lambda_n \sum_{k=1}^p w_k \,|\beta_k|\right) \tag{3.3}$$

with regularization parameter $\lambda_n > 0$, robustification parameter $\alpha_n > 0$ and weights $w_k > 0$ for $k \in \{1,\dots,p\}$, and where the parameter $C_\beta > 0$ (or rather $C_\beta/2$, see Assumption 3.1, (iv)) is some given a-priori bound on the ℓ_2 norm of the true parameter β^*. In (3.3), the empirical loss function $\mathcal{L}^{\mathrm{H}}_{n,\alpha}$ associated with the pseudo Huber loss is defined by

$$\mathcal{L}^{\mathrm{H}}_{n,\alpha}\big(\beta\big) := \frac{1}{n}\sum_{i=1}^n l_\alpha\big(Y_i - \mathbf{X_i}^\top\beta\big)\,, \tag{3.4}$$

and l_α is as in (3.2). We shall call $\widehat{\beta}_n^{\mathrm{WLH}}$ the weighted LASSO Huber estimator (WLHE). It estimates the parameter

$$\beta^*_{\alpha_n} := \mathop{\arg\min}_{\beta\in\mathbb{R}^p,\, \|\beta\|_2 \le C_\beta} \mathbb{E}\Big[l_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top\beta\big)\Big]\,, \tag{3.5}$$

which coincides with β^* in the particular case of a symmetric conditional distribution of ε_1 given $\mathbf{X_1}$, but differs from β^* in general. Later on we assume $\mathbb{E}[\mathbf{X_1}\mathbf{X_1}^\top]$ to be positive definite, hence $\beta^*_{\alpha_n}$ is unique by the strict convexity of l_{α_n}. For a suitable initial estimator $\widehat{\beta}_n^{\mathrm{init}} = (\widehat{\beta}_{n,1}^{\mathrm{init}},\dots,\widehat{\beta}_{n,p}^{\mathrm{init}})^\top \in \mathbb{R}^p$ of β^* such as the LASSO Huber estimator from Fan et al. (2017), choosing the (random) weights

$$w_k = \max\Big\{1/\big|\widehat{\beta}_{n,k}^{\mathrm{init}}\big|, 1\Big\}, \qquad k = 1,\dots,p\,, \tag{3.6}$$

in (3.3) leads to the adaptive LASSO Huber estimator (ALHE) $\widehat{\beta}_n^{\mathrm{ALH}}$ which we shall focus on. Here, if $\big|\widehat{\beta}_{n,k}^{\mathrm{init}}\big| = 0$ so that formally $w_k = \infty$, we require that $\beta_k = 0$ in (3.3).

We shall investigate the sign-consistency as well as the rate of convergence in the ℓ_∞ norm of $\widehat{\beta}_n^{\mathrm{ALH}}$. To this end, let us set up some notation used in the following. Denote the support of the coefficient vector β^* and its regularized version $\beta^*_{\alpha_n}$ in (3.5) by

$$\begin{aligned} S &:= \operatorname{supp}\big(\beta^*\big) = \Big\{k \in \{1, \dots, p\} \mid \beta^*_k \neq 0\Big\}, & s &:= |S|, \\ S_{\alpha_n} &:= \operatorname{supp}\big(\beta^*_{\alpha_n}\big) = \Big\{k \in \{1, \dots, p\} \mid \beta^*_{\alpha_n,k} \neq 0\Big\}, & s_{\alpha_n} &:= \big|S_{\alpha_n}\big|, \end{aligned}$$

where $|S|$ is the cardinality of S. A major additional issue in our investigation will be that the support S of β^*, the object of interest, differs from the support S_{α_n} of $\beta^*_{\alpha_n}$, the parameter which is actually estimated. Indeed, even if β^* is sparse in the sense that S is of small cardinality, this need not be the case for $\beta^*_{\alpha_n}$. However, our analysis will show that the adaptive LASSO penalty reliably sets the small superfluous entries of $\beta^*_{\alpha_n}$ to zero.

Results on support recovery are well-known to depend, in terms of so-called beta-min conditions, on the smallest absolute value of the entries of β^* on its support S, which we denote by

$$\beta^*_{\min} := \min_{k \in S} \big|\beta^*_k\big| \,.$$

3.2. Sign-consistency and rate of convergence in ℓ_∞ norm

In this section we state our main results on sign-consistency and convergence rates in the ℓ_∞ norm of the adaptive LASSO Huber estimator in our setting with heteroscedastic, heavy-tailed and potentially asymmetric errors. Below we give a qualitative version of this result when discarding the constants and focusing on the orders. More precise formulations are provided in Lemma 3.19 combined with Lemmas 3.13 and 3.14 in Section 3.4.

To derive our results we adopt the following assumptions from Fan et al. (2017).

Assumption 3.1.

(i) For $m = 2$ or $m = 3$ and $q > 1$ we have that $\mathbb{E}\big[\mathbb{E}\big[|\varepsilon_1|^m \,\big|\, \mathbf{X_1}\big]^q\big] \leq C_{\epsilon,\mathrm{m}} < \infty$, where $C_{\epsilon,\mathrm{m}} > 0$ is a positive constant.

(ii) For positive constants $0 < c_{\mathbf{X},\mathrm{l}} \leq c_{\mathbf{X},\mathrm{u}}$ we have that $c_{\mathbf{X},\mathrm{l}} \leq \lambda_{\min}\big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]\big) \leq \lambda_{\max}\big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]\big) \leq c_{\mathbf{X},\mathrm{u}} < \infty$, where $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimal and maximal eigenvalues of a symmetric matrix $M \in \mathbb{R}^{d \times d}$.

(iii) For any $v \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$ the variable $v^\top \mathbf{X_1}$ is sub-Gaussian with variance proxy at most $c^2_{\mathbf{X},\mathrm{sub}} \|v\|_2^2$, $c^2_{\mathbf{X},\mathrm{sub}} > 0$, that is $\mathbb{P}\big(|v^\top \mathbf{X_1}| \geq t\big) \leq 2 \exp\big(-t^2/(2\, c^2_{\mathbf{X},\mathrm{sub}} \|v\|_2^2)\big)$ for all $t \geq 0$.

(iv) We have the a-priori upper bound $\|\beta^*\|_2 \leq C_\beta/2$, where $C_\beta \geq 1/8$ is a numerical constant.

The assumptions are essentially those from Fan et al. (2017). In (i), we use a weaker moment assumption, but slightly more than the finite second moment as required in Sun et al. (2020). Further, as in Fan et al. (2017, Section 5) for the loss function from Catoni (2012), in (i) we can only make use of moments up to order 3 when estimating the approximation error. The normalization $C_\beta/2$ is made for later mathematical convenience.

We shall assume that the initial estimator $\widehat{\beta}_n^{\text{init}}$ in the adaptive LASSO achieves the following rates in the ℓ_1 and ℓ_2 norm,

$$\left\|\widehat{\beta}_n^{\text{init}} - \beta^*\right\|_2 \leq C_{\text{init}}\, \lambda_n^{\text{init}} \sqrt{s}\,, \quad \left\|\widehat{\beta}_n^{\text{init}} - \beta^*\right\|_1 \leq C_{\text{init}}\, \lambda_n^{\text{init}}\, s \quad \text{with} \quad \lambda_n^{\text{init}} \simeq \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \tag{3.7}$$

for a positive constant $C_{\text{init}} \geq 1$. The notation suggests that the estimator is based on the regularization parameter λ_n^{init} which then determines its rates. Indeed, under Assumption 3.1 the original LASSO Huber estimator given as a solution of

$$\operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left(\frac{1}{n}\sum_{i=1}^n \tilde{l}_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top \beta\big) + \lambda_n \sum_{k=1}^p |\beta_k|\right)$$

with Huber loss $\tilde{l}_\alpha$ defined in (2.5), satisfies (3.7) for $n \gtrsim s\log(p)$ under the scaling $\alpha_n \simeq (\log(p)/n)^{\frac{1}{2}}$ of the robustification parameter and the choice of the regularization parameter as in (3.7), with probability at least $1-3/p$, see Sun et al. (2020, Theorem 8). From our results in Section 3.4.1 it follows that the same is true when using the pseudo Huber loss function l_α instead.

In the following result, the constants in the order symbols $\simeq$ and $\lesssim$ have to be chosen appropriately to achieve the estimate with the desired probability (3.10), see Lemmas 3.13, 3.14 and 3.19 in Section 3.4 for more details.

Theorem 3.2 (Sign-consistency and rate in the ℓ_∞ norm)**.** *In model* (3.1) *under Assumption 3.1, consider the adaptive LASSO estimator $\widehat{\beta}_n^{\text{ALH}}$ with initial estimator $\widehat{\beta}_n^{\text{init}}$ assumed to satisfy* (3.7)*. Further, suppose that*

$$\left\|\left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1}\right\|_{\text{M},\infty} \leq C_{\text{S},\mathbf{X}}\,, \tag{3.8}$$

where $C_{\text{S},\mathbf{X}} > 0$ is a positive constant, is also satisfied. Assume that the robustification parameter α_n for the adaptive LASSO is chosen of the order

$$\alpha_n \simeq \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}, \tag{3.9}$$

and that the regularization parameter λ_n is chosen of order

$$\lambda_n \simeq \lambda_n^{\text{init}} \left(\frac{|\overline{S}|\log(p)}{n}\right)^{\frac{1}{2}}, \qquad \textit{where} \quad \overline{S} = \left\{k \in \{1,\dots,p\} \,\middle|\, |\widehat{\beta}_{n,k}^{\text{init}}| > \lambda_n^{\text{init}}\right\}$$

and $\lambda_n^{\mathrm{init}} \simeq (\log(p)/n)^{\frac{1}{2}}$ *is as in* (3.7)*. If* $n \gtrsim s^2 \log(p)$ *and* β^* *satisfies a beta-min condition of order* $\beta^*_{\min} \gtrsim s\, \lambda_n^{\mathrm{init}}$*, then with probability at least*

$$1 - c_1 \exp(-c_2 n) - \frac{c_3}{p^2}\,, \tag{3.10}$$

where $c_1, c_2, c_3 > 0$ *are suitable constants, the adaptive LASSO Huber estimator* $\widehat{\beta}_n^{\mathrm{ALH}}$ *as a solution to* (3.3) *with weights* (3.6) *is unique and satisfies*

$$\operatorname{sign}\big(\widehat{\beta}_n^{\mathrm{ALH}}\big) = \operatorname{sign}\big(\beta^*\big) \qquad \text{and} \qquad \Big\|\widehat{\beta}_n^{\mathrm{ALH}} - \beta^*\Big\|_\infty \lesssim \lambda_n^{\mathrm{init}}\,. \tag{3.11}$$

If we drop assumption (3.8) *but instead have* $s \leq \log(p)$*, then we retain the sign-consistency in* (3.11) *but only obtain a* ℓ_∞*-rate of order*

$$\Big\|\widehat{\beta}_n^{\mathrm{ALH}} - \beta^*\Big\|_\infty \lesssim \sqrt{s}\,\lambda_n^{\mathrm{init}}.$$

Remark 3.3. The order in the beta-min condition $\beta^*_{\min} \gtrsim s\,(\log(p)/n)^{\frac{1}{2}}$ as required in our result is the same as in Zhou et al. (2009, equation (4.10)), and quite stronger than the order $\beta^*_{\min} \gtrsim (\log(p)/n)^{\frac{1}{2}}$ required in Loh and Wainwright (2017, Corollary 1, Corollary 3). Potentially, this might be weakened in our setting as well by working with nonconvex regularizers. However, here we preferred to accept this restriction but to have the computationally more efficient adaptive LASSO. The requirement $n \gtrsim s^2 \log(p)$, while being stronger than the $n \gtrsim s \log(p)$ for ordinary least squares in Loh and Wainwright (2017, Corollary 1), is, however, weaker than e.g. the $n \gtrsim s^3 \log(p)$ required in Loh and Wainwright (2017, Corollary 3) for logistic regression. The rate in (3.11) under the additional assumption (3.8) is optimal, while the final bound without this condition is as in Zhou et al. (2009). Somewhat unfortunately, this result requires that $s \leq \log(p)$ and hence is only useful in high dimensions, however, at this stage we were not able to get rid of this assumption. Also, note that the order $\lambda_n \simeq \sqrt{s}\,\log(p)/n$ of the regularization parameter is smaller than the one of the ordinary LASSO. Finally, the bound (3.11) together with the sign-consistency implies that

$$\Big\|\widehat{\beta}_n^{\mathrm{ALH}} - \beta^*\Big\|_2 \lesssim \sqrt{s}\,\lambda_n^{\mathrm{init}} \quad \text{and} \quad \Big\|\widehat{\beta}_n^{\mathrm{ALH}} - \beta^*\Big\|_1 \lesssim s\,\lambda_n^{\mathrm{init}},$$

as for the ordinary LASSO Huber estimator. Our results, in particular Lemmas 3.10 and 3.19 in Section 3.4 imply that this remains true under the weaker set of assumptions in Theorem 3.2, when dropping (3.8).

3.3. Simulations

In this section we numerically compare the performance of the LASSO Huber estimator (LH denotes the LASSO with Huber loss and LPH the LASSO with pseudo Huber loss) and the adaptive LASSO Huber estimator (ALH with Huber loss and ALPH with pseudo Huber loss) with the well-known ordinary LASSO (L) and adaptive LASSO (AL) with quadratic loss function in a simulation setting which is similar to that in Fan et al. (2017). We consider the high-dimensional linear regression model (3.1) with normally distributed covariates $\mathbf{X_1}, \dots, \mathbf{X_n} \sim \mathcal{N}_p\big(\mathbf{0}_p, \mathrm{I}_p\big)$ of dimension $p = 400$ and $n = 200$ observations, and a parameter vector given by

$$\beta^* = \big(3, \dots, 3, 0, \dots, 0\big)^\top$$

with $S = \operatorname{supp}\big(\beta^*\big) = \{1, \dots, 20\}$ and $s = |S| = 20$. In the following we discuss different types of errors (light/heavy tails, symmetric/asymmetric, homo-/heteroscedastic). In the homoscedastic case we assume $\varepsilon_i = \widetilde{\varepsilon}_i$ with $\widetilde{\varepsilon}_1, \dots, \widetilde{\varepsilon}_n$ independent and identically distributed with $\mathbb{E}[\widetilde{\varepsilon}_1] = 0$ and independent of the covariates $\mathbf{X_1}, \dots, \mathbf{X_n}$, while in the heteroscedastic case the errors are

$$\varepsilon_i = \frac{1}{\sqrt{3}\,\|\beta^*\|_2^2}\,\big(\mathbf{X_i}^\top \beta^*\big)^2\,\widetilde{\varepsilon}_i\,.$$

Evidently, $(\mathbf{X_1}^\top, \varepsilon_1)^\top, \dots, (\mathbf{X_n}^\top, \varepsilon_n)^\top$ are independent and identically distributed and $\mathbb{E}[\varepsilon_i \,|\, \mathbf{X_i}] = 0$. Furthermore, the factor $1/\big(\sqrt{3}\,\|\beta^*\|_2^2\big)$ implies

$$\mathbb{E}\big[\varepsilon_1^2\big] = \frac{1}{3\,\|\beta^*\|_2^4}\,\mathbb{E}\Big[\big(\mathbf{X_1}^\top\beta^*\big)^4\Big]\,\mathbb{E}\big[\widetilde{\varepsilon}_1{}^2\big] = \frac{1}{3\,\|\beta^*\|_2^4}\,3\,\|\beta^*\|_2^4\,\mathbb{E}\big[\widetilde{\varepsilon}_1{}^2\big] = \mathbb{E}\big[\widetilde{\varepsilon}_1{}^2\big]$$

since $\mathbf{X_1}^\top\beta^* \sim \mathcal{N}\big(0, \|\beta^*\|_2^2\big)$. Hence the homo- and heteroscedastic errors have the same variance in our simulations.

To compute the estimators in the simulation we use the functions of the packages `glmnet` (ordinary LASSO and adaptive LASSO) and `hqreg` (LASSO with Huber loss and adaptive LASSO with Huber loss). They have a factor of 1/2 in the quadratic loss. Further, the definition of the Huber loss includes an additional scaling of $\alpha/2$ in the package `hqreg`, cf. Yi and Huang (2017). As a consequence, for the Huber loss the regularization parameter λ of the (adaptive) LASSO includes this scaling factor of α as well, therefore we actually displayed λ/α for the Huber loss, which needs to be compared to λ for the ordinary LASSO and the pseudo Huber loss. To compute the estimator for the pseudo Huber loss, we modified the functions of the package `hqreg` which were provided on GitHub by Yi and Huang (2017). This package uses a semismooth Newton coordinate descent algorithm, in contrast to the classical coordinate descent algorithm in `glmnet` or the iterative local adaptive majorize-minimization (I-LAMM) algorithm in Fan et al. (2018).

The parameters α and λ of the estimators are chosen such that the ℓ_2 distance of the respective estimation error is minimal. For this purpose we use 100 independent repetitions where the errors have a specified distribution, and run through a one- or two-dimensional grid for the parameters in each set. In the adaptive versions of the estimators the parameters of the initial estimators are fixed (and equal to the optimal choices for the LASSO), so that we do not require a four-dimensional grid search for the adaptive LASSO. The resulting choices of the robustification parameter α and the regularization parameter λ are displayed in the subsequent tables. Somewhat surprisingly, the tuning parameter for the adaptive version of the estimators differs quite strongly between the ordinary least squares and the estimators based on (pseudo) Huber loss, even for homoscedastic, normally distributed errors.

Next we use these values of the parameters λ and α in a Monte Carlo simulation with 1000 iterations. In addition to the average ℓ_2 and ℓ_∞ norm of the estimation error, we also compute the average percentage of false positives (FP, noise covariates that are selected) and false negatives (FN, signal covariates that are not selected).

The following tables list the results. Overall we have the following main findings. First, for all methods, the version with adaptive weights is superior to that with ordinary weights for both ℓ_2 and ℓ_∞ estimation error, as well as for the proportion of false positives. Second, estimators based on Huber and pseudo Huber loss function perform very similarly. Third, in particular for heteroscedastic errors these estimators have a much better performance than the ordinary LASSO, both in terms of estimation error as well as - in the adaptive versions - for their variable selection properties. Of course, the price to pay is that the additional tuning parameter α has to be chosen.

(a) **Symmetric errors with light tails.**
In this scenario we consider normally distributed errors $\widetilde{\varepsilon}_i \sim \mathcal{N}(0,4)$ with variance equal to 4.

	L	AL	LH	LPH	ALH (LH)	ALPH (LH)	ALPH (LPH)
λ	0.154	0.695	0.157	0.150	0.066	0.067	0.069
α			0.115	0.061	0.153	0.050	0.050
ℓ_2 norm	1.66	0.93	1.67	1.67	0.83	0.83	0.83
ℓ_∞ norm	0.60	0.41	0.61	0.61	0.38	0.38	0.38
FP in %	16.14	1.83	15.76	16.32	1.06	0.99	0.97
FN in %	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3.1.: homoscedastic normally distributed errors.

	L	AL	LH	ALH (LH)	ALPH (LH)
λ	0.150	0.715	0.018	0.0003	0.0003
α			3.476	57.068	55.474
ℓ_2 norm	1.65	0.98	1.12	0.23	0.22
ℓ_∞ norm	0.59	0.41	0.37	0.10	0.09
FP in %	15.81	1.91	21.47	0.96	1.08
FN in %	0.00	0.00	0.00	0.00	0.00

Table 3.2.: heteroscedastic normally distributed errors.

(b) **Symmetric errors with heavy tails.**
Here we consider $\widetilde{\varepsilon}_i = 2\,Q_i$ with $Q_i \sim t_3$ t-distributed with 3 degrees of freedom.

	L	AL	LH	LPH	ALH (LH)	ALPH (LH)	ALPH (LPH)
λ	0.262	0.901	0.142	0.080	0.059	0.040	0.033
α			0.429	0.742	0.563	0.769	0.974
ℓ_2 norm	2.85	1.89	2.34	2.35	1.17	1.18	1.19
ℓ_∞ norm	1.03	0.76	0.85	0.85	0.53	0.53	0.53
FP in %	15.64	2.74	16.66	17.59	1.38	1.39	1.51
FN in %	0.03	0.05	0.00	0.00	0.00	0.00	0.00

Table 3.3.: homoscedastic t-distributed errors.

	L	AL	LH	ALH (LH)	ALPH (LH)
λ	0.226	0.849	0.019	0.0005	0.0006
α			3.574	33.854	29.368
ℓ_2 norm	2.71	1.87	1.37	0.28	0.28
ℓ_∞ norm	0.94	0.72	0.46	0.12	0.11
FP in %	16.36	3.05	20.95	1.13	1.18
FN in %	0.11	0.16	0.00	0.00	0.00

Table 3.4.: heteroscedastic t-distributed errors.

(c) **Asymmetric errors with heavy tails.**
Finally, we consider $\widetilde{\varepsilon}_i = Q_i - \mathbb{E}[Q_i]$ with $Q_i \sim \mathrm{St}(0, 1, 0.6, 3)$ skew t-distributed with location parameter 0, scale parameter 1, skew parameter 0.6 and 3 degrees of freedom. An exact definition can be found in Azzalini and Capitanio (2003) and it is $\mathbb{E}[Q_i] = \big(0.6/\sqrt{1.36}\big)\,\sqrt{3/\pi}\,/\,\Gamma(3/2)$.

	L	AL	LH	LPH	ALH (LH)	ALPH (LH)	ALPH (LPH)
λ	0.118	0.709	0.070	0.058	0.019	0.011	0.010
α			0.863	0.871	1.124	1.842	2.184
ℓ_2 norm	1.33	0.74	1.08	1.12	0.47	0.46	0.47
ℓ_∞ norm	0.48	0.32	0.39	0.40	0.22	0.22	0.23
FP in %	16.37	1.56	16.48	16.49	0.52	0.63	0.53
FN in %	0.01	0.01	0.00	0.00	0.02	0.00	0.01

Table 3.5.: homoscedastic skew t-distributed errors.

	L	AL	LH	ALH (LH)	ALPH (LH)
λ	0.110	0.649	0.009	0.0003	0.0002
α			7.00	33.898	50.684
ℓ_2 norm	1.28	0.77	0.64	0.11	0.11
ℓ_∞ norm	0.45	0.32	0.22	0.05	0.05
FP in %	16.00	1.80	21.18	0.43	0.48
FN in %	0.02	0.02	0.00	0.00	0.00

Table 3.6.: heteroscedastic skew t-distributed errors.

3.4. Main steps of the proof and auxiliary results

In this section we present the results in more technical form together with the main steps of the proofs. Various technical details are deferred to Section 3.6. Let us give an overview of our approach. In Section 3.4.1 we start with various technical preparations, including a bound on the approximation bias and the restricted strong convexity condition for the pseudo Huber loss, similar to Fan et al. (2017, Section 5) for the Catoni loss function. Section 3.4.2 details how to implement the primal-dual witness approach from Wainwright (2009b) in our setting. Compared to Loh and Wainwright (2017) and Zhou et al. (2009), the main additional issue is that $\beta^*_{\alpha_n}$ as defined in (3.5) does not have support S and, indeed, need not to be sparse. Lemmas 3.13 and 3.14 take care of technical expressions, in particular the inverse of the Hessian of the empirical loss function restricted to S, and of a term involving the gradient when checking strict dual feasibility. In Section 3.4.3, we deduce a result for the general weighted LASSO Huber estimator in (3.3), which still involves a mutual incoherence condition. Finally, in Sections 3.4.4 and 3.4.5 this is specialized for the adaptive LASSO, first for an initial estimator satisfying general rate assumptions, and then for one which is assumed to satisfy (3.7), for which we can get rid of the mutual incoherence condition.

We shall use the following additional notation. $\mathbb{X}_n = \big(\mathbf{X_1}, \dots, \mathbf{X_n}\big)^\top \in \mathbb{R}^{n \times p}$ is the design matrix where $\mathbf{X_i} \in \mathbb{R}^p$ is the covariate vector in model (3.1). $w = (w_1, \dots, w_p)^\top$ denotes the vector of weights from (3.3), and we set $w_{\max}\big(S\big) = \max_{k \in S} w_k$ and $w_{\min}\big(S^c\big) = \min_{k \in S^c} w_k$.

3.4.1. Technical preparations

We start with some technical preparations where we extend results from Fan et al. (2017) to the pseudo Huber loss function l_α given in (3.2). See also Fan et al. (2017, Section 5) for similar extensions to the Cantoni loss function (Catoni, 2012). The proofs of the lemmas in this section are provided in Section 3.6.1. To start, straightforward differentiation gives

$$l'_\alpha(x) = \frac{2x}{\sqrt{1+\alpha^2 x^2}} \qquad \text{so that} \qquad \big|l'_\alpha(x)\big| \le \frac{2|x|}{\sqrt{\alpha^2 x^2}} = 2\alpha^{-1}, \tag{3.12}$$

and

$$l''_\alpha(x) = \frac{2\alpha^{-3}}{(\alpha^{-2}+x^2)^{3/2}} \qquad \text{so that} \qquad 0 < l''_\alpha(x) \le \frac{2\alpha^{-3}}{(\alpha^{-2})^{3/2}} = 2\,. \tag{3.13}$$

In particular, l_α is strictly convex. Also note that $\lim_{\alpha\to 0} l_\alpha(x) = x^2$ for all $x \in \mathbb{R}$. For the empirical loss function in (3.4) this gives

$$\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta) = -\frac{1}{n}\sum_{i=1}^{n} l'_\alpha\big(Y_i - \mathbf{X_i}^\top\beta\big)\mathbf{X_i}\,, \qquad \nabla^2 \mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta) = \frac{1}{n}\sum_{i=1}^{n} l''_\alpha\big(Y_i - \mathbf{X_i}^\top\beta\big)\mathbf{X_i}\mathbf{X_i}^\top\,. \tag{3.14}$$

The following result is similar to Fan et al. (2017, Theorem 1 and Theorem 6), however, we work with a weaker moment assumption.

Lemma 3.4 (ℓ_2 norm bound on the approximation error)**.** *Under Assumption 3.1 we have for $\beta^*_{\alpha_n}$ in* (3.5) *that*

$$\big\|\beta^*_{\alpha_n} - \beta^*\big\|_2 \le C_{\mathrm{apx}}\,\alpha_n^{m-1}\,, \tag{3.15}$$

where

$$C_{\mathrm{apx}} = \frac{5\,2^m\,c_{\mathbf{X},\mathrm{sub}}}{c_{\mathbf{X},\mathrm{l}}}\left[\left(\frac{q}{q-1}\,\Gamma\Big(\frac{q}{2(q-1)}\Big)\right)^{\frac{q-1}{q}} (C_{\epsilon,\mathrm{m}})^{\frac{1}{q}} + \Big(2\,\big(2C_\beta^2\,c^2_{\mathbf{X},\mathrm{sub}}\big)^m\,(2m)!\,\Gamma(m)\Big)^{\frac{1}{2}}\right]$$

and $\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)\,dt$*,* $x>0$*, is the gamma function.*

Remark 3.5. The above result leads to $\big\|\beta^*_{\alpha_n} - \beta^*\big\|_2 < C_\beta/2$ for an (appropriate) choice of α_n. Together with the assumption $\|\beta^*\|_2 \le C_\beta/2$ this will imply that $\beta^*_{\alpha_n}$ is strictly feasible for (3.5), that is,

$$\big\|\beta^*_{\alpha_n}\big\|_2 < C_\beta, \tag{3.16}$$

which we will assume from now on.

Next we show along the lines of Fan et al. (2017, Lemmas 2 and 4) that restricted strong convexity, cf. Loh and Wainwright (2017), is satisfied by the pseudo Huber loss function.

Lemma 3.6 (RSC condition)**.** *Under Assumption 3.1 there exist $c_\alpha > 0$ (depending on $c_{\mathbf{X},\mathrm{l}}$, $c_{\mathbf{X},\mathrm{u}}$, $c_{\mathbf{X},\mathrm{sub}}$ and C_β) and $c^{\mathrm{P}}_1, c^{\mathrm{P}}_2 > 0$ (depending on $c_{\mathbf{X},\mathrm{l}}$ and $c_{\mathbf{X},\mathrm{sub}}$) such that for all $\|\beta\|_2 \le 4C_\beta$, $\|\Delta\|_2 \le 8C_\beta$ and $\alpha \le c_\alpha$ with probability at least $1 - c^{\mathrm{P}}_1\exp(-c^{\mathrm{P}}_2 n)$ the empirical pseudo Huber loss function $\mathcal{L}^{\mathrm{H}}_{n,\alpha}$ satisfies the restricted strong convexity condition*

$$\big\langle \nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta+\Delta) - \nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta), \Delta\big\rangle \ge c^{\mathrm{RSC}}_1\,\|\Delta\|_2^2 - c^{\mathrm{RSC}}_2\,\frac{\log(p)}{n}\|\Delta\|_1^2 \tag{3.17}$$

with

$$c_1^{\mathrm{RSC}} = \frac{c_{\mathbf{X},\mathrm{l}}}{16}\,, \qquad c_2^{\mathrm{RSC}} = \frac{1600\, c_{\mathbf{X},\mathrm{sub}}^2 \Big(\max\big(4c_{\mathbf{X},\mathrm{sub}}\sqrt{\log(12c_{\mathbf{X},\mathrm{sub}}^2/c_{\mathbf{X},\mathrm{l}})}, 1\big)\Big)^4}{c_{\mathbf{X},\mathrm{l}}}\,.$$

Restricted strong convexity implies in particular ordinary strong convexity locally on the support S of β^*.

Lemma 3.7. *Under Assumption 3.1, if $\alpha \leq c_\alpha$ and $n \geq c_3^{\mathrm{RSC}} s\log(p)$ with $c_3^{\mathrm{RSC}} = 2c_2^{\mathrm{RSC}}/c_1^{\mathrm{RSC}}$ we have with probability at least $1 - c_1^{\mathrm{P}}\exp(-c_2^{\mathrm{P}} n)$ for $\beta \in \mathbb{R}^p$ with $\|\beta\|_2 \leq 4C_\beta$ that*

$$\lambda_{\min}\Big(\big(\nabla^2\mathcal{L}_{n,\alpha}^{\mathrm{H}}(\beta)\big)_{SS}\Big) \geq \frac{c_1^{\mathrm{RSC}}}{2} = \frac{c_{\mathbf{X},\mathrm{l}}}{32}\,, \tag{3.18}$$

where $\lambda_{\min}(M)$ denotes the minimal eigenvalue of a symmetric matrix $M \in \mathbb{R}^{d\times d}$.

The following result gives a bound on the gradient of the empirical loss function, and is analogous to Fan et al. (2017, Lemma 1).

Lemma 3.8 (ℓ_∞ norm bound on the gradient)**.** *Under Assumption 3.1 there exist $c_1^{\mathrm{Grad}}, c_2^{\mathrm{Grad}} > 0$ (depending on q, $C_{\epsilon,\mathrm{m}}$, $c_{\mathbf{X},\mathrm{sub}}$ and C_β) such that for all $\alpha_n \geq c_1^{\mathrm{Grad}}(\log(p)/n)^{\frac{1}{2}}$ with probability at least $1 - 2/p^2$ the ℓ_∞ norm of the gradient of the empirical pseudo Huber loss function $\mathcal{L}_{n,\alpha_n}^{\mathrm{H}}$ at $\beta_{\alpha_n}^*$ is bounded by*

$$\big\|\nabla\mathcal{L}_{n,\alpha_n}^{\mathrm{H}}(\beta_{\alpha_n}^*)\big\|_\infty \leq c_2^{\mathrm{Grad}}\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}.$$

3.4.2. Primal-dual witness approach

The proof of Theorem 3.2 is based on the primal-dual witness (PDW) approach as originally introduced in Wainwright (2009b). Following Loh and Wainwright (2017) we summarize the three main steps as follows. The key results in this section for implementing this approach in our setting are Lemma 3.12, together with the Lemmas 3.13 and 3.14.

(i) Optimize the restricted program

$$\widehat{\beta}_n^{\mathrm{PDW}} \in \mathop{\arg\min}_{\beta\in\mathbb{R}^p,\mathrm{supp}(\beta)\subseteq S,\|\beta\|_2\leq C_\beta} \left(\mathcal{L}_{n,\alpha_n}^{\mathrm{H}}(\beta) + \lambda_n \sum_{k\in S} w_k\,|\beta_k|\right), \tag{3.19}$$

where we enforce the constraint that $\mathrm{supp}\big(\widehat{\beta}_n^{\mathrm{PDW}}\big) \subseteq S$, and show that all solutions have norm $< C_\beta$.

(ii) Choose $\widehat{\gamma} = \widehat{\gamma}_n \in \mathbb{R}^p$ such that (a.) $\widehat{\gamma}_S \in \partial\big\|\widehat{\beta}_{n,S}^{\mathrm{PDW}}\big\|_1$, (b.) it satisfies the zero-subgradient condition

$$\nabla\mathcal{L}_{n,\alpha_n}^{\mathrm{H}}\big(\widehat{\beta}_n^{\mathrm{PDW}}\big) + \lambda_n\big(w \odot \widehat{\gamma}\big) = \mathbf{0}_p\,, \tag{3.20}$$

and (c.) such that $\widehat{\gamma}_{S^c}$ satisfies the strict dual feasibility condition $\|\widehat{\gamma}_{S^c}\|_\infty < 1$.

(iii) Show that $\widehat{\beta}_n^{\mathrm{PDW}}$ is also a minimum of the full program (3.3),

$$\underset{\beta\in\mathbb{R}^p,\,\|\beta\|_2\le C_\beta}{\arg\min}\left(\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta)+\lambda_n\sum_{k=1}^{p}w_k\,|\beta_k|\right),$$

and, moreover, the uniqueness of the minimizer of this program.

We shall always assume in the following that

$$\alpha_n \le c_\alpha$$

holds, where c_α is given in Lemma 3.6. Later, α_n is chosen of an order tending to zero, so that this is automatically satisfied. The following lemma lists some technical properties of the second derivatives of the empirical loss function.

Lemma 3.9. *We may write*

$$\widehat{Q} := \int_0^1 \nabla^2\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\Big(\beta^*_{\alpha_n}+t\,\big(\widehat{\beta}_n^{\mathrm{PDW}}-\beta^*_{\alpha_n}\big)\Big)dt \;=\frac{2}{n}\sum_{i=1}^{n}d_i\,\mathbf{X_i}\,\mathbf{X_i}^\top=\frac{2}{n}\,\mathbb{X}_n^\top\,D\,\mathbb{X}_n\,, \quad (3.21)$$

where $D=\operatorname{diag}\big(d_1,\ldots,d_n\big)$ *with*

$$d_i=\frac{1}{2}\int_0^1 l''_{\alpha_n}\Big(Y_i-\mathbf{X_i}^\top\big(\beta^*_{\alpha_n}+t\,\big(\widehat{\beta}_n^{\mathrm{PDW}}-\beta^*_{\alpha_n}\big)\big)\Big)dt \qquad \in(0,1]\,.$$

Furthermore, under Assumption 3.1, if $n\ge c_3^{\mathrm{RSC}}s\log(p)$ *with probability at least* $1-c_1^{\mathrm{P}}\exp(-c_2^{\mathrm{P}}n)$ *the submatrix* $\widehat{Q}_{SS}$ *is invertible with minimal eigenvalue bounded below by* $c_{\mathbf{X},\mathrm{l}}/32$ *and we have the bound*

$$\Big\|\big(\widehat{Q}_{SS}\big)^{-1}\Big\|_{\mathrm{M},\infty}\le\frac{32\sqrt{s}}{c_{\mathbf{X},\mathrm{l}}}\,. \quad (3.22)$$

Proof of Lemma 3.9. (3.21) follows from straightforward calculation, see (3.14). Moreover, every point $\beta\in\mathbb{R}^p$ between $\beta^*_{\alpha_n}$ and $\widehat{\beta}_n^{\mathrm{PDW}}$ has ℓ_2 norm smaller than or equal to C_β because $\|\beta^*_{\alpha_n}\|_2, \|\widehat{\beta}_n^{\mathrm{PDW}}\|_2\le C_\beta$. Hence (3.18) implies the invertibility of $\widehat{Q}_{SS}$ and (3.22) follows from

$$\Big\|\big(\widehat{Q}_{SS}\big)^{-1}\Big\|_{\mathrm{M},\infty}\le\sqrt{s}\,\Big\|\big(\widehat{Q}_{SS}\big)^{-1}\Big\|_{\mathrm{M},2}\le\frac{32\sqrt{s}}{c_{\mathbf{X},\mathrm{l}}}\,.$$

□

In the following lemma we show that $\widehat{\beta}_n^{\mathrm{PDW}}$ is strictly feasible for (3.19), meaning $\|\widehat{\beta}_n^{\mathrm{PDW}}\|_2<C_\beta$ holds, for an appropriate choice of λ_n and α_n.

Lemma 3.10 (ℓ_2 norm error bound on the PDW estimator)**.** *Under Assumption 3.1 we have for* $\widehat{\beta}_n^{\mathrm{PDW}}$ *in* (3.19) *with* $\alpha_n \geq c_1^{\mathrm{Grad}} \big(\frac{\log(p)}{n}\big)^{\frac{1}{2}}$ *that*

$$\Big\|\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*\Big\|_2 \leq \left(c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} + w_{\max}\big(S\big)\,\lambda_n + 2C_\beta\, c_2^{\mathrm{RSC}} \frac{\sqrt{s}\log(p)}{n} \right) \frac{\sqrt{s}}{c_1^{\mathrm{RSC}}} \\ + C_{\mathrm{apx}}\, \alpha_n^{m-1}$$

with probability at least $1 - c_1^{\mathrm{P}} \exp(-c_2^{\mathrm{P}} n) - 2/p^2$.

Proof of Lemma 3.10. Let

$$\beta^*_{\alpha_n,\mathrm{supp}} = \operatorname*{arg\,min}_{\substack{\beta\in\mathbb{R}^p,\mathrm{supp}(\beta)\subseteq S,\\ \|\beta\|_2\leq C_\beta}} \mathbb{E}\Big[l_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top\beta\big)\Big] \quad \text{and} \quad \Delta_n^{\mathrm{PDW}} = \widehat{\beta}_n^{\mathrm{PDW}} - \beta^*_{\alpha_n,\mathrm{supp}}\,, \tag{3.23}$$

then $\widehat{\beta}_n^{\mathrm{PDW}}$ in (3.19) is a regularized M-estimator of $\beta^*_{\alpha_n,\mathrm{supp}}$. Following the proof of Lemma 3.4 leads on the one hand to

$$\big\|\beta^*_{\alpha_n,\mathrm{supp}} - \beta^*\big\|_2 \leq C_{\mathrm{apx}}\, \alpha_n^{m-1}\,.$$

In doing so note that

$$\mathbb{E}\Big[l_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top\beta^*_{\alpha_n,\mathrm{supp}}\big)\Big] \leq \mathbb{E}\Big[l_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top\beta^*\big)\Big] \quad \text{and} \quad \big\|\beta^*_{\alpha_n,\mathrm{supp}}\big\|_2 \leq C_\beta$$

because of (3.23), $\mathrm{supp}\big(\beta^*\big) = S$ and $\|\beta^*\|_2 \leq C_\beta$ by (iv) of Assumption 3.1. Further, $\widehat{\beta}_n^{\mathrm{PDW}}$ has to satisfy the first-order necessary condition of a convex constrained optimization problem over a convex set to be a minimum of (3.19), cf. Ruszczynski (2006, Theorem 3.33), that is, there exists $\widehat{\gamma} \in \partial\big\|\widehat{\beta}_{n,S}^{\mathrm{PDW}}\big\|_1$ such that

$$\Big\langle \nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\widehat{\beta}_n^{\mathrm{PDW}}\big) + \lambda_n\big(w \odot \widehat{\gamma}\big), \beta - \widehat{\beta}_n^{\mathrm{PDW}} \Big\rangle \geq 0 \quad \text{for all feasible } \beta \in \mathbb{R}^p\,.$$

Hence by the restricted strong convexity of the empirical pseudo Huber loss function in Lemma 3.6 and the first-order necessary condition it follows that

$$\begin{aligned} c_1^{\mathrm{RSC}} \big\|\Delta_n^{\mathrm{PDW}}\big\|_2^2 - c_2^{\mathrm{RSC}} \frac{\log(p)}{n} \big\|\Delta_n^{\mathrm{PDW}}\big\|_1^2 &\leq \Big\langle \nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\widehat{\beta}_n^{\mathrm{PDW}}) - \nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n,\mathrm{supp}}), \Delta_n^{\mathrm{PDW}} \Big\rangle \\ &\leq \big\langle -\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n,\mathrm{supp}}) - \lambda_n\big(w \odot \widehat{\gamma}\big), \Delta_n^{\mathrm{PDW}} \big\rangle \\ &\leq \big\|\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n,\mathrm{supp}})\big\|_\infty \big\|\Delta_n^{\mathrm{PDW}}\big\|_1 \\ &\qquad + w_{\max}\big(S\big)\,\lambda_n \big\|\Delta_n^{\mathrm{PDW}}\big\|_1 \end{aligned}$$

with probability at least $1 - c_1^{\mathrm{P}} \exp(-c_2^{\mathrm{P}} n)$. Here the last inequality follows since $\beta^*_{\alpha_n,\mathrm{supp}}$ and $\widehat{\beta}_n^{\mathrm{PDW}}$ both have support (contained in) S. Rearranging leads to

$$c_1^{\mathrm{RSC}} \big\|\Delta_n^{\mathrm{PDW}}\big\|_2^2 \leq \Big(\big\|\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n,\mathrm{supp}})\big\|_\infty + w_{\max}\big(S\big)\,\lambda_n \\ + c_2^{\mathrm{RSC}} \frac{\log(p)}{n} \big\|\Delta_n^{\mathrm{PDW}}\big\|_1 \Big) \big\|\Delta_n^{\mathrm{PDW}}\big\|_1\,.$$

We obtain $\left\|\Delta_n^{\mathrm{PDW}}\right\|_1 \leq \sqrt{s}\left\|\Delta_n^{\mathrm{PDW}}\right\|_2$ and $\left\|\Delta_n^{\mathrm{PDW}}\right\|_2 \leq 2C_\beta$ because of (3.19) and (3.23). In addition, by following the proof of Lemma 3.8 we get $\left\|\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n,\mathrm{supp}})\right\|_\infty \leq c_2^{\mathrm{Grad}}\,(\log(p)/n)^{\frac{1}{2}}$ with probability at least $1 - p^2/2$. Hence it follows that

$$\left\|\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*_{\alpha_n,\mathrm{supp}}\right\|_2 \leq \left(c_2^{\mathrm{Grad}}\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} + w_{\max}(S)\,\lambda_n + 2C_\beta\, c_2^{\mathrm{RSC}}\frac{\sqrt{s}\log(p)}{n}\right)\frac{\sqrt{s}}{c_1^{\mathrm{RSC}}}$$

and in total the assertion of the lemma. □

Remark 3.11. The results below will imply that with an (appropriate) choice of λ_n and α_n,

$$\left\|\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*\right\|_2 = \mathcal{O}\left(\left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}}\right)$$

holds with high probability, so that in particular $\left\|\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*\right\|_2 < C_\beta/2$. Together with the assumption $\|\beta^*\|_2 \leq C_\beta/2$ this will imply that $\widehat{\beta}_n^{\mathrm{PDW}}$ is strictly feasible for (3.19), that is,

$$\left\|\widehat{\beta}_n^{\mathrm{PDW}}\right\|_2 < C_\beta\,, \tag{3.24}$$

which we will assume from now on.

Lemma 3.12 (Solving the PDW construction)**.** *Suppose that Assumption 3.1 holds and that β^* satisfies the beta-min condition*

$$\beta^*_{\min} > C_{\mathrm{apx}}\,\alpha_n^{m-1}\,, \tag{3.25}$$

and that $n \geq c_3^{\mathrm{RSC}} s\log(p)$. Let $\widehat{\beta}_n^{\mathrm{PDW}}$ be as in the PDW construction, and suppose that $\widehat{\gamma} \in \mathbb{R}^p$ satisfies $\widehat{\gamma}_S \in \partial\left\|\widehat{\beta}_{n,S}^{\mathrm{PDW}}\right\|_1$ and the zero-subgradient condition (3.20)*. Then, with probability at least $1 - c_1^{\mathrm{P}}\exp(-c_2^{\mathrm{P}} n)$ the strict dual feasibility condition $\|\widehat{\gamma}_{S^c}\|_\infty < 1$ is equivalent to the condition*

$$\begin{aligned}\Bigg|\widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\Big(\lambda_n\big(w_S \odot \widehat{\gamma}_S\big) + \Big(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_S\Big) - \Big(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_{S^c} \\ + \Big(\widehat{Q}_{S^c(S_{\alpha_n}\setminus S)} - \widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\Big)\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\Bigg| < w_{S^c}\,\lambda_n\,.\end{aligned} \tag{3.26}$$

Furthermore, if (3.26) *is satisfied we have that the minimizer $\widehat{\beta}_n^{\mathrm{WLH}}$ in* (3.3) *is unique and given by $\widehat{\beta}_n^{\mathrm{WLH}} = \widehat{\beta}_n^{\mathrm{PDW}}$, so that in particular $\mathrm{supp}\big(\widehat{\beta}_n^{\mathrm{WLH}}\big) \subseteq \mathrm{supp}\big(\beta^*\big)$, and, furthermore, that*

$$\left\|\widehat{\beta}_n^{\mathrm{WLH}} - \beta^*\right\|_\infty \leq \phi_{n,\infty}\,,$$

where

$$\begin{aligned}\phi_{n,\infty} = \left\| \big(\widehat{Q}_{SS}\big)^{-1} \right\|_{\mathrm{M},\infty} \left\| \Big(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_S \right\|_\infty + w_{\max}(S)\, \lambda_n \left\| \big(\widehat{Q}_{SS}\big)^{-1} \right\|_{\mathrm{M},\infty} \\ + \left\| \beta^*_{\alpha_n,S} - \beta^*_S \right\|_\infty + \left\| \big(\widehat{Q}_{SS}\big)^{-1} \right\|_{\mathrm{M},\infty} \left\| \big(\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\, \beta^*_{\alpha_n, S_{\alpha_n}\setminus S} \right\|_\infty . \end{aligned} \tag{3.27}$$

Furthermore, if we have in addition the beta-min condition of the same order

$$\beta^*_{\min} > \phi_{n,\infty}\,, \tag{3.28}$$

then we have the sign-recovery property $\operatorname{sign}\big(\widehat{\beta}^{\mathrm{WLH}}_n\big) = \operatorname{sign}\big(\beta^*\big)$.

The beta-min condition (3.25) is required so that the approximation error $\left\|\beta^*_{\alpha_n} - \beta^*\right\|_\infty$ is smaller than $\beta^*_{\min}$, which implies that the support S_{α_n} of $\beta^*_{\alpha_n}$ contains the support S of β^*. Later, we shall choose α_n and achieve a rate $\phi_{n,\infty}$, which in any case also includes an approximation term $\left\|\beta^*_{\alpha_n,S} - \beta^*_S\right\|_\infty$, such that (3.28) implies (3.25).

Proof of Lemma 3.12. We start by showing that under assumption (3.25) we have $S \subseteq S_{\alpha_n}$. To this end, we estimate

$$\begin{aligned}\left|\beta^*_{\alpha_n,k}\right| = \left|\beta^*_{\alpha_n,k} - \beta^*_k + \beta^*_k\right| \geq \left|\beta^*_k\right| - \left|\beta^*_{\alpha_n,k} - \beta^*_k\right| \geq \beta^*_{\min} - \left\|\beta^*_{\alpha_n} - \beta^*\right\|_\infty \\ \geq \beta^*_{\min} - \left\|\beta^*_{\alpha_n} - \beta^*\right\|_2 \\ \geq \beta^*_{\min} - C_{\mathrm{apx}}\, \alpha_n^{m-1} > 0\end{aligned}$$

for $k \in S$, where the first inequality in the last line follows from (3.15) and the final inequality from (3.25). Now, using

$$\widehat{Q}\,\big(\widehat{\beta}^{\mathrm{PDW}}_n - \beta^*_{\alpha_n}\big) = \nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\widehat{\beta}^{\mathrm{PDW}}_n\big) - \nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\,,$$

see (3.21) for the definition of $\widehat{Q}$, we may rewrite the subgradient condition (3.20), which holds since $\widehat{\beta}^{\mathrm{PDW}}_n$ is strictly feasible as in (3.24), as

$$\widehat{Q}\,\big(\widehat{\beta}^{\mathrm{PDW}}_n - \beta^*_{\alpha_n}\big) + \nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big) + \lambda_n\big(w \odot \widehat{\gamma}\big) = \mathbf{0}_p$$

or in block-form

$$\begin{aligned}\begin{bmatrix} \widehat{Q}_{SS} & \widehat{Q}_{S(S_{\alpha_n}\setminus S)} & \widehat{Q}_{SS^c_{\alpha_n}} \\ \widehat{Q}_{S^cS} & \widehat{Q}_{S^c(S_{\alpha_n}\setminus S)} & \widehat{Q}_{S^cS^c_{\alpha_n}} \end{bmatrix} \begin{pmatrix} \widehat{\beta}^{\mathrm{PDW}}_{n,S} - \beta^*_{\alpha_n,S} \\ -\beta^*_{\alpha_n,S_{\alpha_n}\setminus S} \\ \mathbf{0}_{|S^c_{\alpha_n}|} \end{pmatrix} + \begin{pmatrix} \Big(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_S \\ \Big(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_{S^c} \end{pmatrix} \\ + \lambda_n \begin{pmatrix} w_S \odot \widehat{\gamma}_S \\ w_{S^c} \odot \widehat{\gamma}_{S^c} \end{pmatrix} = \mathbf{0}_p\,,\end{aligned}$$

where we used that $\widehat{\beta}^{\mathrm{PDW}}_{n,S^c} = \mathbf{0}_{p-s}$ by the primal-dual witness construction. By invertibility of $\widehat{Q}_{SS}$, see Lemma 3.9, this leads to

$$\begin{aligned}\widehat{\beta}^{\mathrm{PDW}}_{n,S} - \beta^*_{\alpha_n,S} = \big(\widehat{Q}_{SS}\big)^{-1}\bigg(- \lambda_n\big(w_S \odot \widehat{\gamma}_S\big) - \Big(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_S \\ + \widehat{Q}_{S(S_{\alpha_n}\setminus S)}\, \beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\bigg)\end{aligned} \tag{3.29}$$

and

$$\lambda_n\big(w_{S^c}\odot\widehat{\gamma}_{S^c}\big)=\widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\Big(\lambda_n\big(w_S\odot\widehat{\gamma}_S\big)+\Big(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_S\Big)-\Big(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_{S^c}$$
$$+\Big(\widehat{Q}_{S^c(S_{\alpha_n}\setminus S)}-\widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\Big)\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\,.$$

The second equation shows the equivalence of the strict dual feasibility condition $\|\widehat{\gamma}_{S^c}\|_\infty<1$ and (3.26). Now, if this holds then we obtain that $\widehat{\gamma}\in\partial\big\|\widehat{\beta}^{\mathrm{PDW}}_n\big\|_1$, and since the loss function $\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}$ is convex (and obviously also the weighted ℓ_1 norm), we obtain by (3.20) that $\widehat{\beta}^{\mathrm{PDW}}_n$ is also a solution of (3.3), cf. Ruszczynski (2006, Theorem 3.33) and recall from (3.24) that $\widehat{\beta}^{\mathrm{PDW}}_n$ is (assumed to be) strictly feasible. To conclude $\widehat{\beta}^{\mathrm{WLH}}_n=\widehat{\beta}^{\mathrm{PDW}}_n$ we need to show that this solution is unique. Then apparently $\mathrm{supp}\big(\widehat{\beta}^{\mathrm{WLH}}_n\big)\subseteq\mathrm{supp}\big(\beta^*\big)$ and (3.27) follows from (3.29). If the beta-min condition (3.28) holds, then for $k\in S$

$$\Big|\widehat{\beta}^{\mathrm{WLH}}_{n,k}-\beta^*_k\Big|\le\Big\|\widehat{\beta}^{\mathrm{WLH}}_n-\beta^*\Big\|_\infty<\beta^*_{\min}\le\big|\beta^*_k\big|\,,$$

which implies $\mathrm{sign}\big(\widehat{\beta}^{\mathrm{WLH}}_{n,k}\big)=\mathrm{sign}\big(\beta^*_k\big)$ and hence the sign-consistency of $\widehat{\beta}^{\mathrm{WLH}}_n$. It remains to show uniqueness of the solution of the program (3.3). To this end, we show that all stationary points $\widetilde{\beta}$, that is points satisfying $\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\widetilde{\beta}\big)=-\lambda_n\big(w\odot\widetilde{\gamma}\big)$ with $\widetilde{\gamma}\in\partial\|\widetilde{\beta}\|_1$, have support S, cf. Loh and Wainwright (2017, Lemma 3) or Tibshirani (2013, Section 2.3). Then strict convexity of the loss function restricted to vectors with support S, as implied by (3.18), concludes the proof.

From the form (3.14) of the gradient of the loss function we see that uniqueness of the fitted values $\mathbb{X}_n\,\widetilde{\beta}$ for all stationary points implies uniqueness of the subgradient $\widetilde{\gamma}$, that is $\widetilde{\gamma}=\widehat{\gamma}$. The strict dual feasibility condition $\|\widehat{\gamma}_{S^c}\|_\infty<1$ for $\widehat{\gamma}$ then implies that $\widetilde{\beta}$ must also have support in S, cf. Tibshirani (2013, Section 2.3) or Wainwright (2009b, Lemma 1 (b)). Now, uniqueness of the fitted values follows from the strict convexity of the pseudo Huber loss by using Lemma 1 (ii) in Tibshirani (2013). This concludes the proof of the lemma. □

In the next two technical results we show how to take care of the terms involving the gradient of the loss in the strict dual feasibility assumption (3.26), and how to obtain a sharper bound on the inverse of $\widehat{Q}_{SS}$ then (3.22) under (3.8).

Lemma 3.13 (Strict dual feasibility and norm bound I)**.** *Suppose that Assumption 3.1 and* (3.8) *are satisfied and assume that the robustification parameter α_n is chosen as in* (3.9)*. If $n\ge C_3\,s^2\log(p)$ for a sufficiently large positive constant $C_3>0$, then there exist constants $C_1,C_2,C_{\mathrm{Q,S}}>0$ and $C_{\mathrm{Q},\mathcal{L}}\ge1$ such that*

$$\Big\|\big(\widehat{Q}_{SS}\big)^{-1}\Big\|_{\mathrm{M},\infty}\le C_{\mathrm{Q,S}}\tag{3.30}$$

is satisfied with probability at least $1 - C_1/p^2 - 6/p^{5s}$*, and*

$$\left\| \widehat{Q}_{S^cS} \big(\widehat{Q}_{SS}\big)^{-1} \Big(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_S - \Big(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big)\Big)_{S^c} \right\|_\infty \le C_{\mathrm{Q},\mathcal{L}}\, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \tag{3.31}$$

with probability at least $1 - (4 + C_1 + C_2)/p^2 - 6/p^{5s}$*, where* c_2^{Grad} *is as in Lemma 3.8.*

The proof is deferred to Section 3.6.2. If we drop the requirement (3.8) we still obtain a bound of the form (3.31) under the somewhat restrictive scaling $s \le \log(p)$. The bound (3.30) is no longer valid and needs to be replaced by (3.22).

Lemma 3.14 (Strict dual feasibility and norm bound II)**.** *Suppose that Assumption 3.1 holds and assume that the robustification parameter satisfies* $\alpha_n \ge \sqrt{4/3}\, c_1^{\mathrm{Grad}} (\log(p)/n)^{\frac{1}{2}}$*, where* c_1^{Grad} *is as in Lemma 3.8. Then for* $s \le \log(p)$ *and* $n \ge \max\left\{c_3^{\mathrm{RSC}} s \log(p), 6\log(p)\right\}$ *we still have* (3.31) *with probability at least* $1 - c_1^{\mathrm{P}} \exp(-c_2^{\mathrm{P}} n) - 6/p^2$*.*

The proof is provided in Section 3.6.3.

3.4.3. General result for the weighted LASSO Huber estimator

In the next lemma we consider support recovery and ℓ_∞ bounds for a generic form of the weighted LASSO Huber estimator. This is similar to Zhou et al. (2009, Lemma 8.2). For clarity of formulation we shall impose (3.31), and (3.30) in the second part, as high-level conditions. These are taken care of in the preceding lemmas.

Lemma 3.15 (Weighted LASSO Huber)**.** *Consider model* (3.1) *under Assumption 3.1. Suppose that* (3.31) *holds true, and that the weights satisfy the mutual incoherence condition, that is for some* $\eta \in (0,1)$ *we have that*

$$\left| \widehat{Q}_{S^cS} \big(\widehat{Q}_{SS}\big)^{-1} \big(w_S \odot \widehat{\gamma}_S\big) \right| \le w_{S^c} (1 - \eta)\,, \tag{3.32}$$

where $\widehat{\gamma}$ *is the subgradient of the* ℓ_1 *norm of the estimator in the PDW construction. For the regularization parameter* λ_n *we assume that*

$$w_{\min}\big(S^c\big)\, \lambda_n > \frac{4\, C_{\mathrm{Q},\mathcal{L}}\, c_2^{\mathrm{Grad}}}{\eta} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}. \tag{3.33}$$

Furthermore, suppose that the robustification parameter α_n *is chosen in the range*

$$c_1^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \le \alpha_n \le \left(\frac{c_2^{\mathrm{Grad}}}{80\, C_{\mathrm{apx}}\, c^2_{\mathbf{X},\mathrm{sub}}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \right)^{\frac{1}{m-1}} \tag{3.34}$$

and

$$\beta^*_{\min} > \phi_{n,\infty,s}\,, \quad \text{where} \quad \phi_{n,\infty,s} = \frac{128}{c_{\mathbf{X},\mathrm{l}}} \max\left\{ c_2^{\mathrm{Grad}} \left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}}, \; w_{\max}\big(S\big)\, \lambda_n \sqrt{s} \right\}. \tag{3.35}$$

Then for $n \geq \max\left\{c_3^{\mathrm{RSC}} s \log(p), 6\log(p)\right\}$ *with probability at least*

$$1 - c_1^{\mathrm{P}} \exp(-c_2^{\mathrm{P}} n) - 2\exp(-2n) - \frac{4}{p^2}\,, \tag{3.36}$$

the weighted LASSO Huber estimator as a solution to the program (3.3) *is unique, given by* $\widehat{\beta}_n^{\mathrm{WLH}} = \widehat{\beta}_n^{\mathrm{PDW}}$ *and satisfies*

$$\operatorname{sign}\bigl(\widehat{\beta}_n^{\mathrm{WLH}}\bigr) = \operatorname{sign}\bigl(\beta^*\bigr) \qquad \text{and} \qquad \left\|\widehat{\beta}_n^{\mathrm{WLH}} - \beta^*\right\|_\infty \leq \phi_{n,\infty,s} \tag{3.37}$$

with $\phi_{n,\infty,s}$ *in* (3.35).
If in addition (3.30) *is assumed as well, we may replace* $\phi_{n,\infty,s}$ *in the beta-min condition* (3.35) *and in the* ℓ_∞ *bound in* (3.37) *by*

$$\phi_{n,\infty,f} = 4\,C_{\mathrm{Q,S}}\,\max\left\{c_2^{\mathrm{Grad}}\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}},\, w_{\max}\bigl(S\bigr)\,\lambda_n\right\}. \tag{3.38}$$

Proof of Lemma 3.15. We shall apply Lemma 3.12. Using the mutual incoherence condition (3.32), in order to show strict dual feasibility as in (3.26) it suffices to prove that

$$\begin{aligned}&\left\|\widehat{Q}_{S^cS}\bigl(\widehat{Q}_{SS}\bigr)^{-1}\Bigl(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\bigl(\beta^*_{\alpha_n}\bigr)\Bigr)_S - \Bigl(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\bigl(\beta^*_{\alpha_n}\bigr)\Bigr)_{S^c}\right\|_\infty \\ &\qquad + \left\|\Bigl(\widehat{Q}_{S^c(S_{\alpha_n}\setminus S)} - \widehat{Q}_{S^cS}\bigl(\widehat{Q}_{SS}\bigr)^{-1}\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\Bigr)\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\right\|_\infty < \frac{w_{\min}\bigl(S^c\bigr)\,\eta}{2}\,\lambda_n\,.\end{aligned} \tag{3.39}$$

The first term is bounded by (3.31) (which is satisfied by assumption). We prove in Section 3.6.4 that

$$\left\|\Bigl(\widehat{Q}_{S^c(S_{\alpha_n}\setminus S)} - \widehat{Q}_{S^cS}\bigl(\widehat{Q}_{SS}\bigr)^{-1}\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\Bigr)\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\right\|_\infty \leq 80\,C_{\mathrm{apx}}\,c^2_{\mathbf{X},\mathrm{sub}}\,\alpha_n^{m-1} \tag{3.40}$$

with probability at least $1 - 2\exp(-2n) - 2/p^2$. Then the choices of λ_n and α_n in (3.33) and (3.34) imply (3.39). Since the first beta-min condition in Lemma 3.12 is also satisfied in both cases by the choice of α_n in (3.34), the first part of that lemma up to (3.27) applies. Here we assumed that $\sqrt{s} \geq c_{\mathbf{X},\mathrm{l}}/(2560\,c^2_{\mathbf{X},\mathrm{sub}})$ for (3.35) and $320\,c^2_{\mathbf{X},\mathrm{sub}}\,C_{\mathrm{Q,S}} \geq 1$ for (3.38), which can be arranged by choosing the constants appropriately.
Now we show that $\phi_{n,\infty}$ in (3.27) is bounded by $\phi_{n,\infty,s}$ and, under the additional condition (3.30), is even bounded by $\phi_{n,\infty,f}$. Then (3.35) (or the analogous condition with $\phi_{n,\infty,f}$) implies the beta-min condition (3.28) in Lemma 3.12, which concludes the proof.
To this end, note that $\phi_{n,\infty}$ is bounded by four times the maximum of the summands in (3.27). In addition (3.22) leads to

$$4\,w_{\max}\bigl(S\bigr)\,\lambda_n\left\|\bigl(\widehat{Q}_{SS}\bigr)^{-1}\right\|_{\mathrm{M},\infty} \leq \frac{128\,w_{\max}\bigl(S\bigr)\,\lambda_n\sqrt{s}}{c_{\mathbf{X},\mathrm{l}}}\,,$$

and together with Lemma 3.8 and the lower bound of α_n in (3.34) this leads to

$$4 \left\| \left(\widehat{Q}_{SS}\right)^{-1} \right\|_{\mathrm{M},\infty} \left\| \left(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n}) \right)_S \right\|_\infty \le \frac{128\, c_2^{\mathrm{Grad}}}{c_{\mathbf{X},\mathrm{l}}} \left(\frac{s \log(p)}{n} \right)^{\frac{1}{2}}$$

with probability at least $1 - c_1^{\mathrm{P}} \exp(-c_2^{\mathrm{P}} n) - 2/p^2$. Further, Lemma 3.4 implies

$$4 \left\| \beta^*_{\alpha_n,S} - \beta^*_S \right\|_\infty \le 4 \left\| \beta^*_{\alpha_n} - \beta^* \right\|_2 \le 4\, C_{\mathrm{apx}}\, \alpha_n^{m-1} \le \frac{128\, c_2^{\mathrm{Grad}}}{c_{\mathbf{X},\mathrm{l}}} \left(\frac{s \log(p)}{n} \right)^{\frac{1}{2}}$$

with the choice of α_n in (3.34). Finally, in Section 3.6.4 we also show that

$$\left\| \widehat{Q}_{S(S_\alpha \setminus S)}\, \beta^*_{\alpha, S_\alpha \setminus S} \right\|_\infty \le 80\, C_{\mathrm{apx}}\, c^2_{\mathbf{X},\mathrm{sub}}\, \alpha_n^{m-1} \tag{3.41}$$

with high probability. Together with (3.22) this implies

$$\begin{aligned} 4 \left\| \left(\widehat{Q}_{SS}\right)^{-1} \right\|_{\mathrm{M},\infty} \left\| \widehat{Q}_{S(S_\alpha \setminus S)}\, \beta^*_{\alpha, S_\alpha \setminus S} \right\|_\infty &\le \frac{10240\, C_{\mathrm{apx}}\, c^2_{\mathbf{X},\mathrm{sub}}}{c_{\mathbf{X},\mathrm{l}}} \sqrt{s}\, \alpha_n^{m-1} \\ &\le \frac{128\, c_2^{\mathrm{Grad}}}{c_{\mathbf{X},\mathrm{l}}} \left(\frac{s \log(p)}{n} \right)^{\frac{1}{2}} \end{aligned}$$

by the choice of α_n, which concludes the proof of $\phi_{n,\infty} \le \phi_{n,\infty,s}$. To show $\phi_{n,\infty} \le \phi_{n,\infty,f}$ under the assumption (3.30), after arranging $80\, c^2_{\mathbf{X},\mathrm{sub}}\, C_{\mathrm{Q,S}} \ge 1$ we proceed analogously (and use the estimate (3.30) instead of (3.22) in the previous inequalities). This concludes the proof of the lemma. □

3.4.4. Adaptive LASSO with generic first-stage estimator

The next step is to provide a result on the sign-consistency and ℓ_∞ norm of the estimation error of the adaptive LASSO Huber estimator $\widehat{\beta}_n^{\mathrm{ALH}}$ in (3.3) with weights in (3.6) for a generic initial estimator, similar to Zhou et al. (2009, Theorem 4.3)

Lemma 3.16 (Adaptive LASSO Huber)**.** *Consider model* (3.1) *under Assumption 3.1. Suppose that* (3.31) *holds true, and that α_n is chosen according to* (3.34)*. For the estimation error*

$$\Delta_n := \widehat{\beta}_n^{\mathrm{init}} - \beta^*$$

of the initial estimator $\widehat{\beta}_n^{\mathrm{init}}$, we assume upper bounds of the form

$$\|\Delta_{n,S}\|_\infty \le a_n < 1\,, \qquad \|\Delta_{n,S^c}\|_\infty \le b_n < 1 \tag{3.42}$$

with sequences (a_n) and (b_n) tending to zero. Furthermore, assume that for some $\eta \in (0,1)$ and $C_\lambda > 4/\eta$ the regularization parameter is chosen from the range

$$\frac{4\, C_{\mathrm{Q},\mathcal{L}}\, c_2^{\mathrm{Grad}}\, b_n}{\eta} \left(\frac{\log(p)}{n} \right)^{\frac{1}{2}} < \lambda_n \le C_\lambda\, C_{\mathrm{Q},\mathcal{L}}\, c_2^{\mathrm{Grad}}\, b_n \left(\frac{\log(p)}{n} \right)^{\frac{1}{2}}, \tag{3.43}$$

and, in addition, suppose that there is a sequence $q_n \le (1-\eta)/b_n$*, which may grow if* $b_n \downarrow 0$*, such that*

$$\left\| \widehat{Q}_{S^cS} \big(\widehat{Q}_{SS} \big)^{-1} \right\|_{\mathrm{M},\infty} \le q_n \,. \tag{3.44}$$

Finally, setting

$$\phi_{n,\infty,s,1} = \frac{128}{c_{\mathbf{X},\mathrm{l}}} \max \left\{ c_2^{\mathrm{Grad}} \left(\frac{s \log(p)}{n} \right)^{\frac{1}{2}} , \, \lambda_n \sqrt{s} \right\},$$

suppose that the beta-min assumption

$$\beta^*_{\min} > \max \left\{ 2\, a_n \,,\, \phi_{n,\infty,s,1} \,,\, 2 \max \left\{ \frac{q_n}{1-\eta} \,, C_\lambda \, C_{\mathrm{Q},\mathcal{L}} \right\} b_n \right\} \tag{3.45}$$

is satisfied. Then for $n \ge \max\left\{ c_3^{\mathrm{RSC}} s \log(p), 6 \log(p) \right\}$ *with probability at least equal to* (3.36)*, the adaptive LASSO Huber estimator, given as a solution to the program* (3.3) *with weights in* (3.6)*, is unique, given by* $\widehat{\beta}_n^{\mathrm{ALH}} = \widehat{\beta}_n^{\mathrm{PDW}}$ *and satisfies*

$$\operatorname{sign}\big(\widehat{\beta}_n^{\mathrm{ALH}}\big) = \operatorname{sign}\big(\beta^*\big) \qquad \text{and} \qquad \left\| \widehat{\beta}_n^{\mathrm{ALH}} - \beta^* \right\|_\infty \le \phi_{n,\infty,s,1} \,. \tag{3.46}$$

If in addition (3.30) *is also assumed, we can replace* $\phi_{n,\infty,s,1}$ *in the beta-min condition* (3.45) *and in the upper bound of the* ℓ_∞ *norm of the estimation error by*

$$\phi_{n,\infty,f,1} = 4\, C_{\mathrm{Q,S}} \max \left\{ c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n} \right)^{\frac{1}{2}} , \, \lambda_n \right\} . \tag{3.47}$$

Proof of Lemma 3.16. We shall apply Lemma 3.15. To this end, we start by checking the mutual incoherence condition (3.32) and the condition (3.33) on the regularization parameter λ_n. For (3.33), since $\big|\widehat{\beta}^{\mathrm{init}}_{n,k}\big| = \big|\beta^*_k + \Delta_{n,k}\big| = \big|\Delta_{n,k}\big| \le \big\|\Delta_{n,S^c}\big\|_\infty$ for $k \in S^c$, we obtain from (3.42) that

$$w_{\min}\big(S^c\big) = \min_{k \in S^c} \left\{ \max \left\{ \big(\big|\widehat{\beta}^{\mathrm{init}}_{n,k}\big| \big)^{-1}, 1 \right\} \right\} \ge \big\|\Delta_{n,S^c}\big\|_\infty^{-1} \ge \frac{1}{b_n} \tag{3.48}$$

and hence $w_{\min}\big(S^c\big)\, \lambda_n \ge \lambda_n / b_n$, which together with the assumption (3.43) on λ_n gives (3.33). Next, we turn to the mutual incoherence condition (3.32), for which it suffices to prove

$$\left\| \widehat{Q}_{S^cS} \big(\widehat{Q}_{SS} \big)^{-1} \right\|_{\mathrm{M},\infty} \le \frac{w_{\min}\big(S^c\big)}{w_{\max}\big(S\big)} \big(1-\eta\big) \,. \tag{3.49}$$

From the beta-min condition (3.45) and the bounds in (3.42) we have in particular that $\beta^*_{\min}/2 > a_n \ge \|\Delta_{n,S}\|_\infty \ge \big|\Delta_{n,k}\big|$ and hence that

$$\big|\widehat{\beta}^{\mathrm{init}}_{n,k}\big| = \big|\beta^*_k + \Delta_{n,k}\big| \ge \big|\beta^*_k\big| - \big|\Delta_{n,k}\big| > \beta^*_{\min} - \frac{\beta^*_{\min}}{2} = \frac{\beta^*_{\min}}{2}$$

for $k \in S$. This together with the definition of the weights implies

$$w_{\max}\big(S\big) = \max_{k \in S} \Big\{ \max\big\{ \big(|\widehat{\beta}^{\,\mathrm{init}}_{n,k}|\big)^{-1}, 1\big\}\Big\} \le \max\{2/\beta^*_{\min}, 1\}\,. \tag{3.50}$$

In order to conclude (3.49) we consider two cases. If $\beta^*_{\min} \le 2$, then we have $w_{\max}\big(S\big) \le 2/\beta^*_{\min}$ because of (3.50) and hence with (3.48) and the last term in the beta-min condition (3.45) we obtain

$$\frac{w_{\min}\big(S^c\big)}{w_{\max}\big(S\big)}\big(1-\eta\big) \ge \frac{\beta^*_{\min}\,\big(1-\eta\big)}{2\,b_n} > q_n \ge \Big\|\widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\Big\|_{\mathrm{M},\infty}$$

by (3.44). If $\beta^*_{\min} > 2$, then $w_{\max}\big(S\big) \le 1$ and by (3.44), (3.48) and the choice of q_n it follows that

$$\frac{w_{\min}\big(S^c\big)}{w_{\max}\big(S\big)}\big(1-\eta\big) \ge \frac{1-\eta}{b_n} \ge q_n = \Big\|\widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\Big\|_{\mathrm{M},\infty}\,,$$

so that (3.49) is satisfied in both cases.

Next, we show that $\phi_{n,\infty,s} \le \phi_{n,\infty,s,1}$, then the beta-min condition (3.45) directly implies (3.35). Comparing $\phi_{n,\infty,s,1}$ and $\phi_{n,\infty,s}$ it remains to show that

$$\frac{128\,w_{\max}\big(S\big)\,\lambda_n\,\sqrt{s}}{c_{\mathbf{X},\mathrm{l}}} \le \frac{128}{c_{\mathbf{X},\mathrm{l}}}\,\max\left\{c_2^{\mathrm{Grad}}\left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}}, \lambda_n\,\sqrt{s}\right\}. \tag{3.51}$$

To this end, note that the last lower bound in the inequality (3.45) implies

$$\frac{128\,c_2^{\mathrm{Grad}}}{c_{\mathbf{X},\mathrm{l}}}\left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}} > \frac{128\,c_2^{\mathrm{Grad}}}{c_{\mathbf{X},\mathrm{l}}}\left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}}\frac{2\,C_\lambda\,C_{\mathrm{Q},\mathcal{L}}\,b_n}{\beta^*_{\min}} \ge \frac{256\,\lambda_n\sqrt{s}}{c_{\mathbf{X},\mathrm{l}}\,\beta^*_{\min}}$$

by the choice of the regularization parameter λ_n in (3.43). This together with (3.50) implies (3.51). So Lemma 3.15 applies and we conclude that the ℓ_∞ bound in (3.37) can be reduced to (3.46).

For the sharper bound, $\phi_{n,\infty,f} \le \phi_{n,\infty,f,1}$, under assumption (3.30), one argues similarly. This concludes the proof. □

3.4.5. Adaptive LASSO with the LASSO in the first stage

We start with the following lemma, which is analogous to Zhou et al. (2009, Lemma 4.2) and gives a superset $\overline{S}$ of the support S, the cardinality of which is of the same order s. This is used to determine the order of regularization in the adaptive LASSO Huber estimator in the lemma to follow.

Lemma 3.17 (Thresholding procedure)**.** *If the initial estimator $\widehat{\beta}^{\,\mathrm{init}}_n$ satisfies* (3.7)*, and if the following beta-min condition*

$$\beta^*_{\min} > 2\,C_{\mathrm{init}}\,\lambda^{\mathrm{init}}_n\sqrt{s} \tag{3.52}$$

holds, then the set $\overline{S} = \big\{k \in \{1,\dots,p\}\,\big|\,\big|\widehat{\beta}^{\,\mathrm{init}}_{n,k}\big| > \lambda^{\mathrm{init}}_n\big\}$ satisfies

$$S \subseteq \overline{S} \qquad \text{and} \qquad s \le \big|\overline{S}\big| \le 2\,C_{\mathrm{init}}\,s\,.$$

Proof of Lemma 3.17. Let $\Delta_n = \widehat{\beta}_n^{\text{init}} - \beta^*$. Then from (3.7) it follows that

$$\|\Delta_{n,S}\|_\infty \le \|\Delta_n\|_\infty \le \|\Delta_n\|_2 \le C_{\text{init}}\,\lambda_n^{\text{init}}\sqrt{s}\,,$$

and hence for all $k \in S$ that

$$\big|\widehat{\beta}_{n,k}^{\text{init}}\big| = \big|\beta_k^* + \Delta_{n,k}\big| \ge \big|\beta_k^*\big| - \big|\Delta_{n,k}\big| \ge \beta_{\min}^* - \|\Delta_{n,S}\|_\infty > C_{\text{init}}\,\lambda_n^{\text{init}}\sqrt{s}$$

because of inequality (3.52). As a consequence, the definition of the set $\overline{S}$ implies the membership $S \subseteq \overline{S}$. Furthermore, for $k \in S^c$ (since $\beta_k^* = 0$) it is

$$\big|\widehat{\beta}_{n,k}^{\text{init}}\big| = \big|\beta_k^* + \Delta_{n,k}\big| = \big|\Delta_{n,k}\big|$$

and the upper bound of the ℓ_1 norm of the estimation error in (3.7) leads to

$$\Big\|\widehat{\beta}_{n,S^c}^{\text{init}}\Big\|_1 = \|\Delta_{n,S^c}\|_1 \le \|\Delta_n\|_1 \le C_{\text{init}}\,\lambda_n^{\text{init}}\,s\,.$$

Hence, we include at most $C_{\text{init}}\,s$ more entries from S^c in $\overline{S}$, thus

$$s \le \big|\overline{S}\big| \le s + C_{\text{init}}\,s \le 2\,C_{\text{init}}\,s\,,$$

which completes the proof. □

Lemma 3.18. *Suppose Assumption 3.1 and* $n \ge \max\big\{c_3^{\text{RSC}} s\log(p), 6\log(p)\big\}$ *hold. Then*

$$\max_{k\in\{1\dots,p-s\}} \left\|\Big(e_k^\top \widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\Big)^\top\right\|_2 \le \frac{33\,c_{\mathbf{X},\text{sub}}}{\sqrt{c_{\mathbf{X},\text{l}}}} \qquad \textit{and}$$
$$\left\|\widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\right\|_{\text{M},\infty} \le \frac{33\,c_{\mathbf{X},\text{sub}}\sqrt{s}}{\sqrt{c_{\mathbf{X},\text{l}}}}$$

hold true with probability at least $1 - c_1^{\text{P}}\exp(-c_2^{\text{P}} n) - 2/p^2$.

The technical proof of this lemma is deferred to Section 3.6.3.

For clarity of formulation in the following result we shall again impose (3.31), and (3.30) in the second part, as high-level conditions. Theorem 3.2 then follows from the following Lemma 3.19 together with the Lemmas 3.13 and 3.14.

Lemma 3.19 (Adaptive LASSO Huber with LASSO in the first stage)**.** *Consider model* (3.1) *under Assumption 3.1. Suppose that* (3.31) *holds true, and that the robustification parameter* α_n *is chosen according to* (3.34)*. Suppose that the initial estimator* $\widehat{\beta}_n^{\text{init}}$ *satisfies* (3.7) *with* $\lambda_n^{\text{init}} = C_{\lambda,\text{init}}\,(\log(p)/n)^{\frac{1}{2}}$ *for some constant* $C_{\lambda,\text{init}} \ge 16\,c_2^{\text{Grad}}/c_{\mathbf{X},\text{l}}$*, and that for suitable* $\eta \in (0,1)$ *and* $C_{\lambda,\text{L}} > 4\,(2\,C_{\text{init}})^{\frac{1}{2}}/\eta$ *the regularization parameter is chosen from the range*

$$\frac{4\,C_{\text{Q},\mathcal{L}}\,c_2^{\text{Grad}}\,C_{\text{init}}\,\lambda_n^{\text{init}}}{\eta}\left(\frac{\big|\overline{S}\big|\log(p)}{n}\right)^{\frac{1}{2}} < \lambda_n \le C_{\lambda,\text{L}}\,C_{\text{Q},\mathcal{L}}\,c_2^{\text{Grad}}\,\lambda_n^{\text{init}}\left(\frac{C_{\text{init}}\big|\overline{S}\big|\log(p)}{2\,n}\right)^{\frac{1}{2}} \tag{3.53}$$

with $\overline{S} = \{k \in \{1,\dots,p\} \,|\, |\widehat{\beta}_{n,k}^{\text{init}}| > \lambda_n^{\text{init}}\}$ *as above. In addition, suppose that the sample size satisfies*

$$n \geq \max\left\{ \left(\frac{33\, c_{\mathbf{X},\text{sub}}\, C_{\text{init}}\, C_{\lambda,\text{init}}}{(1-\eta)\sqrt{c_{\mathbf{X},\text{l}}}}\right)^2 s^2 \log(p)\,, \right.$$

$$\left. \max\left\{ c_3^{\text{RSC}}\,,\, \left(\frac{64\, c_2^{\text{Grad}}}{c_{\mathbf{X},\text{l}}}\right)^2 \right\} s \log(p)\,,\, 6\log(p) \right\}, \tag{3.54}$$

and that we have the beta-min condition

$$\beta^*_{\min} > 2\max\left\{ \frac{33\, c_{\mathbf{X},\text{sub}} \sqrt{s}}{\sqrt{c_{\mathbf{X},\text{l}}}\,(1-\eta)}, C_{\lambda,\text{L}}\, C_{\text{Q},\mathcal{L}} \right\} C_{\text{init}}\, \lambda_n^{\text{init}} \sqrt{s}\,. \tag{3.55}$$

Then with probability at least

$$1 - c_1^{\text{P}} \exp(-c_2^{\text{P}} n) - 2\exp(-2n) - \frac{4}{p^2}$$

the adaptive LASSO Huber estimator, given as a solution to the program (3.3) *with weights in* (3.6)*, is unique, given by* $\widehat{\beta}_n^{\text{ALH}} = \widehat{\beta}_n^{\text{PDW}}$ *and satisfies*

$$\operatorname{sign}\big(\widehat{\beta}_n^{\text{ALH}}\big) = \operatorname{sign}\big(\beta^*\big) \qquad \text{and} \qquad \left\|\widehat{\beta}_n^{\text{ALH}} - \beta^*\right\|_\infty \leq 2\, C_{\lambda,\text{L}}\, C_{\text{Q},\mathcal{L}}\, C_{\text{init}}\, \lambda_n^{\text{init}} \sqrt{s}\,. \tag{3.56}$$

If in addition (3.30) *is assumed as well, the upper bound of the* ℓ_∞ *norm of the estimation error in* (3.56) *reduces to*

$$\left\|\widehat{\beta}_n^{\text{ALH}} - \beta^*\right\|_\infty \leq \max\left\{ \frac{4\, C_{\text{Q,S}}\, c_2^{\text{Grad}}}{C_{\lambda,\text{init}}}\,,\, \frac{C_{\lambda,\text{L}}\, C_{\text{Q,S}}\, C_{\text{Q},\mathcal{L}}\, C_{\text{init}}\, c_{\mathbf{X},\text{l}}}{16} \right\} \lambda_n^{\text{init}}\,. \tag{3.57}$$

Proof of Lemma 3.19. We shall apply Lemma 3.16. To check the assumptions, for (3.42) using (3.7) we get $\|\Delta_n\|_\infty \leq \|\Delta_n\|_2 \leq C_{\text{init}}\, \lambda_n^{\text{init}} \sqrt{s} = a_n = b_n$. For the lower bound in (3.43), using (3.53), Lemma 3.17 and the choice of b_n we estimate

$$\begin{aligned}\lambda_n &> \frac{4\, C_{\text{Q},\mathcal{L}}\, c_2^{\text{Grad}}\, C_{\text{init}}\, \lambda_n^{\text{init}}}{\eta} \left(\frac{|\overline{S}|\log(p)}{n}\right)^{\frac{1}{2}} \geq \frac{4\, C_{\text{Q},\mathcal{L}}\, c_2^{\text{Grad}}\, C_{\text{init}}\, \lambda_n^{\text{init}}}{\eta} \left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}} \\ &= \frac{4\, C_{\text{Q},\mathcal{L}}\, c_2^{\text{Grad}}\, b_n}{\eta} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}},\end{aligned}$$

and similarly for the upper bound

$$\lambda_n \leq C_{\lambda,\text{L}}\, C_{\text{Q},\mathcal{L}}\, c_2^{\text{Grad}}\, C_{\text{init}}\, \lambda_n^{\text{init}} \left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}} = C_{\lambda,\text{L}}\, C_{\text{Q},\mathcal{L}}\, c_2^{\text{Grad}}\, b_n \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}$$

with $C_{\lambda,\mathrm{L}} > 4/\eta$. Next, (3.44) follows from Lemma 3.18 with $q_n = 33\, c_{\mathbf{X},\mathrm{sub}}\, \sqrt{s}/\sqrt{c_{\mathbf{X},\mathrm{l}}}$ with high probability. In addition, the choice of b_n and the lower bound (3.54) of the sample size implies

$$q_n \le \frac{33\, c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}} \, \frac{(1-\eta)\, \sqrt{c_{\mathbf{X},\mathrm{l}}}}{33\, c_{\mathbf{X},\mathrm{sub}}\, C_{\mathrm{init}}\, C_{\lambda,\mathrm{init}}\, (s\log(p)/n)^{\frac{1}{2}}} = \frac{1-\eta}{b_n}\,.$$

So, finally we have to check the beta-min condition in (3.45), which concludes the proof of the lemma in this setting. The last term in the maximum is given by (3.55) and the choice of b_n and q_n, and $\beta^*_{\min} \ge 2\, a_n$ is clear because of the choice of a_n and (3.55). Hence for applying Lemma 3.16 it remains to show that

$$\phi_{n,\infty,s,1} \le 2\, C_{\lambda,\mathrm{L}}\, C_{\mathrm{Q},\mathcal{L}}\, C_{\mathrm{init}}\, \lambda_n^{\mathrm{init}} \sqrt{s}\,.$$

This bound implies then also (3.56) because of (3.46). It is

$$\begin{aligned} \frac{128\, c_2^{\mathrm{Grad}}}{c_{\mathbf{X},\mathrm{l}}} \left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}} &= \frac{128\, c_2^{\mathrm{Grad}}}{c_{\mathbf{X},\mathrm{l}}\, C_{\lambda,\mathrm{init}}}\, \lambda_n^{\mathrm{init}} \sqrt{s} \\ &\le 8\, C_{\mathrm{Q},\mathcal{L}}\, C_{\mathrm{init}}\, \lambda_n^{\mathrm{init}} \sqrt{s} \le 2\, C_{\lambda,\mathrm{L}}\, C_{\mathrm{Q},\mathcal{L}}\, C_{\mathrm{init}}\, \lambda_n^{\mathrm{init}} \sqrt{s} \end{aligned}$$

since $16\, c_2^{\mathrm{Grad}} \le c_{\mathbf{X},\mathrm{l}}\, C_{\lambda,\mathrm{init}}\, C_{\mathrm{Q},\mathcal{L}}\, C_{\mathrm{init}}$. Moreover, (3.53) and (3.54) together with Lemma 3.17 lead to

$$\begin{aligned} \frac{128\, \lambda_n\, \sqrt{s}}{c_{\mathbf{X},\mathrm{l}}} &\le \frac{128\, C_{\lambda,\mathrm{L}}\, C_{\mathrm{Q},\mathcal{L}}\, c_2^{\mathrm{Grad}}\, \lambda_n^{\mathrm{init}}}{c_{\mathbf{X},\mathrm{l}}} \left(\frac{C_{\mathrm{init}}\, \lvert\overline{S}\rvert \log(p)}{2\, n}\right)^{\frac{1}{2}} \frac{c_{\mathbf{X},\mathrm{l}}}{64\, c_2^{\mathrm{Grad}}} \left(\frac{n}{\log(p)}\right)^{\frac{1}{2}} \\ &\le 2\, C_{\lambda,\mathrm{L}}\, C_{\mathrm{Q},\mathcal{L}}\, C_{\mathrm{init}}\, \lambda_n^{\mathrm{init}} \sqrt{s}\,. \end{aligned}$$

Under the stronger assumption (3.30) we show

$$\phi_{n,\infty,f,1} \le \max\left\{ \frac{4\, C_{\mathrm{Q},\mathrm{S}}\, c_2^{\mathrm{Grad}}}{C_{\lambda,\mathrm{init}}}\,,\ \frac{C_{\lambda,\mathrm{L}}\, C_{\mathrm{Q},\mathrm{S}}\, C_{\mathrm{Q},\mathcal{L}}\, C_{\mathrm{init}}\, c_{\mathbf{X},\mathrm{l}}}{16} \right\} \lambda_n^{\mathrm{init}}\,,$$

which implies (3.57) because of (3.47). Note that the upper bound is obviously also smaller than the right term in (3.55). It is easy to see that

$$4\, C_{\mathrm{Q},\mathrm{S}}\, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} = \frac{4\, C_{\mathrm{Q},\mathrm{S}}\, c_2^{\mathrm{Grad}}}{C_{\lambda,\mathrm{init}}}\, \lambda_n^{\mathrm{init}}$$

and

$$\begin{aligned} 4\, C_{\mathrm{Q},\mathrm{S}}\, \lambda_n &\le 4\, C_{\lambda,\mathrm{L}}\, C_{\mathrm{Q},\mathrm{S}}\, C_{\mathrm{Q},\mathcal{L}}\, C_{\mathrm{init}}\, \lambda_n^{\mathrm{init}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \frac{c_{\mathbf{X},\mathrm{l}}}{64} \left(\frac{n}{\log(p)}\right)^{\frac{1}{2}} \\ &= \frac{C_{\lambda,\mathrm{L}}\, C_{\mathrm{Q},\mathrm{S}}\, C_{\mathrm{Q},\mathcal{L}}\, C_{\mathrm{init}}\, c_{\mathbf{X},\mathrm{l}}}{16}\, \lambda_n^{\mathrm{init}} \end{aligned}$$

by (3.53), (3.54) and Lemma 3.17, which concludes the proof. □

3.5. Conclusions

In their recent paper, Sun et al. (2020) extended the analysis from Fan et al. (2017) to fixed designs, as well as to conditional moments of ε_1 of order strictly smaller than 2, in which case they showed that the rates of convergence deteriorate. Results on support estimation, rates of convergence in the ℓ_∞ norm together with a data-driven choice of the robustification parameter would be of some interest in this setting as well.

Furthermore, Fan et al. (2016) and Sun et al. (2020) suggest also robustification of the covariates, which would be of major interest as well, especially for the estimation of the second moments in the linear random coefficient regression model. For more details see Section 6.2.

Another possible extension or modification of our method would be the use of nonconvex penalty functions such as smoothly clipped absolute deviation (SCAD) as in Loh and Wainwright (2017) and Loh (2017), with the methodological aim to achieve milder beta-min conditions.

Another extension of some interest would be to robustify asymmetric versions of least squares regression (Newey and Powell, 1987; Gu and Zou, 2016), that is, high-dimensional expectile regression.

3.6. Technical proofs

At first we introduce further notations. For a random variable $Y \in \mathbb{R}$ we write $Y \sim \operatorname{subG}(\tau)$ with $\tau > 0$ if $\mathbb{P}(|Y| \geq t) \leq 2\,\exp\big(-t^2/(2\,\tau^2)\big)$ for all $t \geq 0$, and for a random vector $\mathbf{Y} \in \mathbb{R}^d$ we write $\mathbf{Y} \sim \operatorname{subG}_d(\tau)$ if $\mathbb{P}(|v^\top \mathbf{Y}| \geq t) \leq 2\,\exp\big(-t^2/(2\,\tau^2\|v\|_2^2)\big)$ for all $v \in \mathbb{R}^d \setminus \{\mathbf{0}_d\}$ and $t \geq 0$. In addition, a random variable $Y \sim \operatorname{subE}(\tau, b)$ is called sub-Exponential with $\tau, b > 0$ if $\mathbb{E}[Y] = 0$ and $\mathbb{E}\big[\exp(t\,Y)\big] \leq \exp\big(t^2\tau^2/2\big)$ for all $|t| < 1/b$. Furthermore, we denote by $\vec{X}_1, \dots, \vec{X}_p \in \mathbb{R}^n$ the columns of $\mathbb{X}_n$ and the rows are $\mathbf{X_i} = (X_{i,1}, \dots, X_{i,p})^\top \in \mathbb{R}^p$.

3.6.1. Proofs for Section 3.4.1

Proof of Lemma 3.4. Let $l(x) = x^2$, then by (ii) of Assumption 3.1 we get

$$\begin{aligned}\mathbb{E}\Big[l\big(Y_1 - \mathbf{X_1}^\top \beta^*_{\alpha_n}\big) - l\big(Y_1 - \mathbf{X_1}^\top \beta^*\big)\Big] &= \big(\beta^*_{\alpha_n} - \beta^*\big)^\top \mathbb{E}\Big[\mathbf{X_1}\mathbf{X_1}^\top\Big]\,\big(\beta^*_{\alpha_n} - \beta^*\big) \\ &\geq c_{\mathbf{X},\mathrm{l}}\,\big\|\beta^*_{\alpha_n} - \beta^*\big\|_2^2\,. \end{aligned} \tag{3.58}$$

Let $g_{\alpha_n}(x) = l(x) - l_{\alpha_n}(x) = x^2 - 2\alpha_n^{-2}\big(\sqrt{1+\alpha_n^2 x^2} - 1\big)$, then

$$\begin{aligned}\mathbb{E}\Big[l\big(Y_1 - \mathbf{X_1}^\top \beta^*_{\alpha_n}\big) - l\big(Y_1 - \mathbf{X_1}^\top \beta^*\big)\Big] &= \mathbb{E}\Big[l\big(Y_1 - \mathbf{X_1}^\top \beta^*_{\alpha_n}\big) - l_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top \beta^*_{\alpha_n}\big)\\ &\qquad + l_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top \beta^*_{\alpha_n}\big) - l_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top \beta^*\big)\\ &\qquad + l_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top \beta^*\big) - l\big(Y_1 - \mathbf{X_1}^\top \beta^*\big)\Big]\\ &\le \mathbb{E}\Big[g_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top \beta^*_{\alpha_n}\big) - g_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top \beta^*\big)\Big]\end{aligned} \tag{3.59}$$

because $\beta^*_{\alpha_n}$ minimizes $\mathbb{E}\big[l_{\alpha_n}(Y_1 - \mathbf{X_1}^\top\beta)\big]$ over $\|\beta\|_2 \le C_\beta$ and $\|\beta^*\|_2 \le C_\beta$ by (iv) of Assumption 3.1. Furthermore, the mean value theorem implies

$$\begin{aligned}\mathbb{E}\Big[g_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top \beta^*_{\alpha_n}\big) - g_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top \beta^*\big)\Big] &= \mathbb{E}\Big[g'_{\alpha_n}(Z)\big(\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big)\Big]\\ &\le \mathbb{E}\Big[\big|g'_{\alpha_n}(Z)\big|\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\,\mathbb{1}\{|Z| \ge \alpha_n^{-1}\}\Big]\\ &\quad + \mathbb{E}\Big[\big|g'_{\alpha_n}(Z)\big|\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\,\mathbb{1}\{|Z| < \alpha_n^{-1}\}\Big]\end{aligned} \tag{3.60}$$

with $Z = Y_1 - \mathbf{X_1}^\top\widetilde{\beta}$ and $\widetilde{\beta}$ between β^* and $\beta^*_{\alpha_n}$. Note that $\widetilde{\beta}$ is also a random vector. For the first summand we obtain from (3.12) that

$$\begin{aligned}&\mathbb{E}\Big[\big|g'_{\alpha_n}(Z)\big|\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\,\mathbb{1}\{|Z| \ge \alpha_n^{-1}\}\Big]\\ &\qquad \le 2\,\mathbb{E}\bigg[|Z|\bigg(1 - \frac{1}{\sqrt{1+\alpha_n^2 Z^2}}\bigg)\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\,\mathbb{1}\{|Z| \ge \alpha_n^{-1}\}\bigg].\end{aligned}$$

Let $\mathbb{P}_\varepsilon$ be distribution of ε_1 conditional on $\mathbf{X_1}$ and $\mathbb{E}_\varepsilon$ the corresponding conditional expectation. Then we get the inequality

$$\begin{aligned}\mathbb{E}_\varepsilon\bigg[|Z|\bigg(1 - \frac{1}{\alpha_n^2 Z^2}\bigg)\mathbb{1}\{|Z| \ge \alpha_n^{-1}\}\bigg] &\le \mathbb{E}_\varepsilon\big[|Z|\,\mathbb{1}\{|Z| \ge \alpha_n^{-1}\}\big]\\ &= \int_0^\infty \mathbb{P}_\varepsilon\big(|Z| \ge \alpha_n^{-1}\,,\,|Z| > t\big)\,dt\\ &= \int_{\alpha_n^{-1}}^\infty \mathbb{P}_\varepsilon\big(|Z| > t\big)\,dt + \int_0^{\alpha_n^{-1}} \mathbb{P}_\varepsilon\big(|Z| \ge \alpha_n^{-1}\big)\,dt\\ &\le \int_{\alpha_n^{-1}}^\infty \frac{\mathbb{E}_\varepsilon\big[|Z|^m\big]}{t^m}\,dt + \int_0^{\alpha_n^{-1}} \frac{\mathbb{E}_\varepsilon\big[|Z|^m\big]}{\alpha_n^{-m}}\,dt\\ &= \frac{\alpha_n^{m-1}}{m-1}\,\mathbb{E}_\varepsilon\big[|Z|^m\big] + \alpha_n^{m-1}\,\mathbb{E}_\varepsilon\big[|Z|^m\big]\\ &\le 2\,\alpha_n^{m-1}\,\mathbb{E}_\varepsilon\big[|Z|^m\big]\,,\end{aligned}$$

where $m \in \{2,3\}$ is given in Assumption 3.1, and in consequence

$$\mathbb{E}\Big[\big|g'_{\alpha_n}(Z)\big|\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\,\mathbb{1}\{|Z| \geq \alpha_n^{-1}\}\Big] \leq 4\,\alpha_n^{m-1}\,\mathbb{E}\Big[|Z|^m\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\Big]\,. \tag{3.61}$$

Now we analyze the second term in (3.60). Taking the derivative in the series expansion

$$g_{\alpha_n}(x) = -2\sum_{k=2}^{\infty}\binom{1/2}{k}\alpha_n^{2k-2}\,x^{2k}\,, \qquad \alpha^2 x^2 \leq 1\,, \tag{3.62}$$

implies that

$$\big|g'_{\alpha_n}(x)\big| = \left|\,-2\sum_{k=2}^{\infty}\binom{1/2}{k}2k\,\alpha_n^{2k-2}\,x^{2k-1}\right| \leq \big|\alpha_n^2\,x^3\big| = \alpha_n^2\,|x|^3\,,$$

and hence that

$$\mathbb{E}\Big[\big|g'_{\alpha_n}(Z)\big|\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\,\mathbb{1}\{|Z| < \alpha_n^{-1}\}\Big] \leq \alpha_n^2\,\mathbb{E}\Big[|Z|^3\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\,\mathbb{1}\{|Z| < \alpha_n^{-1}\}\Big]$$

because $\alpha_n|Z| < 1$. Moreover, it is

$$\begin{aligned}\alpha_n^2\,\mathbb{E}_\varepsilon\big[|Z|^3\,\mathbb{1}\{|Z| < \alpha_n^{-1}\}\big] &= \alpha_n^2\,\mathbb{E}_\varepsilon\big[|Z|^m\,|Z|^{3-m}\,\mathbb{1}\{|Z| < \alpha_n^{-1}\}\big] \\ &\leq \alpha_n^{2+m-3}\,\mathbb{E}_\varepsilon\big[|Z|^m\,\mathbb{1}\{|Z| < \alpha_n^{-1}\}\big] \leq \alpha_n^{m-1}\,\mathbb{E}_\varepsilon\big[|Z|^m\big]\,,\end{aligned}$$

and in consequence

$$\mathbb{E}\Big[\big|g'_{\alpha_n}(Z)\big|\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\,\mathbb{1}\{|Z| < \alpha_n^{-1}\}\Big] \leq \alpha_n^{m-1}\,\mathbb{E}\Big[|Z|^m\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\Big]\,. \tag{3.63}$$

So in total we obtain by (3.58) - (3.63) the inequality

$$\big\|\beta^*_{\alpha_n} - \beta^*\big\|_2^2 \leq \frac{5}{c_{\mathbf{X},\mathrm{l}}}\,\mathbb{E}\Big[|Z|^m\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\Big]\,\alpha_n^{m-1}\,. \tag{3.64}$$

The mean on the right-hand side can be upper bounded by

$$\begin{aligned}\mathbb{E}\Big[|Z|^m\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\Big] &= \mathbb{E}\Big[\big|\varepsilon_1 + \mathbf{X_1}^\top(\beta^* - \widetilde{\beta})\big|^m\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\Big] \\ &\leq 2^{m-1}\bigg(\mathbb{E}\Big[|\varepsilon_1|^m\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\Big] \\ &\qquad + \mathbb{E}\Big[\big|\mathbf{X_1}^\top(\beta^* - \widetilde{\beta})\big|^m\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\Big]\bigg)\,.\end{aligned} \tag{3.65}$$

Moreover, for the first term in the brackets we obtain by Hölder's inequality and (i) of Assumption 3.1

$$\begin{aligned}\mathbb{E}\Big[|\varepsilon_1|^m\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\Big] &= \mathbb{E}\Big[\mathbb{E}\big[|\varepsilon_1|^m\big|\mathbf{X_1}\big]\,\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|\Big] \\ &\leq \mathbb{E}\Big[\mathbb{E}\big[|\varepsilon_1|^m\big|\mathbf{X_1}\big]^q\Big]^{\frac{1}{q}}\,\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|^{\frac{q}{q-1}}\Big]^{\frac{q-1}{q}} \\ &\leq (C_{\epsilon,\mathrm{m}})^{\frac{1}{q}}\,\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|^{\frac{q}{q-1}}\Big]^{\frac{q-1}{q}}\,.\end{aligned}$$

In addition, note that $\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n}) \sim \operatorname{subG}\big(c_{\mathbf{X},\text{sub}} \left\|\beta^* - \beta^*_{\alpha_n}\right\|_2\big)$ by (iii) of Assumption 3.1, and that the moments of a sub-Gaussian random variable $Q \sim \operatorname{subG}(\tau)$ with $\tau > 0$ are bounded by

$$\mathbb{E}\big[|Q|^r\big] \leq \big(2\tau^2\big)^{\frac{r}{2}}\, r\, \Gamma\Big(\frac{r}{2}\Big), \quad \mathbb{E}\big[|Q|^r\big]^{\frac{1}{r}} \leq \sqrt{2}\left(r\, \Gamma\Big(\frac{r}{2}\Big)\right)^{\frac{1}{r}} \tau \tag{3.66}$$

for $r > 1$. This can be proven analogously to Rigollet and Hütter (2019, Lemma 1.4). Hence

$$\mathbb{E}\Big[|\varepsilon_1|^m \left|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\right|\Big] \leq \sqrt{2}\,(C_{\epsilon,\mathrm{m}})^{\frac{1}{q}} \left(\frac{q}{q-1}\,\Gamma\Big(\frac{q}{2(q-1)}\Big)\right)^{\frac{q-1}{q}} c_{\mathbf{X},\text{sub}} \left\|\beta^* - \beta^*_{\alpha_n}\right\|_2 . \tag{3.67}$$

For the second term in the brackets in (3.65) the Cauchy-Schwarz inequality implies

$$\begin{aligned}\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\beta^* - \widetilde{\beta})\big|^m \left|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\right|\Big] &\leq \left(\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\beta^* - \widetilde{\beta})\big|^{2m}\Big]\,\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\big|^2\Big]\right)^{\frac{1}{2}} \\ &\leq 2\,\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\beta^* - \widetilde{\beta})\big|^{2m}\Big]^{\frac{1}{2}} c_{\mathbf{X},\text{sub}} \left\|\beta^* - \beta^*_{\alpha_n}\right\|_2 .\end{aligned} \tag{3.68}$$

To give a upper bound for the remaining expected value we consider at first a tail bound for the appropriate random variable. Let L be the line between β^* and $\beta^*_{\alpha_n}$, then L is also the convex hull of $\mathcal{V}(L) = \{\beta^*, \beta^*_{\alpha_n}\}$ and we obtain

$$\mathbb{P}\Big(\big|(\beta^* - \widetilde{\beta})^\top \mathbf{X_1}\big| > x\Big) \leq \mathbb{P}\big(\max_{u \in L} \left|u^\top \mathbf{X_1}\right| > x\big)$$

for $x \geq 0$ because $\widetilde{\beta}$ lies between β^* and $\beta^*_{\alpha_n}$. Moreover, $\mathbf{X_1}^\top\beta^*$ and $\mathbf{X_1}^\top\beta^*_{\alpha_n}$ are sub-Gaussian with variance proxy $C_\beta^2\, c^2_{\mathbf{X},\text{sub}}$ by (iii) and (iv) of Assumption 3.1 and $\left\|\beta^*_{\alpha_n}\right\|_2 \leq C_\beta$ by (3.5). Hence Rigollet and Hütter (2019, Theorem 1.16) leads to

$$\mathbb{P}\big(\big|(\beta^* - \widetilde{\beta})^\top \mathbf{X_1}\big| > x\big) \leq \mathbb{P}\big(\max_{u \in L} \left|u^\top \mathbf{X_1}\right| > x\big) \leq 4\, \exp\left(-\frac{x^2}{2C_\beta^2\, c^2_{\mathbf{X},\text{sub}}}\right).$$

In addition, Rigollet and Hütter (2019, Lemma 1.4) and the corresponding proof imply

$$\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\beta^* - \widetilde{\beta})\big|^{2m}\Big] \leq 2\left(2C_\beta^2\, c^2_{\mathbf{X},\text{sub}}\right)^m (2m)!\,\Gamma(m)\,. \tag{3.69}$$

In total (3.64) - (3.69) leads to

$$\left\|\beta^*_{\alpha_n} - \beta^*\right\|_2 \leq C_{\text{apx}}\, \alpha_n^{m-1}$$

with

$$C_{\mathrm{apx}} = \frac{5\, 2^m\, c_{\mathbf{X},\mathrm{sub}}}{c_{\mathbf{X},\mathrm{l}}} \left((C_{\epsilon,\mathrm{m}})^{\frac{1}{q}} \left(\frac{q}{q-1}\, \Gamma\Big(\frac{q}{2(q-1)} \Big) \right)^{\frac{q-1}{q}} \right.$$
$$\left. + \left(2 \left(2C_\beta^2\, c_{\mathbf{X},\mathrm{sub}}^2 \right)^m (2m)!\, \Gamma(m) \right)^{\frac{1}{2}} \right).$$

□

Proof of Lemma 3.6. We obtain

$$\left\langle \nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta+\Delta) - \nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta), \Delta \right\rangle = \frac{1}{n} \sum_{i=1}^{n} \Big(l'_\alpha\big(Y_i - \mathbf{X_i}^\top \beta\big) - l'_\alpha\big(Y_i - \mathbf{X_i}^\top (\beta+\Delta)\big) \Big)\, \mathbf{X_i}^\top \Delta$$

for $\beta, \Delta \in \mathbb{R}^p$ by (3.14). Firstly we show that

$$\left\langle \nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta+\Delta) - \nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta), \Delta \right\rangle \geq \frac{1}{2n} \sum_{i=1}^{n} \varphi_{\tau\|\Delta\|_2}\Big(\mathbf{X_i}^\top \Delta\, \mathbb{1}\{|Y_i - \mathbf{X_i}^\top \beta| \leq T\} \Big) \tag{3.70}$$

for all $\alpha \leq 1/(T + 8\tau\, C_\beta)$ and $(\beta, \Delta) \in A := \big\{ (\beta,\Delta) : \|\beta\|_2 \leq 4C_\beta \text{ and } \|\Delta\|_2 \leq 8C_\beta \big\}$, where

$$\varphi_t(u) = u^2\, \mathbb{1}\{|u| \leq t/2\} + \big(t - |u|\big)^2\, \mathbb{1}\{t/2 < |u| \leq t\}$$

and

$$T = 96\, \frac{c_{\mathbf{X},\mathrm{sub}}^2 \sqrt{c_{\mathbf{X},\mathrm{u}}}\, C_\beta}{c_{\mathbf{X},\mathrm{l}}}, \qquad \tau = \max\left\{ 4c_{\mathbf{X},\mathrm{sub}} \sqrt{\log(12 c_{\mathbf{X},\mathrm{sub}}^2 / c_{\mathbf{X},\mathrm{l}})}, 1 \right\}.$$

The function φ_t satisfies obviously $\varphi_t(u) \leq u^2\, \mathbb{1}\{|u| \leq t\}$. Let $i \in \{1, \ldots, n\}$ be fixed, then we get on the one hand

$$\varphi_{\tau\|\Delta\|_2}\Big(\mathbf{X_i}^\top \Delta\, \mathbb{1}\{|Y_i - \mathbf{X_i}^\top \beta| \leq T\} \Big) = 0$$

if $|\mathbf{X_i}^\top \Delta| > \tau \|\Delta\|_2$ or $|Y_i - \mathbf{X_i}^\top \beta| > T$. In addition, we have always

$$\Big(l'_\alpha\big(Y_i - \mathbf{X_i}^\top \beta\big) - l'_\alpha\big(Y_i - \mathbf{X_i}^\top (\beta+\Delta)\big) \Big)\, \mathbf{X_i}^\top \Delta \geq 0$$

because of the convexity of $g(\beta) = l_\alpha(Y_i - \mathbf{X_i}^\top \beta)$. On the other hand, if $|\mathbf{X_i}^\top \Delta| \leq \tau \|\Delta\|_2$ and $|Y_i - \mathbf{X_i}^\top \beta| \leq T$, we get

$$\big| Y_i - \mathbf{X_i}^\top \beta \big| \leq T \leq \alpha^{-1}$$

and

$$\big| Y_i - \mathbf{X_i}^\top (\beta+\Delta) \big| \leq \big| Y_i - \mathbf{X_i}^\top \beta \big| + \big| \mathbf{X_i}^\top \Delta \big| \leq T + \tau \|\Delta\|_2 \leq T + 8\tau\, C_\beta \leq \alpha^{-1}$$

because $(\beta, \Delta) \in A$ and the choice of α. In addition, the mean value theorem implies

$$\begin{aligned} l'_\alpha\big(Y_i - \mathbf{X_i}^\top\beta\big) - l'_\alpha\big(Y_i - \mathbf{X_i}^\top(\beta+\Delta)\big) &= l''_\alpha(c)\Big(Y_i - \mathbf{X_i}^\top\beta - Y_i + \mathbf{X_i}^\top(\beta+\Delta)\Big) \\ &= l''_\alpha(c)\,\mathbf{X_i}^\top\Delta \end{aligned}$$

with $c \in \big(Y_i - \mathbf{X_i}^\top\beta, Y_i - \mathbf{X_i}^\top(\beta+\Delta)\big)$ since the pseudo Huber loss l_α is twice differentiable. The above conditions lead to $|c| \le \alpha^{-1}$ as well. Moreover, note that

$$l''_\alpha(c) = \frac{2\alpha^{-3}}{(\alpha^{-2}+c^2)^{3/2}} \ge \frac{2\alpha^{-3}}{(2\alpha^{-2})^{3/2}} = \frac{2}{2^{3/2}} \ge \frac{1}{2}$$

for all $|c| \le \alpha^{-1}$. Hence it follows that

$$\begin{aligned} \Big(l'_\alpha\big(Y_i - \mathbf{X_i}^\top\beta\big) - l'_\alpha\big(Y_i - \mathbf{X_i}^\top(\beta+\Delta)\big)\Big)\,\mathbf{X_i}^\top\Delta = l''_\alpha(c)\,\big(\mathbf{X_i}^\top\Delta\big)^2 &\ge \frac{1}{2}\,\big(\mathbf{X_i}^\top\Delta\big)^2 \\ &\ge \frac{1}{2}\,\varphi_{\tau\|\Delta\|_2}\Big(\mathbf{X_i}^\top\Delta\,\mathbb{1}\{|Y_i - \mathbf{X_i}^\top\beta| \le T\}\Big) \end{aligned}$$

if $|\mathbf{X_i}^\top\Delta| \le \tau\|\Delta\|_2$ and $|Y_i - \mathbf{X_i}^\top\beta| \le T$. So in total inequality (3.70) is satisfied for all $(\beta, \Delta) \in A$ and $\alpha \le 1/(T + 8\tau C_\beta)$. Furthermore, the condition of α reduces to $\alpha \le c_\alpha$, where c_α is a positive constant depending on $c_{\mathbf{X},\mathrm{l}}$, $c_{\mathbf{X},\mathrm{u}}$, $c_{\mathbf{X},\mathrm{sub}}$ and C_β, because of the choice of T and τ. The proof of Fan et al. (2017, Lemma 2) provides

$$\frac{1}{n}\sum_{i=1}^{n}\varphi_{\tau\|\Delta\|_2}\Big(\mathbf{X_i}^\top\Delta\,\mathbb{1}\{|Y_i - \mathbf{X_i}^\top\beta| \le T\}\Big) \ge c_1\|\Delta\|_2\left(\|\Delta\|_2 - c_2\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}\|\Delta\|_1\right) \tag{3.71}$$

with $c_1 = c_{\mathbf{X},\mathrm{l}}/4$ and $c_2 = 160\,\tau^2 c_{\mathbf{X},\mathrm{sub}}/c_{\mathbf{X},\mathrm{l}}$. Additionally, the proof of Fan et al. (2017, Lemma 4) leads to

$$c_1\|\Delta\|_2\left(\|\Delta\|_2 - c_2\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}\|\Delta\|_1\right) \ge \frac{c_1}{2}\|\Delta\|_2^2 - \frac{c_1c_2^2}{2}\frac{\log(p)}{n}\|\Delta\|_1^2\,. \tag{3.72}$$

All in all the inequalities (3.70) - (3.72) imply the assertion of Lemma 3.6. □

Proof of Lemma 3.7. For $v \in B_S := \big\{v \in \mathbb{R}^p \mid \mathrm{supp}(v) \subseteq S, \|v\|_2 = 1\big\}$ we have that

$$\big(\nabla^2\mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta)\big)\,v = \lim_{t\to 0}\frac{\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha}\big(\beta + t\,v\big) - \nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta)}{t}\,,$$

and hence that

$$v^\top\big(\nabla^2\mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta)\big)v = \lim_{t\to 0}\frac{\big\langle\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha}\big(\beta + t\,v\big) - \nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta), t\,v\big\rangle}{t^2}\,. \tag{3.73}$$

The RSC condition (3.17) implies that for $t \leq 1$ and $v \in B_S$ we obtain

$$\begin{aligned}\left\langle \nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha}\big(\beta + t\,v\big) - \nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta), t\,v \right\rangle &\geq t^2 \left(c_1^{\mathrm{RSC}} \|v\|_2^2 - c_2^{\mathrm{RSC}} \frac{\log(p)}{n} \|v\|_1^2 \right) \\ &\geq t^2 \left(c_1^{\mathrm{RSC}} - c_2^{\mathrm{RSC}} \frac{s\log(p)}{n} \right),\end{aligned}$$

where we used $\|v\|_1 \leq \sqrt{s}\,\|v\|_2$ since $\operatorname{supp}(v) \subseteq S$ and $\|v\|_2 = 1$. Plugging this into (3.73) together with the condition $n \geq c_3^{\mathrm{RSC}} s\log(p)$ gives

$$v^\top \big(\nabla^2 \mathcal{L}^{\mathrm{H}}_{n,\alpha}(\beta)\big) v \geq c_1^{\mathrm{RSC}} - \frac{c_1^{\mathrm{RSC}}}{2} = \frac{c_1^{\mathrm{RSC}}}{2},$$

which is equivalent to the estimate (3.18). □

Proof of Lemma 3.8. By (3.12) and (3.14) we obtain

$$\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\big(\beta^*_{\alpha_n}\big) = -\frac{1}{n}\sum_{i=1}^{n} l'_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top \beta^*_{\alpha_n}\big)\mathbf{X_i}$$

with $\big|l'_{\alpha_n}(x)\big| \leq 2\alpha_n^{-1}$ and $\big|l'_{\alpha_n}(x)\big| \leq 2|x|$ for all $x \in \mathbb{R}$. Furthermore, by (3.66) in the proof of Lemma 3.4 it follows that

$$\begin{aligned}\mathbb{E}\big[|Q|^{ru}\big]^{\frac{1}{r}} &\leq \left(\big(2\tau^2\big)^{\frac{ru}{2}}\, ru\, \Gamma\!\left(\frac{ru}{2}\right) \right)^{\frac{1}{r}} \leq \big(2\tau^2\big)^{\frac{u}{2}} \Big((ru)! \Big)^{\frac{1}{r}} \\ &\leq \big(2\tau^2\big)^{\frac{u}{2}} \Big((u!)^r r^{ru} \Big)^{\frac{1}{r}} = \big(2\tau^2\big)^{\frac{u}{2}} u!\, r^u \qquad (3.74)\end{aligned}$$

for $Q \sim \mathrm{subG}(\tau)$ with $\tau > 0$ and $u, r \in \mathbb{N}$ with $u \geq 2$ and $r/2 \in \mathbb{N}$. In the last inequality we bound the r largest factors of $(ru)!$ by ru, then the next r largest factors by $r(u-1)$ and so on. Now we choose $1 < q_1 \leq q$, where q is given in Assumption 3.1, such that $r_1 = q_1/(q_1 - 1) \in \mathbb{N}$ and r_1 is even. Then we obtain

$$\begin{aligned}\mathbb{E}\left[\Big(l'_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top \beta^*_{\alpha_n}\big) X_{i,k}\Big)^2\right] &\leq 4\,\mathbb{E}\Big[\big(\varepsilon_i - \mathbf{X_i}^\top(\beta^* - \beta^*_{\alpha_n})\big)^2 X_{i,k}^2\Big] \\ &\leq 8\,\mathbb{E}\left[\Big(\varepsilon_i^2 + \big(\mathbf{X_i}^\top(\beta^* - \beta^*_{\alpha_n})\big)^2\Big) X_{i,k}^2\right] \\ &= 8\,\mathbb{E}\Big[\mathbb{E}\big[\varepsilon_i^2\big|\mathbf{X_i}\big]\, X_{i,k}^2 + \big(\mathbf{X_i}^\top(\beta^* - \beta^*_{\alpha_n})\big)^2 X_{i,k}^2\Big] \\ &\leq 8\,\mathbb{E}\left[\Big(1 + \mathbb{E}\big[|\varepsilon_i|^m\big|\mathbf{X_i}\big]\Big) X_{i,k}^2 + \big(\mathbf{X_i}^\top(\beta^* - \beta^*_{\alpha_n})\big)^2 X_{i,k}^2\right] \\ &\leq c_3^{\mathrm{Grad}}\end{aligned}$$

with

$$c_3^{\mathrm{Grad}} = 32\, c_{\mathbf{X},\mathrm{sub}}^2 \Big(1 + (1 + C_{\epsilon,\mathrm{m}})^{\frac{1}{q_1}} r_1^2 + 2^2\, 64\, c_{\mathbf{X},\mathrm{sub}}^2\, C_\beta^2\Big)$$

for $k = 1, \dots, p$. In the last inequality we used the Hölder and Cauchy-Schwarz inequality, (3.74) with $u = 2$ and $r \in \{2, r_1\}$, and the fact that $X_{i,k} \sim \mathrm{subG}(c_{\mathbf{X},\mathrm{sub}})$ and $\mathbf{X_i}^\top(\beta^* - \beta^*_{\alpha_n}) \sim \mathrm{subG}(2C_\beta\, c_{\mathbf{X},\mathrm{sub}})$ by (iii) of Assumption 3.1. Analogously we obtain for higher moments, $u \geq 3$, using $\left|l'_{\alpha_n}(x)\right|^u \leq 4\,(2\alpha_n^{-1})^{u-2}\,x^2$, the estimate

$$\begin{aligned}
\mathbb{E}\Big[\Big|l'_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top\beta^*_{\alpha_n}\big)\,X_{i,k}\Big|^u\Big] &\leq 4\left(\frac{2}{\alpha_n}\right)^{u-2}\mathbb{E}\Big[\big(\varepsilon_i - \mathbf{X_i}^\top(\beta^* - \beta^*_{\alpha_n})\big)^2\,\big|X_{i,k}\big|^u\Big] \\
&\leq 8\left(\frac{2}{\alpha_n}\right)^{u-2}\mathbb{E}\Big[\Big(1 + \mathbb{E}\big[|\varepsilon_i|^m\big|\mathbf{X_i}\big]\Big)\,\big|X_{i,k}\big|^u \\
&\qquad\qquad + \big(\mathbf{X_i}^\top(\beta^* - \beta^*_{\alpha_n})\big)^2\,\big|X_{i,k}\big|^u\Big] \\
&\leq 8\left(\frac{2}{\alpha_n}\right)^{u-2} 2^{\frac{u}{2}}\,c^u_{\mathbf{X},\mathrm{sub}}\,u!\,\Big(1 + (1 + C_{\epsilon,\mathrm{m}})^{\frac{1}{q_1}}\,r_1^u \\
&\qquad\qquad + 2^u\,64\,c^2_{\mathbf{X},\mathrm{sub}}\,C_\beta^2\Big) \\
&= u!\left(\frac{\sqrt{2}\,2\,c_{\mathbf{X},\mathrm{sub}}}{\alpha_n}\right)^{u-2} 16\,c^2_{\mathbf{X},\mathrm{sub}}\Big(1 + (1 + C_{\epsilon,\mathrm{m}})^{\frac{1}{q_1}}\,r_1^u \\
&\qquad\qquad + 2^u\,64\,c^2_{\mathbf{X},\mathrm{sub}}\,C_\beta^2\Big) \\
&\leq \frac{u!}{2}\left(\frac{2\,c_4^{\mathrm{Grad}}}{\alpha_n}\right)^{u-2} c_3^{\mathrm{Grad}}
\end{aligned}$$

with

$$c_4^{\mathrm{Grad}} = \sqrt{2}\,\max(r_1, 2)\,c_{\mathbf{X},\mathrm{sub}}\,.$$

In addition, note that $\mathbb{E}\big[l'_{\alpha_n}(Y_i - \mathbf{X_i}^\top\beta^*_{\alpha_n})\,X_{i,k}\big] = 0$ because of (3.5) and (3.16). Now Bernstein's inequality, cf. Massart (2007, Proposition 2.9), leads to

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n l'_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top\beta^*_{\alpha_n}\big)\,X_{i,k}\right| \geq \left(\frac{2c_3^{\mathrm{Grad}}\,x}{n}\right)^{\frac{1}{2}} + \frac{2c_4^{\mathrm{Grad}}\,x}{\alpha_n\,n}\right) \leq 2\exp(-x)$$

for $x > 0$ since the terms of the sum are independent. Let $x = 3\log(p)$ and $c_1^{\mathrm{Grad}} = \sqrt{96/c_3^{\mathrm{Grad}}}\,c_4^{\mathrm{Grad}}/4$, then by the choice of α_n we get

$$\frac{2c_4^{\mathrm{Grad}}\,x}{\alpha_n\,n} \leq \frac{24c_4^{\mathrm{Grad}}}{c_4^{\mathrm{Grad}}}\left(\frac{c_3^{\mathrm{Grad}}\,\log(p)}{96n}\right)^{\frac{1}{2}} = \left(\frac{2c_3^{\mathrm{Grad}}\,x}{n}\right)^{\frac{1}{2}}$$

and hence

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n l'_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top\beta^*_{\alpha_n}\big)\,X_{i,k}\right| \geq 2\left(\frac{6c_3^{\mathrm{Grad}}\,\log(p)}{n}\right)^{\frac{1}{2}}\right) \leq 2\exp\big(-3\log(p)\big)\,.$$

Union bound implies

$$\mathbb{P}\left(\big\|\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n})\big\|_\infty \geq 2\left(\frac{6c_3^{\mathrm{Grad}}\log(p)}{n}\right)^{\frac{1}{2}}\right) \leq 2\exp\big(-3\log(p)+\log(p)\big) = \frac{2}{p^2}$$

and $c_2^{\mathrm{Grad}} = 2(6c_3^{\mathrm{Grad}})^{\frac{1}{2}}$. □

3.6.2. Proof of Lemma 3.13

The proof of Lemma 3.13 relies on the following two technical results.

Lemma 3.20. *Suppose Assumption 3.1 and* $\alpha_n = C_\alpha(\log(p)/n)^{\frac{1}{2}}$ *for some positive constant* $C_\alpha > 0$ *hold. If in addition* $n \geq \max\big\{(576\log(6)\,C_\alpha\,C_\beta^2\,c^2_{\mathbf{X},\mathrm{sub}})^2 s^2\log(p), 16\log(24)\,s\log(p)\big\}$*, then there exist positive constants* $C_1, C_2, C_3 > 0$ *such that*

$$\begin{aligned}\Big\|\widehat{Q}_{SS} - \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big\|_{\mathrm{M},2} &\leq C_2\max\Bigg\{\left(\frac{s}{n}\right)^{\frac{1}{2}}, \frac{s}{n}, \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}, \left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}}, \\ &\qquad\qquad \alpha_n^{\frac{m}{2}}, \alpha_n^{m-\frac{1}{2}}, \alpha_n\Bigg\} \\ &\leq \frac{C_3}{\sqrt{s}}\end{aligned}$$

with probability at least $1 - C_1/p^2 - 6/p^{5s}$.

Proof of Lemma 3.20. The following proof uses elements of the proof of Lemma 1 in Sun et al. (2020). Let $\mathcal{B}_2^s = \big\{u \in \mathbb{R}^s \mid \|u\|_2 \leq 1\big\}$. Then using (3.21) in Lemma 3.9 we have

$$\begin{aligned}\left\|\frac{2}{n}\sum_{i=1}^n \big(\mathbf{X_i}\big)_S\big(\mathbf{X_i}\big)_S^\top - \widehat{Q}_{SS}\right\|_{\mathrm{M},2} &= \max_{u\in\mathcal{B}_2^s} u^\top\left(\frac{2}{n}\sum_{i=1}^n (1-d_i)\,\big(\mathbf{X_i}\big)_S\big(\mathbf{X_i}\big)_S^\top\right)u \\ &= \max_{u\in\mathcal{B}_2^s}\Big(Z_n^1(u) + Z_n^2(u)\Big) \\ &\leq \max_{u\in\mathcal{B}_2^s} Z_n^1(u) + \max_{u\in\mathcal{B}_2^s} Z_n^2(u) \qquad (3.75)\end{aligned}$$

with

$$\begin{aligned}Z_n^1(u) = \frac{1}{n}\sum_{i=1}^n \Bigg(\int_0^1 &\Big(2 - l''_{\alpha_n}\Big(Y_i - \mathbf{X_i}^\top\big(\beta^*_{\alpha_n} + t\,\big(\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*_{\alpha_n}\big)\big)\Big)\Big) \\ &\cdot \mathbb{1}_{[0,\alpha_n^{-1/2}]}\Big(\Big|Y_i - \mathbf{X_i}^\top\big(\beta^*_{\alpha_n} + t\,\big(\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*_{\alpha_n}\big)\big)\Big|\Big)dt\Bigg)\Big(u^\top\big(\mathbf{X_i}\big)_S\Big)^2,\end{aligned}$$

$$\begin{aligned} Z_n^2(u) = \frac{1}{n}\sum_{i=1}^{n}\Bigg(\int_0^1 \Big(2 - l''_{\alpha_n}\Big(Y_i - \mathbf{X_i}^\top\big(\beta^*_{\alpha_n} + t\,\big(\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*_{\alpha_n}\big)\big)\Big)\Big) \\ \cdot\, \mathbb{1}_{(\alpha_n^{-1/2},\infty)}\Big(\Big|Y_i - \mathbf{X_i}^\top\big(\beta^*_{\alpha_n} + t\,\big(\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*_{\alpha_n}\big)\big)\Big|\Big)\,dt\Bigg)\Big(u^\top \big(\mathbf{X_i}\big)_S\Big)^2 . \end{aligned}$$

To handle the first sum in (3.75) we consider the series expansion in (3.62), which implies

$$\big|2 - l''_{\alpha_n}(x)\big| = \left| -2\sum_{k=2}^{\infty}\binom{1/2}{k} 2k\,(2k-1)\,\alpha_n^{2k-2}\,x^{2k-2}\right| \le \big|3\alpha_n^2\,x^2\big| = 3\alpha_n^2\,x^2\,,$$

if $\alpha_n^2\,x^2 < 1$. Hence for small α_n we get

$$\max_{u\in\mathcal{B}_2^s} Z_n^1(u) \le \max_{u\in\mathcal{B}_2^s} \frac{3\alpha_n}{n}\sum_{i=1}^{n}\Big(u^\top \big(\mathbf{X_i}\big)_S\Big)^2 .$$

Standard spectral norm bounds on the sample covariance matrix (with independent and identically distributed sub-Gaussian rows), cf. Wainwright (2019, Theorem 6.5), and (ii) of Assumption 3.1 lead to

$$\begin{aligned} \max_{u\in\mathcal{B}_2^s} \frac{1}{n}\sum_{i=1}^{n}\Big(u^\top\big(\mathbf{X_i}\big)_S\Big)^2 &= \left\|\frac{1}{n}\sum_{i=1}^{n}\,\big(\mathbf{X_i}\big)_S\big(\mathbf{X_i}\big)_S^\top\right\|_{\mathrm{M},2} \\ &\le \big\|\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\big\|_{\mathrm{M},2} + \left\|\frac{1}{n}\sum_{i=1}^{n}\,\big(\mathbf{X_i}\big)_S\big(\mathbf{X_i}\big)_S^\top - \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right\|_{\mathrm{M},2} \\ &\le c_{\mathbf{X},\mathrm{u}} + C_4\left(\left(\frac{s}{n}\right)^{\frac{1}{2}} + \frac{s}{n} + \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}\right) \qquad (3.76) \end{aligned}$$

with probability at least $1 - C_1/p^2$ for some positive constants $C_1, C_4 > 0$. Hence

$$\max_{u\in\mathcal{B}_2^s} Z_n^1(u) \le 3\big(c_{\mathbf{X},\mathrm{u}} + 3C_4\big)\,\alpha_n \qquad (3.77)$$

with high probability. For the second sum in (3.75) we firstly estimate

$$\begin{aligned} \max_{u\in\mathcal{B}_2^s} Z_n^2(u) \le \max_{u\in\mathcal{B}_2^s} \frac{2}{n}\sum_{i=1}^{n}\Bigg(\int_0^1 \mathbb{1}_{(\alpha_n^{-1/2},\infty)}\Big(\Big|Y_i - \mathbf{X_i}^\top\Big(\beta^*_{\alpha_n} + t\,\big(\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*_{\alpha_n}\big)\Big)\Big|\Big)dt \\ \cdot\Big(u^\top\big(\mathbf{X_i}\big)_S\Big)^2\Bigg) \qquad (3.78) \end{aligned}$$

because of (3.13). Now we can rearrange the term in the indicator function as

$$\Big|Y_i - \mathbf{X_i}^\top\Big(\beta^*_{\alpha_n} + t\,\big(\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*_{\alpha_n}\big)\Big)\Big| = \Big|\varepsilon_i + (1-t)\,\mathbf{X_i}^\top\big(\beta^* - \beta^*_{\alpha_n}\big) + t\,\mathbf{X_i}^\top\big(\beta^* - \widehat{\beta}_n^{\mathrm{PDW}}\big)\Big|.$$

Using the inequality

$$\mathbb{1}_{(\alpha_n^{-1/2},\infty)}\Big(\big|Q_1+Q_2+Q_3\big|\Big) \le \mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\Big(\big|Q_1\big|\Big) + \mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\Big(\big|Q_2\big|\Big) + \mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\Big(\big|Q_3\big|\Big)$$

for random variables Q_1, Q_2 and Q_3 leads to

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\left(\int_0^1 \mathbb{1}_{(\alpha_n^{-1/2},\infty)}\left(\left|Y_i-\mathbf{X_i}^\top\left(\beta^*_{\alpha_n}+t\left(\widehat{\beta}_n^{\mathrm{PDW}}-\beta^*_{\alpha_n}\right)\right)\right|\right)dt\right)\left(u^\top\left(\mathbf{X_i}\right)_S\right)^2\\
&\quad\le \frac{2}{n}\sum_{i=1}^{n}\left(\int_0^1\left[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\big(|\varepsilon_i|\big)+\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|(1-t)\,\mathbf{X_i}^\top\left(\beta^*-\beta^*_{\alpha_n}\right)\right|\right)\right.\right.\\
&\qquad\left.\left.+\,\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|t\,\mathbf{X_i}^\top\left(\beta^*-\widehat{\beta}_n^{\mathrm{PDW}}\right)\right|\right)\right]dt\right)\left(u^\top\left(\mathbf{X_i}\right)_S\right)^2\\
&\quad\le \frac{2}{n}\sum_{i=1}^{n}\left(\int_0^1\left[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\big(|\varepsilon_i|\big)+\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X_i}^\top\left(\beta^*-\beta^*_{\alpha_n}\right)\right|\right)\right.\right.\\
&\qquad\left.\left.+\,\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X_i}^\top\left(\beta^*-\widehat{\beta}_n^{\mathrm{PDW}}\right)\right|\right)\right]dt\right)\left(u^\top\left(\mathbf{X_i}\right)_S\right)^2\\
&\quad= \frac{2}{n}\sum_{i=1}^{n}\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\big(|\varepsilon_i|\big)\left(u^\top\left(\mathbf{X_i}\right)_S\right)^2\\
&\qquad+\frac{2}{n}\sum_{i=1}^{n}\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X_i}^\top\left(\beta^*-\beta^*_{\alpha_n}\right)\right|\right)\left(u^\top\left(\mathbf{X_i}\right)_S\right)^2\\
&\qquad+\frac{2}{n}\sum_{i=1}^{n}\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X_i}^\top\left(\beta^*-\widehat{\beta}_n^{\mathrm{PDW}}\right)\right|\right)\left(u^\top\left(\mathbf{X_i}\right)_S\right)^2.
\end{aligned} \tag{3.79}$$

We consider each of the three terms separately. By (iii) of Assumption 3.1 we get for fixed $u\in\mathcal{B}_2^s$ that $u^\top\left(\mathbf{X_i}\right)_S\sim\mathrm{subG}(c_{\mathbf{X},\mathrm{sub}})$, and following the proof of Rigollet and Hütter (2019, Lemma 1.12) together with $\left(\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}(|\varepsilon_i|)\right)^2=\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}(|\varepsilon_i|)$ leads to

$$\begin{aligned}
Q_i(u)&=\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\big(|\varepsilon_i|\big)\left(u^\top\left(\mathbf{X_i}\right)_S\right)^2-\mathbb{E}\left[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\big(|\varepsilon_i|\big)\left(u^\top\left(\mathbf{X_i}\right)_S\right)^2\right]\\
&\sim\mathrm{subE}\big(16\,c^2_{\mathbf{X},\mathrm{sub}},16\,c^2_{\mathbf{X},\mathrm{sub}}\big)\,.
\end{aligned}$$

Bernstein's inequality, cf. Rigollet and Hütter (2019, Theorem 1.13), implies

$$\mathbb{P}\left(\left|\frac{2}{n}\sum_{i=1}^{n}Q_i(u)\right|>x\right)\le 2\max\left\{\exp\left(-\frac{x^2\,n}{2048\,c^4_{\mathbf{X},\mathrm{sub}}}\right),\exp\left(-\frac{x\,n}{64\,c^2_{\mathbf{X},\mathrm{sub}}}\right)\right\}$$

for $x > 0$ and fixed $u \in \mathcal{B}_2^s$. Now we proceed with a covering argument. Consider a 1/8-cover A of cardinality $N = N(1/8; \mathcal{B}_2^s, \|\cdot\|_2) \leq 24^s$ of the unit Euclidean ball of $\mathbb{R}^s$ with respect to the Euclidean distance, cf. in Rigollet and Hütter (2019, Lemma 1.18) or Wainwright (2019, Example 5.8). We can argue similarly to the proof in Wainwright (2019, Theorem 6.5) since we consider also a quadratic form, and obtain for $x = 256\,\sqrt{\log(24)}\,c_{\mathbf{X},\mathrm{sub}}^2\,(s\log(p)/n)^{\frac{1}{2}}$ that

$$\begin{aligned}
\mathbb{P}\Bigg(\max_{u\in\mathcal{B}_2^s}\left|\frac{2}{n}\sum_{i=1}^n Q_i(u)\right| > x\Bigg) &\leq \mathbb{P}\Bigg(\max_{u\in A}\left|\frac{2}{n}\sum_{i=1}^n Q_i(u)\right| > \frac{x}{2}\Bigg)\\
&\leq 2\,\big|N\big|\,\max\Bigg\{\exp\big(-8\log(24)\,s\log(p)\big)\,,\\
&\qquad\qquad\exp\Big(-\big(4\log(24)\,s\log(p)\,n\big)^{\frac{1}{2}}\Big)\Bigg\}\\
&\leq 2\exp\big(\log(24)\,s - 8\log(24)\,s\log(p)\big)\\
&\leq 2\exp\big(-4\log(24)\,s\log(p)\big)\\
&\leq \frac{2}{p^{5s}} \qquad (3.80)
\end{aligned}$$

since $4\log(p) \geq 1$ if $p \geq 2$, and by assumption $n \geq 16\log(24)\,s\log(p)$. In addition, we obtain

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^n \mathbb{E}\Big[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\big(|\varepsilon_i|\big)\,\Big(u^\top(\mathbf{X_i})_S\Big)^2\Big] &= 2\,\mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\big(|\varepsilon_1|\big)\Big|\mathbf{X_1}\Big]\Big(u^\top(\mathbf{X_1})_S\Big)^2\Big]\\
&\leq 2\,\big(1+C_{\epsilon,\mathrm{m}}\big)\,\big(9\alpha_n\big)^{\frac{m}{2}}\,\mathbb{E}\Big[\Big(u^\top(\mathbf{X_1})_S\Big)^2\Big]\\
&\leq 2\,\big(1+C_{\epsilon,\mathrm{m}}\big)\,c_{\mathbf{X},\mathrm{sub}}^2\,\big(9\alpha_n\big)^{\frac{m}{2}}
\end{aligned}$$

by Assumption 3.1 and an application of the conditional version of Markov's inequality,

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\big(|\varepsilon_1|\big)\Big|\mathbf{X_1}\Big] = \mathbb{P}\Bigg(|\varepsilon_1| > \frac{1}{3\alpha_n^{\frac{1}{2}}}\Bigg|\mathbf{X_1}\Bigg) &\leq (9\alpha_n)^{\frac{m}{2}}\,\mathbb{E}\Big[\mathbb{E}\big[|\varepsilon_1|^m\big|\mathbf{X_1}\big]\Big]\\
&\leq (9\alpha_n)^{\frac{m}{2}}\,\Big(1+\mathbb{E}\Big[\mathbb{E}\big[|\varepsilon_1|^m\big|\mathbf{X_1}\big]^q\Big]\Big)\\
&\leq \big(1+C_{\epsilon,\mathrm{m}}\big)\,\big(9\alpha_n\big)^{\frac{m}{2}}\,.
\end{aligned}$$

By building the maximum of the expected values over $u \in \mathcal{B}_2^s$ and collecting terms we find that

$$\begin{aligned}
\max_{u\in\mathcal{B}_2^s}\frac{2}{n}\sum_{i=1}^n \mathbb{1}_{(\alpha_n^{-1}/3,\infty)}\big(|\varepsilon_i|\big)\,\Big(u^\top(\mathbf{X_i})_S\Big)^2 &\leq 256\,\sqrt{\log(24)}\,c_{\mathbf{X},\mathrm{sub}}^2\left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}}\\
&\quad + 2\,\big(1+C_{\epsilon,\mathrm{m}}\big)\,c_{\mathbf{X},\mathrm{sub}}^2\,\big(9\alpha_n\big)^{\frac{m}{2}} \qquad (3.81)
\end{aligned}$$

with probability at least $1 - 2/p^{5s}$. We proceed similarly for the second and third sum in (3.79), hence it is sufficient to consider the rates of the expected values

$$\mathbb{E}\left[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X_1}^\top(\beta^* - \beta)\right|\right)\left(u^\top(\mathbf{X_1})_S\right)^2\right]$$

with $\beta = \beta^*_{\alpha_n}$ and $\beta = \widehat{\beta}_n^{\mathrm{PDW}}$. Obviously it is

$$\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\right|\right) \leq 3\left|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\right|\alpha_n^{\frac{1}{2}}$$

and hence by Assumption 3.1, Rigollet and Hütter (2019, Lemma 1.4) and the Cauchy-Schwarz inequality

$$\begin{aligned}\mathbb{E}&\left[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\right|\right)\left(u^\top(\mathbf{X_1})_S\right)^2\right] \\ &\leq 3\alpha_n^{\frac{1}{2}}\,\mathbb{E}\left[\left|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\right|\left(u^\top(\mathbf{X_1})_S\right)^2\right] \\ &\leq 3\alpha_n^{\frac{1}{2}}\left(\mathbb{E}\left[\left(\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\right)^2\right]\mathbb{E}\left[\left(u^\top(\mathbf{X_1})_S\right)^4\right]\right)^{\frac{1}{2}} \\ &\leq 12\, c^3_{\mathbf{X},\mathrm{sub}}\left\|\beta^* - \beta^*_{\alpha_n}\right\|_2\alpha_n^{\frac{1}{2}}\,.\end{aligned}$$

Lemma 3.4 implies

$$\mathbb{E}\left[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X_1}^\top(\beta^* - \beta^*_{\alpha_n})\right|\right)\left(u^\top(\mathbf{X_1})_S\right)^2\right] \leq 12\, C_{\mathrm{apx}}\, c^3_{\mathbf{X},\mathrm{sub}}\,\alpha_n^{m-\frac{1}{2}}\,. \tag{3.82}$$

The vector $\widehat{\beta}_n^{\mathrm{PDW}}$ has support S and satisfies $\left\|\widehat{\beta}_n^{\mathrm{PDW}} - \beta^*\right\|_2 \leq 2C_\beta$ by (3.19) and (iv) of Assumption 3.1, hence it follows that

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X_1}^\top(\beta^* - \widehat{\beta}_n^{\mathrm{PDW}})\right|\right)\right] &= \mathbb{P}\left(\left|\mathbf{X_1}^\top(\beta^* - \widehat{\beta}_n^{\mathrm{PDW}})\right| > \frac{1}{3\alpha_n^{\frac{1}{2}}}\right) \\ &\leq \mathbb{P}\left(\max_{u\in\mathbb{R}^s:\|u\|_2\leq 2C_\beta}\left|u^\top(\mathbf{X_1})_S\right| > \frac{1}{3\alpha_n^{\frac{1}{2}}}\right) \\ &= \mathbb{P}\left(\max_{u\in\mathcal{B}_2^s}\left|u^\top\left(2C_\beta(\mathbf{X_1})_S\right)\right| > \frac{1}{3\alpha_n^{\frac{1}{2}}}\right) \\ &\leq \exp\left(\log(6)\,s - \frac{1}{288\,C_\beta^2\,c^2_{\mathbf{X},\mathrm{sub}}\,\alpha_n}\right)\end{aligned}$$

by Rigollet and Hütter (2019, Theorem 1.19) together with Assumption 3.1. By the choice of α_n and the sample size n we obtain

$$\begin{aligned}\exp\left(\log(6)\, s - \frac{1}{288\, C_\beta^2\, c_{\mathbf{X},\mathrm{sub}}^2\, \alpha_n}\right) &= \exp\left(\log(6)\, s - \frac{\sqrt{n}}{576\, C_\beta^2\, c_{\mathbf{X},\mathrm{sub}}^2\, C_\alpha\, \sqrt{\log(p)}}\right.\\ &\qquad\qquad \left. - \frac{1}{576\, C_\beta^2\, c_{\mathbf{X},\mathrm{sub}}^2\, \alpha_n}\right)\\ &\leq \exp\big(\log(6)\, s - \log(6)\, s\big)\, \exp\left(-\frac{\alpha_n^{-1}}{576\, C_\beta^2\, c_{\mathbf{X},\mathrm{sub}}^2}\right)\\ &= 2\, \big(576\, C_\beta^2\, c_{\mathbf{X},\mathrm{sub}}^2\, \alpha_n\big)^2\end{aligned}$$

since $\exp(x) \geq x^2/2$ for $x > 0$. Therefore

$$\mathbb{E}\left[\mathbb{1}_{(\alpha_n^{-1/2}/3,\infty)}\left(\left|\mathbf{X}_{\mathbf{1}}^\top\big(\beta^* - \widehat{\beta}_n^{\mathrm{PDW}}\big)\right|\right)\left(u^\top \big(\mathbf{X}_{\mathbf{1}}\big)_S\right)^2\right] \leq \sqrt{2}\, 2304\, C_\beta^2\, c_{\mathbf{X},\mathrm{sub}}^4\, \alpha_n \tag{3.83}$$

by the Cauchy-Schwarz inequality. So finally the previous considerations in (3.75) - (3.83) showed that

$$\begin{aligned}\left\|\frac{2}{n}\sum_{i=1}^{n}\big(\mathbf{X}_{\mathbf{i}}\big)_S\big(\mathbf{X}_{\mathbf{i}}\big)_S^\top - \widehat{Q}_{SS}\right\|_{\mathrm{M},2} &\leq 768\, \sqrt{\log(24)}\, c_{\mathbf{X},\mathrm{sub}}^2 \left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}}\\ &\quad + 2\, \big(1 + C_{\epsilon,\mathrm{m}}\big)\, c_{\mathbf{X},\mathrm{sub}}^2\, \big(9\alpha_n\big)^{\frac{m}{2}} + 24\, C_{\mathrm{apx}}\, c_{\mathbf{X},\mathrm{sub}}^3\, \alpha_n^{m-\frac{1}{2}}\\ &\quad + \Big(3\big(c_{\mathbf{X},\mathrm{u}} + 3C_4\big) + \sqrt{2}\, 4608\, C_\beta^2\, c_{\mathbf{X},\mathrm{sub}}^4\Big)\alpha_n\\ &\leq C_5\, \max\left\{\left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}},\, \alpha_n^{\frac{m}{2}},\, \alpha_n^{m-\frac{1}{2}},\, \alpha_n\right\}\end{aligned}$$

for a positive constant $C_5 > 0$ with probability at least $1 - C_1/p^2 - 6/p^{5s}$. Furthermore, repeated application of the spectral norm bound in (3.76) leads to

$$\begin{aligned}\left\|\widehat{Q}_{SS} - \mathbb{E}\big[\mathbf{X}_{\mathbf{1}}\mathbf{X}_{\mathbf{1}}^\top\big]_{SS}\right\|_{\mathrm{M},2} &\leq \left\|\widehat{Q}_{SS} - \frac{2}{n}\sum_{i=1}^{n}\big(\mathbf{X}_{\mathbf{i}}\big)_S\big(\mathbf{X}_{\mathbf{i}}\big)_S^\top\right\|_{\mathrm{M},2}\\ &\quad + 2\left\|\frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{X}_{\mathbf{i}}\big)_S\big(\mathbf{X}_{\mathbf{i}}\big)_S^\top - \mathbb{E}\big[\mathbf{X}_{\mathbf{1}}\mathbf{X}_{\mathbf{1}}^\top\big]_{SS}\right\|_{\mathrm{M},2}\\ &\leq C_2\, \max\left\{\left(\frac{s}{n}\right)^{\frac{1}{2}},\, \frac{s}{n},\, \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}},\, \left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}},\right.\\ &\qquad\qquad \left.\alpha_n^{\frac{m}{2}},\, \alpha_n^{m-\frac{1}{2}},\, \alpha_n\right\}\end{aligned}$$

for a positive constant $C_2 > 0$. By the choices of α_n and $n \gtrsim s^2\log(p)$ together with

$m \in \{2,3\}$, finally, it follows that

$$\left\| \widehat{Q}_{SS} - \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS} \right\|_{\mathrm{M},2} \leq C_6 \left(\frac{s\log(p)}{n}\right)^{\frac{1}{2}} \leq \frac{C_3}{\sqrt{s}}$$

for some positive constants $C_3, C_6 > 0$ with probability at least $1 - C_1/p^2 - 6/p^{5s}$. □

Lemma 3.21. *Let $M \in \mathbb{R}^{|A|\times|B|}$ be a matrix with $A, B \subseteq \{1,\dots,p\}$ and $\max_{k\in\{1,\dots,|A|\}} \big\|M^\top e_k\big\|_2 \leq C_M$ for some positive constant $C_M > 0$. Suppose Assumption 3.1 and $\alpha_n \geq c_1^{\mathrm{Grad}} \,(\log(p)/n)^{\frac{1}{2}}$ holds, then with probability at least $1 - 2/p^2$ the ℓ_∞ norm of $M\big(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n})\big)_B$ is bounded by*

$$\left\| M\Big(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n})\Big)_B \right\|_\infty \leq C_M\, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}.$$

Proof of Lemma 3.21. We follow the proof of Lemma 3.8. It is

$$\begin{aligned} M\Big(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n})\Big)_B &= M\bigg(-\frac{1}{n}\sum_{i=1}^n l'_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top\beta^*_{\alpha_n}\big)\big(\mathbf{X_i}\big)_B\bigg) \\ &= -\frac{1}{n}\sum_{i=1}^n l'_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top\beta^*_{\alpha_n}\big)\mathbf{Z_i} \end{aligned}$$

with $\mathbf{Z_i} = M\big(\mathbf{X_i}\big)_B$. The random vectors $l'_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top\beta^*_{\alpha_n}\big)\mathbf{Z_1}, \dots, l'_{\alpha_n}\big(Y_n - \mathbf{X_n}^\top\beta^*_{\alpha_n}\big)\mathbf{Z_n}$ are independent and identically distributed because $(\mathbf{X_1},\varepsilon_1),\dots,(\mathbf{X_n},\varepsilon_n)$ are independent and identically distributed. In addition (iii) of Assumption 3.1 and $\max_{k\in\{1,\dots,|A|\}} \big\|M^\top e_k\big\|_2 \leq C_M$ imply that the entries $Z_{i,k} = e_k^\top\mathbf{Z_i}$ of $\mathbf{Z_i}$ are sub-Gaussian with variance proxy $C_M^2\, c^2_{\mathbf{X},\mathrm{sub}}$. This leads to

$$\mathbb{E}\bigg[\Big(l'_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top\beta^*_{\alpha_n}\big)Z_{i,k}\Big)^2\bigg] \leq C_M^2\, c_3^{\mathrm{Grad}}$$

and

$$\mathbb{E}\bigg[\Big|l'_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top\beta^*_{\alpha_n}\big)\, Z_{i,k}\Big|^u\bigg] \leq \frac{u!}{2}\left(\frac{2\, C_M\, c_4^{\mathrm{Grad}}}{\alpha_n}\right)^{u-2} c_3^{\mathrm{Grad}}$$

for $u \in \mathbb{N}$, $u \geq 3$, where c_3^{Grad} and c_4^{Grad} are given in the proof of Lemma 3.8. Moreover, we obtain

$$\mathbb{E}\Big[l'_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top\beta^*_{\alpha_n}\big)\mathbf{Z_1}\Big] = M\,\mathbb{E}\Big[l'_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top\beta^*_{\alpha_n}\big)\big(\mathbf{X_1}\big)_B\Big] = \mathbf{0}_{|A|}$$

since $\mathbb{E}\big[l'_{\alpha_n}\big(Y_1 - \mathbf{X_1}^\top\beta^*_{\alpha_n}\big)\mathbf{X_1}\big] = \mathbf{0}_p$ (see proof of Lemma 3.8). Arguing as in the proof of Lemma 3.8 concludes the proof. □

Proof of Lemma 3.13. For the first part we invoke Lemma 3.20 and obtain (if $C_3 \geq \max\big\{(576\,\log(6)\,C_\alpha\,C_\beta^2\,c_{\mathbf{X},\mathrm{sub}}^2)^2, 16\log(24)\big\}$ in Lemma 3.13)

$$\Big\|\widehat{Q}_{SS} - \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big\|_{\mathrm{M},2} \leq \frac{C_4}{\sqrt{s}}$$

with probability at least $1 - C_1/p^2 - 6/p^{5s}$ for some positive constants $C_1, C_4 > 0$. Moreover, we have

$$\Big\|\Big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big)^{-1}\Big\|_{\mathrm{M},2} \leq \Big\|\Big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big)^{-1}\Big\|_{\mathrm{M},\infty} \leq C_{\mathrm{S},\mathbf{X}} \tag{3.84}$$

by (3.8) and the symmetry of the matrix. Hence by Loh and Wainwright (2017, Lemma 11) we conclude that

$$\Big\|\big(\widehat{Q}_{SS}\big)^{-1} - \Big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big)^{-1}\Big\|_{\mathrm{M},2} \leq 2\,C_{\mathrm{S},\mathbf{X}}^2\,\Big\|\widehat{Q}_{SS} - \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big\|_{\mathrm{M},2} \leq \frac{2\,C_4\,C_{\mathrm{S},\mathbf{X}}^2}{\sqrt{s}} \tag{3.85}$$

with high probability if $\sqrt{s} \geq 2\,C_4\,C_{\mathrm{S},\mathbf{X}}$. Finally, the triangle inequality and once again (3.8) lead to

$$\begin{aligned}
\Big\|\big(\widehat{Q}_{SS}\big)^{-1}\Big\|_{\mathrm{M},\infty} &\leq \Big\|\Big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big)^{-1}\Big\|_{\mathrm{M},\infty} + \Big\|\big(\widehat{Q}_{SS}\big)^{-1} - \Big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big)^{-1}\Big\|_{\mathrm{M},\infty} \\
&\leq C_{\mathrm{S},\mathbf{X}} + \sqrt{s}\,\Big\|\big(\widehat{Q}_{SS}\big)^{-1} - \Big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big)^{-1}\Big\|_{\mathrm{M},2} \\
&\leq C_{\mathrm{S},\mathbf{X}} + 2\,C_4\,C_{\mathrm{S},\mathbf{X}}^2
\end{aligned}$$

with probability at least $1 - C_1/p^2 - 6/p^{5s}$.

To prove the second part of this lemma we follow the inequalities

$$\begin{aligned}
&\Big\|\widehat{Q}_{S^cS}\,\big(\widehat{Q}_{SS}\big)^{-1}\Big(\nabla\mathcal{L}_{n,\alpha_n}^{\mathrm{H}}\big(\beta_{\alpha_n}^*\big)\Big)_S\Big\|_\infty \\
&\quad\leq \Big\|\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{S^cS}\,\Big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big)^{-1}\Big(\nabla\mathcal{L}_{n,\alpha_n}^{\mathrm{H}}\big(\beta_{\alpha_n}^*\big)\Big)_S\Big\|_\infty \\
&\qquad + \Big\|\Big(\widehat{Q}_{S^cS}\,\big(\widehat{Q}_{SS}\big)^{-1} - \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{S^cS}\,\Big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\Big)^{-1}\Big) \\
&\qquad\qquad \cdot\Big(\nabla\mathcal{L}_{n,\alpha_n}^{\mathrm{H}}\big(\beta_{\alpha_n}^*\big)\Big)_S\Big\|_\infty
\end{aligned} \tag{3.86}$$

and

$$\begin{aligned}
&\left\| \left(\widehat{Q}_{S^cS} \left(\widehat{Q}_{SS}\right)^{-1} - \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{S^cS} \left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1} \right) \left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S \right\|_\infty \\
&\quad \le \max_{k\in\{1,\dots,p-s\}} \left\| \left(e_k^\top \left(\widehat{Q}_{S^cS} \left(\widehat{Q}_{SS}\right)^{-1} - \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{S^cS} \left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1} \right) \right)^\top \right\|_2 \\
&\qquad\qquad\qquad\qquad \cdot \left\| \left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S \right\|_2 \\
&\quad \le \max_{k\in\{1,\dots,p-s\}} \left(\left\| \left(e_k^\top\, \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{S^cS}\, \Delta_1 \right)^\top \right\|_2 + \left\| \left(e_k^\top\, \Delta_2^\top \left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1} \right)^\top \right\|_2 \right. \\
&\qquad\qquad\qquad \left. + \left\| \left(e_k^\top\, \Delta_2^\top\, \Delta_1 \right)^\top \right\|_2 \right) \left\| \left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S \right\|_2 \\
&\quad \le \max_{k\in\{1,\dots,p-s\}} \left(\|\Delta_1\|_{\mathrm{M},2} \left\| \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS^c}\, e_k \right\|_2 + \left\| \left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1} \right\|_{\mathrm{M},2} \|\Delta_2\, e_k\|_2 \right. \\
&\qquad\qquad\qquad \left. + \|\Delta_1\|_{\mathrm{M},2} \|\Delta_2\, e_k\|_2 \right) \left\| \left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S \right\|_2
\end{aligned} \tag{3.87}$$

with

$$\Delta_1 = \left(\widehat{Q}_{SS}\right)^{-1} - \left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1} \quad \text{and} \quad \Delta_2 = \widehat{Q}_{SS^c} - \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS^c}$$

in Loh and Wainwright (2017, Corollary 3). Note that (3.85) implies $\|\Delta_1\|_{\mathrm{M},2} \le 2\,C_4\,C^2_{\mathrm{S},\mathbf{X}}/\sqrt{s}$. For the first term in (3.86) we shall apply Lemma 3.21 with $M = \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{S^cS} \left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1}$. We obtain

$$\begin{aligned}
\max_{k\in\{1,\dots,p-s\}} \left\| \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS^c}\, e_k \right\|_2 &\le \max_{k\in S^c} \left\| \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]\, e_k \right\|_2 \le \max_{u\in\mathbb{R}^p,\, \|u\|_2=1} \left\| \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]\, u \right\|_2 \\
&\le c_{\mathbf{X},\mathrm{u}}
\end{aligned}$$

by (ii) of Assumption 3.1, and hence together with (3.84) the estimate

$$\begin{aligned}
&\max_{k\in\{1,\dots,p-s\}} \left\| \left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1} \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS^c}\, e_k \right\|_2 \\
&\qquad \le \max_{k\in\{1,\dots,p-s\}} \left\| \left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1} \right\|_{\mathrm{M},2} \left\| \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS^c}\, e_k \right\|_2 \\
&\qquad \le C_{\mathrm{S},\mathbf{X}}\, c_{\mathbf{X},\mathrm{u}}\,.
\end{aligned}$$

Lemma 3.21 and the choice of α_n in (3.9) lead to

$$\left\| \mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{S^cS} \left(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{SS}\right)^{-1} \left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S \right\|_\infty \le C_{\mathrm{S},\mathbf{X}}\, c_{\mathbf{X},\mathrm{u}}\, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \tag{3.88}$$

with probability at least $1 - 2/p^2$. In addition, we get by Lemma 3.8 also

$$\left\| \left(\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n}) \right)_S \right\|_2 \leq \sqrt{s} \left\| \left(\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n}) \right)_S \right\|_\infty \leq c_2^{\mathrm{Grad}} \left(\frac{s \log(p)}{n} \right)^{\frac{1}{2}} \tag{3.89}$$

with the same probability. The final task is now to study the rate of $\max_{k \in \{1,\dots,p-s\}} \|\Delta_2 \, e_k\|_2$. First of all it is

$$\begin{aligned} \max_{k \in \{1,\dots,p-s\}} & \left\| \left(\widehat{Q}_{SS^c} - \mathbb{E}\big[\mathbf{X_1} \mathbf{X_1}^\top\big]_{SS^c} \right) e_k \right\|_2 \\ & \leq \sqrt{s} \max_{k \in \{1,\dots,p-s\}} \left\| \left(\widehat{Q}_{SS^c} - \mathbb{E}\big[\mathbf{X_1} \mathbf{X_1}^\top\big]_{SS^c} \right) e_k \right\|_\infty \\ & = \sqrt{s} \max_{\substack{l \in \{1,\dots,s\}, \\ k \in \{1,\dots,p-s\}}} \left| e_l^\top \left(\widehat{Q}_{SS^c} - \mathbb{E}\big[\mathbf{X_1} \mathbf{X_1}^\top\big]_{SS^c} \right) e_k \right| \\ & \leq \sqrt{s} \max_{k,l \in \{1,\dots,p\}} \left| e_l^\top \left(\widehat{Q} - \mathbb{E}\big[\mathbf{X_1} \mathbf{X_1}^\top\big] \right) e_k \right| . \end{aligned}$$

We proceed similarly to the proof of Lemma 3.20 but here we have only the maximum over p^2 elements in comparison to the 24^s elements in the mentioned proof. In addition we use the fact that the centered product of two sub-Gaussian random variables is sub-Exponential, cf. Vershynin (2018, Lemma 2.7.7), and that also the centered product of two sub-Gaussian random variables and a bounded random variable is sub-Exponential. Hence we do not have the rates depending on s in (3.76) and in (3.80) the factor s can be dropped. It follows that there exist positive constants $C_2, C_5, C_6 > 0$ such that

$$\max_{k,l \in \{1,\dots,p\}} \left| e_l^\top \left(\widehat{Q} - \mathbb{E}\big[\mathbf{X_1} \mathbf{X_1}^\top\big] \right) e_k \right| \leq C_5 \max \left\{ \left(\frac{\log(p)}{n} \right)^{\frac{1}{2}}, \alpha_n^{\frac{m}{2}}, \alpha_n^{m - \frac{1}{2}}, \alpha_n \right\} \leq \frac{C_6}{s}$$

with probability at least $1 - C_2/p^2$ by the choices of α_n in (3.9) and $n \gtrsim s^2 \log(p)$ together with $m \in \{2, 3\}$. Hence

$$\max_{k \in \{1,\dots,p-s\}} \|\Delta_2 \, e_k\|_2 \leq \frac{C_6}{\sqrt{s}} \tag{3.90}$$

with high probability and in total we obtain by (3.86) - (3.90) the inequality

$$\begin{aligned} \left\| \widehat{Q}_{S^c S} \left(\widehat{Q}_{SS} \right)^{-1} \left(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}(\beta^*_{\alpha_n}) \right)_S \right\|_\infty & \leq C_{\mathrm{S},\mathbf{X}} \, c_{\mathbf{X},\mathrm{u}} \, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n} \right)^{\frac{1}{2}} + c_2^{\mathrm{Grad}} \left(\frac{s \log(p)}{n} \right)^{\frac{1}{2}} \\ & \quad \cdot \left(\frac{2 \, C_4 \, c_{\mathbf{X},\mathrm{u}} \, C_{\mathrm{S},\mathbf{X}}^2}{\sqrt{s}} + \frac{C_6 \, C_{\mathrm{S},\mathbf{X}}}{\sqrt{s}} + \frac{2 \, C_4 \, C_6 \, C_{\mathrm{S},\mathbf{X}}^2}{s} \right) \\ & \leq C_7 \, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n} \right)^{\frac{1}{2}} \end{aligned}$$

with probability at least $1-(4+C_1+C_2)/p^2-6/p^{5s}$ for some positive constant $C_7>0$. Renewed application of Lemma 3.8 and the triangular inequality lead to

$$\left\|\widehat{Q}_{S^cS}\left(\widehat{Q}_{SS}\right)^{-1}\left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S-\left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_{S^c}\right\|_\infty\le\left(1+C_7\right)c_2^{\mathrm{Grad}}\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}.$$

□

3.6.3. Proofs of Lemmas 3.14 and 3.18

We start with proving Lemma 3.18. For this purpose we need a technical result concerning the column normalization of the design matrix $\mathbb{X}_n$.

Lemma 3.22. *Let $\mathbb{X}_n=\left(\mathbf{X_1},\dots,\mathbf{X_n}\right)^\top\in\mathbb{R}^{n\times p}$ be a matrix with independent and identically distributed rows $\mathbf{X_i}\sim\mathrm{subG}_p(c_{\mathbf{X},\mathrm{sub}})$ with variance proxy $c^2_{\mathbf{X},\mathrm{sub}}>0$. Then for $n\ge 6\log(p)$ the columns $\vec{X}_k$ of $\mathbb{X}_n$ satisfy with probability at least $1-2/p^2$*

$$\frac{1}{n}\max_{k\in\{1,\dots,p\}}\left\|\vec{X}_k\right\|_2^2\le 17\,c^2_{\mathbf{X},\mathrm{sub}}\,.$$

Proof of Lemma 3.22. We have $X_{i,k}=e_k^\top\,\mathbf{X_i}\sim\mathrm{subG}(c_{\mathbf{X},\mathrm{sub}})$ for all $i=1,\dots,n$ and $k=1,\dots,p$ by the definition of a sub-Gaussian random vector. Rigollet and Hütter (2019, Lemma 1.12) implies $X^2_{i,k}-\mathbb{E}\left[X^2_{i,k}\right]\sim\mathrm{subE}(16\,c^2_{\mathbf{X},\mathrm{sub}},16\,c^2_{\mathbf{X},\mathrm{sub}})$ and with Bernstein's inequality, cf. Rigollet and Hütter (2019, Theorem 1.13), it follows that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\left(X^2_{i,k}-\mathbb{E}\left[X^2_{i,k}\right]\right)\right|>x\right)\le 2\max\left\{\exp\left(-\frac{x^2\,n}{512\,c^4_{\mathbf{X},\mathrm{sub}}}\right),\exp\left(-\frac{x\,n}{32\,c^2_{\mathbf{X},\mathrm{sub}}}\right)\right\}$$

for all $x>0$ and $k=1,\dots,p$ since $X_{1,k},\dots,X_{n,k}$ are independent and identically distributed. By the union bound and the condition $n\ge 6\log(p)$ we obtain

$$\mathbb{P}\left(\max_{k\in\{1,\dots,p\}}\left|\frac{1}{n}\sum_{i=1}^n\left(\left(e_k^\top\,\mathbf{X_i}\right)^2-\mathbb{E}\left[\left(e_k^\top\,\mathbf{X_i}\right)^2\right]\right)\right|>16\,c^2_{\mathbf{X},\mathrm{sub}}\right)\le 2\,p\exp\left(-\frac{n}{2}\right)\le\frac{2}{p^2}\,.$$

Furthermore, we have for all $k=1,\dots,p$ the estimate

$$\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[X^2_{i,k}\right]=\mathbb{E}\left[X^2_{1,k}\right]\le c^2_{\mathbf{X},\mathrm{sub}}$$

since $X_{1,k}$ is sub-Gaussian with variance proxy $c^2_{\mathbf{X},\mathrm{sub}}$, and therefore we get

$$\begin{aligned}\max_{k\in\{1,\dots,p\}}\frac{1}{n}\left\|\vec{X}_k\right\|_2^2&\le\max_{k\in\{1,\dots,p\}}\left|\frac{1}{n}\sum_{i=1}^n\left(X^2_{i,k}-\mathbb{E}\left[X^2_{i,k}\right]\right)\right|+\max_{k\in\{1,\dots,p\}}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[X^2_{i,k}\right]\\&\le 16\,c^2_{\mathbf{X},\mathrm{sub}}+c^2_{\mathbf{X},\mathrm{sub}}=17\,c^2_{\mathbf{X},\mathrm{sub}}\end{aligned}$$

with high probability. □

Proof of Lemma 3.18. We follow the proof of Zhou et al. (2009, Lemma 10.3). It is

$$\widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1} = \frac{2}{n}\,\mathbb{X}_{n,S^c}^\top\, D\,\mathbb{X}_{n,S}\left(\frac{2}{n}\,\mathbb{X}_{n,S}^\top\, D\,\mathbb{X}_{n,S}\right)^{-1} = \mathbb{X}_{n,S^c}^\top\, D\,\mathbb{X}_{n,S}\big(\mathbb{X}_{n,S}^\top\, D\,\mathbb{X}_{n,S}\big)^{-1},$$

see Lemma 3.9. For $k \in S^c$ let

$$r_k = \big(\mathbb{X}_{n,S}^\top\, D\,\mathbb{X}_{n,S}\big)^{-1}\mathbb{X}_{n,S}^\top\, D\,\vec{X}_k \qquad \in \mathbb{R}^s\,,$$

then we have

$$\left\|\widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\right\|_{\mathrm{M},\infty} = \max_{k\in S^c}\|r_k\|_1\,.$$

Furthermore, on the one hand the column normalization in Lemma 3.22 under the condition $n \geq 6\log(p)$ and the submultiplicativity of the spectral norm lead to

$$\begin{aligned}\max_{k\in S^c}\left\|D^{\frac{1}{2}}\,\mathbb{X}_{n,S}\,r_k\right\|_2 &\leq \max_{k\in S^c}\left(\left\|D^{\frac{1}{2}}\,\mathbb{X}_{n,S}\left(\mathbb{X}_{n,S}^\top\, D\,\mathbb{X}_{n,S}\right)^{-1}\mathbb{X}_{n,S}^\top\, D^{\frac{1}{2}}\right\|_{\mathrm{M},2}\left\|D^{\frac{1}{2}}\right\|_{\mathrm{M},2}\left\|\vec{X}_k\right\|_2\right)\\ &\leq \max_{k\in S^c}\left\|\vec{X}_k\right\|_2 \leq \sqrt{17}\,c_{\mathbf{X},\mathrm{sub}}\sqrt{n}\end{aligned}$$

with probability at least $1-2/p^2$ since $D^{\frac{1}{2}}\,\mathbb{X}_{n,S}\left(\mathbb{X}_{n,S}^\top\, D\,\mathbb{X}_{n,S}\right)^{-1}\mathbb{X}_{n,S}^\top\, D^{\frac{1}{2}}$ is an orthogonal projection matrix and D a diagonal matrix with entries smaller than or equal to 1. On the other hand under the condition $n \geq c_3^{\mathrm{RSC}} s\log(p)$ the smallest eigenvalue of $\widehat{Q}_{SS} = \frac{2}{n}\,\mathbb{X}_{n,S}^\top\, D\,\mathbb{X}_{n,S}$ is bounded below by $c_{\mathbf{X},\mathrm{l}}/32$ with probability at least $1 - c_1^{\mathrm{P}}\exp(-c_1^{\mathrm{P}}n)$, see Lemma 3.9, which implies

$$\left\|D^{\frac{1}{2}}\,\mathbb{X}_{n,S}\,r_k\right\|_2^2 = r_k^\top\,\mathbb{X}_{n,S}^\top\, D\,\mathbb{X}_{n,S}\,r_k = \frac{n}{2}\,r_k^\top\left(\frac{2}{n}\,\mathbb{X}_{n,S}^\top\, D\,\mathbb{X}_{n,S}\right)r_k \geq \frac{c_{\mathbf{X},\mathrm{l}}}{64}\,\|r_k\|_2^2\,n$$

for all $k \in S^c$. Hence we obtain by the last inequalities the estimate

$$\max_{k\in S^c}\|r_k\|_2 \leq \max_{k\in S^c}\left(\frac{64}{c_{\mathbf{X},\mathrm{l}}\,n}\right)^{\frac{1}{2}}\left\|D^{\frac{1}{2}}\,\mathbb{X}_{n,S}\,r_k\right\|_2 \leq \frac{33\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}}\,,$$

and in total

$$\left\|\widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\right\|_{\mathrm{M},\infty} = \max_{k\in S^c}\|r_k\|_1 \leq \max_{k\in S^c}\sqrt{s}\,\|r_k\|_2 \leq \frac{33\,c_{\mathbf{X},\mathrm{sub}}\sqrt{s}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}}$$

with high probability. □

Lemma 3.14 immediately follows from the following lemma and Lemma 3.8.

Lemma 3.23. *Suppose Assumption 3.1 and* $\alpha_n \geq \sqrt{4/3}\,c_1^{\mathrm{Grad}}\,(\log(p)/n)^{\frac{1}{2}}$ *hold. Then for* $s \leq \log(p)$ *and* $n \geq \max\left\{c_3^{\mathrm{RSC}} s\log(p), 6\log(p)\right\}$ *we have that*

$$\left\|\widehat{Q}_{S^cS}\,\big(\widehat{Q}_{SS}\big)^{-1}\left(\nabla\mathcal{L}_{n,\alpha_n}^{\mathrm{H}}\left(\beta_{\alpha_n}^*\right)\right)_S\right\|_\infty \leq \frac{\sqrt{4}\,66\,c_{\mathbf{X},\mathrm{sub}}^2}{\sqrt{3\,c_{\mathbf{X},\mathrm{l}}}}\,c_2^{\mathrm{Grad}}\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}$$

with probability at least $1 - c_1^{\mathrm{P}}\exp(-c_2^{\mathrm{P}}n) - 4/p^2$.

Proof of Lemma 3.23. Set

$$\mathcal{T} = \left\{ \left\| \widehat{Q}_{S^cS} \left(\widehat{Q}_{SS}\right)^{-1} \left(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S \right\|_\infty \leq \frac{\sqrt{4}\, 66\, c^2_{\mathbf{X},\mathrm{sub}}}{\sqrt{3\, c_{\mathbf{X},\mathrm{l}}}}\, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \right\}.$$

Then

$$\begin{aligned}
\mathbb{P}\left(\mathcal{T}^c\right) &\leq \mathbb{P}\left(\mathcal{T}^c \cap \left\{ \max_{k\in\{1\ldots,p-s\}} \left\| \left(e_k^\top \widehat{Q}_{S^cS}\left(\widehat{Q}_{SS}\right)^{-1}\right)^\top \right\|_2 \leq \frac{33\, c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}} \right\}\right) \\
&\qquad + \mathbb{P}\left(\max_{k\in\{1\ldots,p-s\}} \left\| \left(e_k^\top \widehat{Q}_{S^cS}\left(\widehat{Q}_{SS}\right)^{-1}\right)^\top \right\|_2 > \frac{33\, c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}} \right) \\
&\leq \mathbb{P}\left(\mathcal{T}^c \cap \left\{ \max_{k\in\{1\ldots,p-s\}} \left\| \left(e_k^\top \widehat{Q}_{S^cS}\left(\widehat{Q}_{SS}\right)^{-1}\right)^\top \right\|_2 \leq \frac{33\, c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}} \right\}\right) \\
&\qquad + c_1^{\mathrm{P}} \exp(-c_2^{\mathrm{P}} n) + 2/p^2 \qquad (3.91)
\end{aligned}$$

because of Lemma 3.18. Further, by definition of the event $\mathcal{T}$,

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{T}^c \cap \left\{ \max_{k\in\{1\ldots,p-s\}} \left\| \left(e_k^\top \widehat{Q}_{S^cS}\left(\widehat{Q}_{SS}\right)^{-1}\right)^\top \right\|_2 \leq \frac{33\, c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}} \right\}\right) \\
&\quad = \mathbb{P}\Bigg(\left\{ \max_{k\in\{1\ldots,p-s\}} \left| e_k^\top \widehat{Q}_{S^cS} \left(\widehat{Q}_{SS}\right)^{-1} \left(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S \right| \right. \\
&\qquad\qquad\qquad \left. > \frac{\sqrt{4}\, 66\, c^2_{\mathbf{X},\mathrm{sub}}}{\sqrt{3\, c_{\mathbf{X},\mathrm{l}}}}\, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \right\} \\
&\qquad\qquad \cap \left\{ \max_{k\in\{1\ldots,p-s\}} \left\| \left(e_k^\top \widehat{Q}_{S^cS}\left(\widehat{Q}_{SS}\right)^{-1}\right)^\top \right\|_2 \leq \frac{33\, c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}} \right\}\Bigg) \\
&\quad \leq \mathbb{P}\left(\max_{u\in\mathbb{R}^s:\|u\|_2 \leq \frac{33\, c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}}} \left| u^\top \left(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S \right| > \frac{\sqrt{4}\, 66\, c^2_{\mathbf{X},\mathrm{sub}}}{\sqrt{3\, c_{\mathbf{X},\mathrm{l}}}}\, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \right) \\
&\quad = \mathbb{P}\Bigg(\max_{u\in\mathbb{R}^s:\|u\|_2\leq 1} \left| u^\top \left(\frac{33\, c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}} \left(\nabla \mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S \right) \right| \\
&\qquad\qquad\qquad > \frac{\sqrt{4}\, 66\, c^2_{\mathbf{X},\mathrm{sub}}}{\sqrt{3\, c_{\mathbf{X},\mathrm{l}}}}\, c_2^{\mathrm{Grad}} \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \Bigg). \qquad (3.92)
\end{aligned}$$

In the following we proceed with a covering argument. Let A denote a 1/2-cover of cardinality $N = N(1/2; \mathcal{B}^s_2, \|\cdot\|_2)$ of the unit Euclidean ball $\mathcal{B}^s_2 = \left\{u \in \mathbb{R}^s : \|u\|_2 \leq 1\right\}$ of $\mathbb{R}^s$ with respect to the Euclidean distance, cf. Rigollet and Hütter (2019, Definition 1.17) or Wainwright (2019, Definition 5.1). Then, as in the proof of Rigollet and Hütter

(2019, Theorem 1.19), we obtain

$$\mathbb{P}\left(\max_{u\in\mathcal{B}_2^s}\left|u^\top\left(\frac{33\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}}\left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S\right)\right|>\frac{\sqrt{4}\,66\,c^2_{\mathbf{X},\mathrm{sub}}}{\sqrt{3\,c_{\mathbf{X},\mathrm{l}}}}\,c_2^{\mathrm{Grad}}\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}\right)$$
$$\leq\mathbb{P}\left(\max_{u\in A}\left|u^\top\left(\frac{33\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}}\left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S\right)\right|>\frac{\sqrt{4}\,33\,c^2_{\mathbf{X},\mathrm{sub}}}{\sqrt{3\,c_{\mathbf{X},\mathrm{l}}}}\,c_2^{\mathrm{Grad}}\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}\right). \tag{3.93}$$

Now we can write for fixed $u\in A$, analogously to the proof of Lemma 3.21,

$$u^\top\left(\frac{33\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}}\left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S\right)=-\frac{1}{n}\sum_{i=1}^n l'_{\alpha_n}\left(Y_i-\mathbf{X_i}^\top\beta^*_{\alpha_n}\right)Z_i$$

with $Z_i=33\,c_{\mathbf{X},\mathrm{sub}}/\sqrt{c_{\mathbf{X},\mathrm{l}}}\;u^\top\left(\mathbf{X_i}\right)_S$. The random variables $l'_{\alpha_n}\left(Y_1-\mathbf{X_1}^\top\beta^*_{\alpha_n}\right)Z_1,\dots,$ $l'_{\alpha_n}\left(Y_n-\mathbf{X_n}^\top\beta^*_{\alpha_n}\right)Z_n$ are independent and identically distributed and have mean equal to zero, see proof of Lemma 3.21 for more details. In addition (iii) of Assumption 3.1 implies that the random variables $Z_1,\dots,Z_n$ are sub-Gaussian with variance proxy $1089\,c^4_{\mathbf{X},\mathrm{sub}}/c_{\mathbf{X},\mathrm{l}}$. This leads to

$$\mathbb{E}\left[\left(l'_{\alpha_n}\left(Y_i-\mathbf{X_i}^\top\beta^*_{\alpha_n}\right)Z_i\right)^2\right]\leq\frac{1089\,c^2_{\mathbf{X},\mathrm{sub}}\,c_3^{\mathrm{Grad}}}{c_{\mathbf{X},\mathrm{l}}}$$

and

$$\mathbb{E}\left[\left|l'_{\alpha_n}\left(Y_i-\mathbf{X_i}^\top\beta^*_{\alpha_n}\right)Z_i\right|^u\right]\leq\frac{u!}{2}\left(\frac{2\,33\,c_{\mathbf{X},\mathrm{sub}}\,c_4^{\mathrm{Grad}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}\,\alpha_n}\right)^{u-2}c_3^{\mathrm{Grad}}$$

for $u\in\mathbb{N}$, $u\geq 3$, where c_3^{Grad} and c_4^{Grad} are given in the proof of Lemma 3.8. Bernstein's inequality and the choice of α_n leads for fixed $u\in A$ to

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n l'_{\alpha_n}\left(Y_i-\mathbf{X_i}^\top\beta^*_{\alpha_n}\right)Z_i\right|\geq\frac{66\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}}\left(\frac{8c_3^{\mathrm{Grad}}\,\log(p)}{n}\right)^{\frac{1}{2}}\right)\leq 2\exp\left(-4\log(p)\right),$$

see proof of Lemma 3.8 for more details. By the union bound and the definition of c_2^{Grad} in the proof of Lemma 3.8 we get

$$\mathbb{P}\left(\max_{u\in A}\left|u^\top\left(\frac{33\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{c_{\mathbf{X},\mathrm{l}}}}\left(\nabla\mathcal{L}^{\mathrm{H}}_{n,\alpha_n}\left(\beta^*_{\alpha_n}\right)\right)_S\right)\right|>\frac{\sqrt{4}\,33\,c^2_{\mathbf{X},\mathrm{sub}}}{\sqrt{3\,c_{\mathbf{X},\mathrm{l}}}}\,c_2^{\mathrm{Grad}}\left(\frac{\log(p)}{n}\right)^{\frac{1}{2}}\right)$$
$$\leq 2\,N\,\exp\left(-4\log(p)\right)\leq 2\exp\left(-4\log(p)+s\log(6)\right)$$
$$\leq 2\exp\left(-4\log(p)+2\log(p)\right)=\frac{2}{p^2} \tag{3.94}$$

since the 1/2-covering-number N can be upper bounded by 6^s, cf. Rigollet and Hütter (2019, Lemma 1.18) or Wainwright (2019, Example 5.8), and we assumed $s\leq\log(p)$. In conclusion the inequalities (3.91) - (3.94) imply the assertion of the lemma. □

3.6.4. Proofs of (3.40) and (3.41)

From (3.21) in Lemma 3.9 we obtain

$$\begin{aligned}
&\Big(\widehat{Q}_{S^c(S_{\alpha_n}\setminus S)} - \widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\Big)\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\\
&\quad= \Big(\frac{2}{n}\,\mathbb{X}^\top_{n,S^c}\,D\,\mathbb{X}_{n,S_{\alpha_n}\setminus S} - \frac{2}{n}\mathbb{X}^\top_{n,S^c}\,D\,\mathbb{X}_{n,S}\big(\mathbb{X}^\top_{n,S}\,D\,\mathbb{X}_{n,S}\big)^{-1}\mathbb{X}^\top_{n,S}\,D\,\mathbb{X}_{n,S_{\alpha_n}\setminus S}\Big)\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\\
&\quad= \frac{2}{n}\,\mathbb{X}^\top_{n,S^c}\,D^{\frac{1}{2}}\Big(\mathrm{I}_n - D^{\frac{1}{2}}\,\mathbb{X}_{n,S}\big(\mathbb{X}^\top_{n,S}\,D\,\mathbb{X}_{n,S}\big)^{-1}\mathbb{X}^\top_{n,S}\,D^{\frac{1}{2}}\Big)D^{\frac{1}{2}}\,\mathbb{X}_{n,S_{\alpha_n}\setminus S}\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\,.
\end{aligned}$$

The matrix in brackets, which we will denote by P, is an orthogonal projection matrix. Therefore using Lemma 3.22, on an event with probability at least $1 - 2/p^2$ we obtain

$$\begin{aligned}
\max_{k\in\{1,\dots,p-s\}}\left\|\Big(e_k^\top\,\frac{2}{n}\,\mathbb{X}^\top_{n,S^c}\,D^{\frac{1}{2}}\,\mathrm{P}\,D^{\frac{1}{2}}\Big)^\top\right\|_2 &\le \max_{k\in S^c}\Big(\frac{2}{n}\,\big\|D^{\frac{1}{2}}\big\|_{\mathrm{M},2}\,\big\|\mathrm{P}\big\|_{\mathrm{M},2}\,\big\|D^{\frac{1}{2}}\big\|_{\mathrm{M},2}\big\|\vec{X}_k\big\|_2\Big)\\
&\le \frac{2\sqrt{17}\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{n}}
\end{aligned}$$

since the entries of the diagonal matrix D are smaller than or equal to 1. Setting $Q = \mathbb{X}_{n,S_{\alpha_n}\setminus S}\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}$, for $x > 0$ this leads to

$$\begin{aligned}
&\mathbb{P}\Bigg(\max_{k\in\{1,\dots,p-s\}}\left|e_k^\top\,\frac{2}{n}\,\mathbb{X}^\top_{n,S^c}\,D^{\frac{1}{2}}\,\mathrm{P}\,D^{\frac{1}{2}}\,\mathbb{X}_{n,S_{\alpha_n}\setminus S}\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\right| > x\,,\\
&\qquad\qquad\qquad\qquad\max_{k\in\{1,\dots,p\}}\frac{1}{\sqrt{n}}\,\big\|\vec{X}_k\big\|_2 \le \sqrt{17}\,c_{\mathbf{X},\mathrm{sub}}\Bigg)\\
&\quad\le \mathbb{P}\Bigg(\max_{k\in\{1,\dots,p-s\}}\left|e_k^\top\,\frac{2}{n}\,\mathbb{X}^\top_{n,S^c}\,D^{\frac{1}{2}}\,\mathrm{P}\,D^{\frac{1}{2}}\,Q\right| > x\,,\\
&\qquad\qquad\qquad\max_{k\in\{1,\dots,p-s\}}\left\|\Big(e_k^\top\,\frac{2}{n}\,\mathbb{X}^\top_{n,S^c}\,D^{\frac{1}{2}}\,\mathrm{P}\,D^{\frac{1}{2}}\Big)^\top\right\|_2 \le \frac{2\sqrt{17}\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{n}}\Bigg)\\
&\quad\le \mathbb{P}\Bigg(\max_{u\in\mathbb{R}^n:\|u\|_2\le\frac{2\sqrt{17}\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{n}}}\big|u^\top Q\big| > x\Bigg)\\
&\quad= \mathbb{P}\Bigg(\max_{u\in\mathbb{R}^n:\|u\|_2\le 1}\left|u^\top\Big(\frac{2\sqrt{17}\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{n}}\,Q\Big)\right| > x\Bigg).
\end{aligned}$$

The vector Q has independent and sub-Gaussian entries

$$\big(\mathbf{X_i}\big)^\top_{S_{\alpha_n}\setminus S}\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S} = \mathbf{X_i}^\top\begin{pmatrix}\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\\ \mathbf{0}_{|(S_{\alpha_n}\setminus S)^c|}\end{pmatrix} \sim \mathrm{subG}\Big(c_{\mathbf{X},\mathrm{sub}}\,\big\|\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\big\|_2\Big)$$

by (iii) of Assumption 3.1, and Rigollet and Hütter (2019, Theorem 1.6) implies

$$\frac{2\sqrt{17}\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{n}}\,Q \sim \mathrm{subG}_n\left(\frac{2\sqrt{17}\,c^2_{\mathbf{X},\mathrm{sub}}\,\big\|\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\big\|_2}{\sqrt{n}}\right).$$

Finally, Rigollet and Hütter (2019, Theorem 1.19) with the choice $\delta = \exp(-2n)$ leads to

$$\mathbb{P}\left(\max_{u\in\mathbb{R}^n:\|u\|_2\le 1}\left|u^\top\left(\frac{2\sqrt{17}\,c_{\mathbf{X},\mathrm{sub}}}{\sqrt{n}}\,Q\right)\right| > 16\sqrt{17}\,c^2_{\mathbf{X},\mathrm{sub}}\,\big\|\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\big\|_2\right) \le \exp\big(-2n\big)\,,$$

so that we obtain overall

$$\begin{aligned}\mathbb{P}\bigg(\Big\|\Big(\widehat{Q}_{S^c(S_{\alpha_n}\setminus S)} - \widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\Big)\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\Big\|_\infty&\\ > 16\sqrt{17}\,c^2_{\mathbf{X},\mathrm{sub}}\,\big\|\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\big\|_2\bigg)&\\ \le \exp\big(-2n\big) + 2/p^2\,.&\end{aligned} \tag{3.95}$$

Similarly, for the vector $\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S} = \frac{2}{n}\,\mathbb{X}^\top_{n,S}\,D\,\mathbb{X}_{n,S_{\alpha_n}\setminus S}\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}$, arguing as for (3.95) we obtain

$$\mathbb{P}\left(\Big\|\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\Big\|_\infty > 16\sqrt{17}\,c^2_{\mathbf{X},\mathrm{sub}}\,\big\|\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\big\|_2\right) \le \exp\big(-2n\big) + 2/p^2\,.$$

From Lemma 3.4 we get

$$\Big\|\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\Big\|_2 = \Big\|\beta^*_{\alpha_n,S_{\alpha_n}\setminus S} - \beta^*_{S_{\alpha_n}\setminus S}\Big\|_2 \le \big\|\beta^*_{\alpha_n} - \beta^*\big\|_2 \le C_{\mathrm{apx}}\,\alpha_n^{m-1}$$

since $\beta^*_l = 0$ for all $l \in S_{\alpha_n}\setminus S$. So in total we have

$$\begin{aligned}\Big\|\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\Big\|_\infty,\ \Big\|\Big(\widehat{Q}_{S^c(S_{\alpha_n}\setminus S)} - \widehat{Q}_{S^cS}\big(\widehat{Q}_{SS}\big)^{-1}\widehat{Q}_{S(S_{\alpha_n}\setminus S)}\Big)\,\beta^*_{\alpha_n,S_{\alpha_n}\setminus S}\Big\|_\infty&\\ \le 80\,C_{\mathrm{apx}}\,c^2_{\mathbf{X},\mathrm{sub}}\,\alpha_n^{m-1}&\end{aligned}$$

with probability at least $1 - 2\exp(-2n) - 2/p^2$, which yields the claimed inequalities in (3.40) and (3.41).

Part II.

Variable selection for the means and covariances in linear random coefficient regression models with the adaptive LASSO

4. Identifiability of the mean vector and covariance matrix

In this chapter we consider the linear regression model with random coefficients introduced in Section 2.4. From a nonparametric point of view the identification and estimation of the joint distribution or density of the coefficients is intensively studied, cf. Beran and Hall (1992), Hoderlein et al. (2010) and Holzmann and Meister (2020). For that purpose a large support of the covariates is required. However, often explanatory variables only have compact or even finite support. Hence our main interest in this chapter is the identifiability of the mean vector and covariance matrix of the (potentially) random coefficients from regressors with possibly only finite support. Some consistent estimators for this parametric approach are also provided in Hildreth and Houck (1968) where the authors, however, restrict themselves to mutually independent coefficients. Note that the first and second moments would determine the density under a normality assumption, but we do not assume this here.

Now we repeat briefly the concept of identifiability in our statistical model. Consider the linear regression model (2.6) with random coefficients. Let $\theta \in \Theta$ be the unknown parameters of the conditional distribution $\mathbb{P}_\theta$ of the response variable Y. Then identifiability of the parameters means that $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$ leads to $\theta_1 = \theta_2$ for $\theta_1, \theta_2 \in \Theta$. In the following results the choice of the parameter space Θ will be clear from the context. For example, in the final theorems the unknown parameters are the means, variances and covariances of the random coefficients, that means $\theta = (\mu^\top, \sigma^\top)^\top$ and $\Theta = \mathbb{R}^p \times \mathbb{V}_p^+$ with

$$\mathbb{V}_p^+ = \big\{\mathrm{vec}(M) \,\big|\, M \in \mathbb{R}^{p\times p} \text{ symmetric and positive semi-definite}\big\} \quad \subset \mathbb{R}^{\frac{p(p+1)}{2}}.$$

This chapter is structured as follows. In Sections 4.1 and 4.2 we provide sufficient conditions for the identifiability of the first and second central moments in the linear random coefficient regression model (2.6), where the detailed proofs are deferred to Section 4.3.

4.1. Preliminary results

In previous papers about the estimation of the moments, e.g. Hildreth and Houck (1968), the authors often a-priori assumed that the covariance matrix Σ in the linear random coefficient regression model (2.6) is a diagonal matrix,

$$\Sigma = \mathrm{diag}\big(\sigma_1^2, \dots, \sigma_p^2\big). \tag{4.1}$$

However, first of all correlations between the random coefficients may be of applied interest. Secondly, even if the main interest lies on the variances of the coefficients, in particular on their potential randomness, it is important not to exclude correlations without very good reasons, since otherwise one may draw wrong conclusions about the (non-)randomness of the coefficients. Let us illustrate the second point for a univariate regressor W_1, which means that the model (2.6) simplifies to

$$Y = B_0 + W_1\, B_1\,. \tag{4.2}$$

The following proofs are all provided in Section 4.3.1.

Proposition 4.1. *Suppose that in model (4.2) the random variable $W_1 \in \{0,1\}$ is binary, and denote the identified standard deviations by*

$$s_1 = \sqrt{\mathbb{V}\mathrm{ar}(B_0)}\,, \qquad s_2 = \sqrt{\mathbb{V}\mathrm{ar}(B_0+B_1)}\,.$$

Then each value

$$\sqrt{\mathbb{V}\mathrm{ar}(B_1)} \in \big[|s_1 - s_2|, s_1 + s_2\big]$$

is consistent with s_1 and s_2, provided the correlation $\rho = \mathbb{C}\mathrm{or}(B_0, B_1)$ is chosen for $s_1 > 0$ and $\sqrt{\mathbb{V}\mathrm{ar}(B_1)} > 0$ as

$$\rho = \frac{s_2^2 - s_1^2 - \mathbb{V}\mathrm{ar}(B_1)}{2\, s_1 \sqrt{\mathbb{V}\mathrm{ar}(B_1)}} \in \begin{cases} [-1,1], & \text{if } s_1 < s_2\,, \\ \big[-1, -\sqrt{s_1^2 - s_2^2}/s_1\big] & \text{if } s_1 \geq s_2\,. \end{cases} \tag{4.3}$$

Thus, to conclude from $\mathbb{V}\mathrm{ar}(B_0) = \mathbb{V}\mathrm{ar}(B_0 + B_1)$ that $\mathbb{V}\mathrm{ar}(B_1) = 0$ fully relies on the assumption of a diagonal covariance matrix, without this assumption B_1 can well be random. Hence we mainly focus on the case of a general covariance matrix Σ.

Let us proceed with some preliminary results. Consider at first the simple model (4.2) again.

Proposition 4.2. *In model* (4.2)*, if W_1 has $n+1$ support points and $\mathbb{E}\big[|B_0|^n\big] < \infty$, $\mathbb{E}\big[|B_1|^n\big] < \infty$, then all mixed moments $\mathbb{E}\big[B_0^j\, B_1^k\big]$, $j,k \geq 0$, $j+k \leq n$ are identified.*

Remark 4.3. The proposition shows that three distinct support points of W_1 are enough to identify the means $\mathbb{E}[B_j]$, the variances $\mathbb{V}\mathrm{ar}(B_j)$, $j = 0,1$, and the covariance $\mathbb{C}\mathrm{ov}(B_0, B_1)$. From Proposition 4.1 and not surprisingly, two support points are insufficient for this purpose.

Let us now turn to general dimensions for the regressors. Recall that points $\mathbf{w_1}, \ldots, \mathbf{w_d} \in \mathbb{R}^{d-1}$ are said to be in general position if $\sum_{k=1}^d \alpha_k \mathbf{w_k} = \mathbf{0}_{d-1}$ for $\alpha_k \in \mathbb{R}$, $\sum_{k=1}^d \alpha_k = 0$, implies that $\alpha_1 = \ldots = \alpha_d = 0$.

Lemma 4.4 (General position)**.** *Points $\mathbf{w_1}, \ldots, \mathbf{w_d} \in \mathbb{R}^{d-1}$ are in general position if and only if one of the following conditions holds.*

1. *$\mathbf{w_2} - \mathbf{w_1}, \ldots, \mathbf{w_d} - \mathbf{w_1}$ are linearly independent.*
2. *For each $j \in \{1, \ldots, d\}$ the point $\mathbf{w_j}$ is not contained in $\big\{\sum_{k=1,k\neq j}^d \alpha_k \mathbf{w_k}$, $\sum_{k=1,k\neq j}^d \alpha_k = 1\big\}$, the hyperplane generated by $\mathbf{w_k}, k \neq j$.*

Before going to the general case, let us briefly consider diagonal covariance matrices for the random coefficients. For a vector $\mathbf{w} = (w_1, \ldots, w_d)^\top \in \mathbb{R}^d$ we write $\mathbf{w}^2 = (w_1^2, \ldots, w_d^2)^\top \in \mathbb{R}^d$.

Proposition 4.5 (Identifiability of the means and variances)**.** *Consider model* (2.6) *with a diagonal covariance matrix as in* (4.1)*. If the support of* $\mathbf{W}$ *contains* p *points* $\mathbf{w_1}, \ldots, \mathbf{w_p} \in \mathbb{R}^{p-1}$ *in general position, the means* μ *are identified. If, moreover, the support also contains possibly distinct points* $\bar{\mathbf{w}}_\mathbf{1}, \ldots, \bar{\mathbf{w}}_\mathbf{p} \in \mathbb{R}^{p-1}$ *for which* $\bar{\mathbf{w}}_\mathbf{1}{}^2, \ldots, \bar{\mathbf{w}}_\mathbf{p}{}^2$ *are in general position, the variances* $\sigma_1^2, \ldots, \sigma_p^2$ *are identified as well.*

Remark 4.6. Consider model (2.6) and let $p = 3$, that is $Y = B_0 + W_1 B_1 + W_2 B_2$. Consider the support points $(1,1)^\top$, $(2,4)^\top$, $(\sqrt{5}, \sqrt{21})^\top$, which are in general position, while the squares $(1,1)^\top$, $(4,16)^\top$, $(5,21)^\top$ are not. However, if we have an additional support point $(w_1, w_2)^\top$ such that $(w_1^2, w_2^2)^\top$ is not on the line $(1,1)^\top + \lambda\,(1,5)^\top$, $\lambda \in \mathbb{R}$, then $(w_1, w_2)^\top$ will be in general position with at least one pair of the original three points, and the squares of these three points will be in general position as well.

4.2. Full point identification

Now we turn to the case of a general covariance matrix for the random coefficients $\mathbf{A}$. If the means μ of $\mathbf{A}$ are identified, we identify for the variances and covariances by

$$\mathbb{V}\mathrm{ar}\left(Y \,\middle|\, \mathbf{W} = \mathbf{w}\right) = (1, \mathbf{w}^\top)\,\Sigma\,(1, \mathbf{w}^\top)^\top = \mathrm{v}\Big((1, \mathbf{w}^\top)^\top\Big)^\top \sigma\,, \tag{4.4}$$

where $\mathbf{w}$ ranges over the support of $\mathbf{W}$, the half-vectorization σ of the covariance matrix Σ of the coefficients $\mathbf{A}$ is given in (2.7) and the associated vector transformation v is defined by

$$\begin{aligned} \mathrm{v} : \mathbb{R}^d &\to \mathbb{R}^{\frac{d(d+1)}{2}}\,, \\ \mathbf{x} &\mapsto \left(x_1^2, \ldots, x_d^2, 2x_1x_2, \ldots, 2x_1x_d, 2x_2x_3, \ldots, 2x_2x_p, \ldots, 2x_{d-1}x_d\right)^\top. \end{aligned} \tag{4.5}$$

Note that we can also write

$$\mathrm{v}\Big((1, \mathbf{w}^\top)^\top\Big) = \left(1, (\mathbf{w}^2)^\top, 2\mathbf{w}^\top, 2w_1w_2, \ldots, 2w_1w_{p-1}, 2w_2w_3, \ldots, 2w_{p-2}w_{p-1}\right)^\top.$$

Based on equation (4.4) we can establish a linear system for the $p(p+1)/2$ entries of σ (respectively Σ). With the above notation, we may state the following basic result. The proofs of this section are deferred to Section 4.3.2.

Theorem 4.7 (Identifiability of the means, variances and covariances I)**.** *In model* (2.6) *a sufficient condition for identification of the mean vector* μ *and the covariance matrix* Σ *of the random coefficients is the existence of* $p(p+1)/2$ *points* $\mathbf{w_1}, \ldots, \mathbf{w}_{\frac{\mathbf{p(p+1)}}{\mathbf{2}}} \in \mathbb{R}^{p-1}$ *in the support of* $\mathbf{W}$*, for which the matrix*

$$S = \left[\mathrm{v}\Big((1, \mathbf{w_1}^\top)^\top\Big), \ldots, \mathrm{v}\Big((1, \mathbf{w}_{\frac{\mathbf{p(p+1)}}{\mathbf{2}}}^\top)^\top\Big)\right]^\top \quad \in \mathbb{R}^{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}} \tag{4.6}$$

is of full rank. This condition is also necessary among the full-rank covariance matrices.

Note that certain singular covariance matrices may also be identified from lower-rank matrices S.

Remark 4.8. For which points is the matrix S in (4.6) of full rank? Its determinant is a sum of monomials in the coordinate variables of the vectors $\mathbf{w_1}, \dots, \mathbf{w_{\frac{p(p+1)}{2}}} \in \mathbb{R}^{p-1}$ of degree $(p-1)^2 + 2(p-1)$. Hence characterizing its zero set, that are those points for which the requirement of full rank is not satisfied, is a formidable task even for $p = 2$. Of course, given $p(p+1)/2$ support points, one may simply form the matrix S and compute its determinant. More generally, given $m \geq p(p+1)/2$ support points, one may directly check whether

$$S_m = \left[\mathrm{v}\Big((1, \mathbf{w_1}^\top)^\top \Big), \dots, \mathrm{v}\Big((1, \mathbf{w_m}^\top)^\top \Big) \right]^\top \qquad \in \mathbb{R}^{m \times \frac{p(p+1)}{2}}$$

has full rank $p(p+1)/2$, which is equivalent to the $p(p+1)/2$ column vectors of S_m being linearly independent in $\mathbb{R}^m$, or to $S_m^\top S_m$ being invertible.

Example 4.9. Let $\mathbf{W} = (W_1, \dots, W_{p-1})^\top$, and suppose that W_1 has only two support points a and b. The joint support of $\mathbf{W}$ is assumed to be finite (the example remains true without this assumption), but of unrestricted cardinality. Then the matrix S_m, where m is total number of support points, has rank at most $p(p+1)/2 - 1$. Thus, from Theorem 4.7, full-rank covariance matrices Σ are not identified. Indeed, the matrix $S_m^\top$ contains the submatrix

$$\begin{bmatrix} 1 & \dots & 1 & 1 & \dots & 1 \\ a & \dots & a & b & \dots & b \\ a^2 & \dots & a^2 & b^2 & \dots & b^2 \end{bmatrix} \quad \in \mathbb{R}^{3 \times m}.$$

Evidently, this matrix has column rank at most 2, since there are only two distinct columns. Thus, its row rank is also at most two, which implies that the corresponding three columns in S_m are linearly dependent.

Hence it is of some interest to have a simple, sufficient condition for identifiability of the covariance matrix. We provide the following general result.

Proposition 4.10. *Consider model* (2.6) *and suppose that the support of* $\mathbf{W}$ *contains points satisfying the following properties.*

1. *The p points $\mathbf{w_1}, \dots, \mathbf{w_p} \in \mathbb{R}^{p-1}$ are in general position.*
2. *For each $j \in \{1, \dots, p\}$ there exist points $\mathbf{w_{j,1}}, \dots, \mathbf{w_{j,p-1}} \in \mathbb{R}^{p-1}$, possibly equal to those in 1., such that*
   - *$\mathbf{w_j}, \mathbf{w_{j,1}}, \dots, \mathbf{w_{j,p-1}}$ are in general position,*
   - *and for each $j \in \{1, \dots, p\}$, $k \in \{1, \dots, p-1\}$ there is a $\mathbf{z_{j,k}} \in \mathbb{R}^{p-1}$ for which $\mathbf{w_j}, \mathbf{w_{j,k}}, \mathbf{z_{j,k}}$ are all distinct but generate only a one-dimensional affine space, that means they are all contained in a line.*

Then the design matrix S in (4.6) *formed from all the points $\mathbf{w_j}, \mathbf{w_{j,k}}$ and $\mathbf{z_{j,k}}$ has full rank $p(p+1)/2$ and hence, the mean vector μ and the covariance matrix Σ of the random coefficients are identified.*

The minimal number of support points required in Proposition 4.10 is $p + p(p-1)/2 = p(p+1)/2$, which corresponds to the number of free parameters in Σ. Our major application of this proposition is to covariates having Cartesian products as supports. The following theorem makes this precise.

Theorem 4.11 (Identifiability of the means, variances and covariances II)**.** *Consider model* (2.6). *Suppose that the support of the covariate vector* $\mathbf{W} = (W_1, \dots, W_{p-1})^\top$ *contains a Cartesian product of three distinct points in each coordinate. Then there exist* $p(p+1)/2$ *support points such that the matrix* S *in* (4.6) *has full rank* $p(p+1)/2$ *and consequently, the mean vector* μ *and the covariance matrix* Σ *of the random coefficients are identified. Conversely, if there is a* W_j *having only two support points, then in the full-rank covariance matrices identification fails.*

4.3. Detailed proofs

4.3.1. Proofs for Section 4.1

Proof of Proposition 4.1. Set $u = \sqrt{\mathbb{V}\mathrm{ar}(B_1)}$. From $s_2^2 = s_1^2 + u^2 + 2\,\rho\, s_1\, u$ and $|\rho| \leq 1$ we obtain the inequalities

$$(u - s_1)^2 \leq s_2^2 \leq (u + s_1)^2.$$

By equating $s_2^2 = (u+s_1)^2$ we obtain the solutions $\pm s_2 - s_1$ for u, which yields $u \geq s_2 - s_1$ if $s_2 > s_1$. If $s_2 \leq s_1$ we obviously have only the bound $u \geq 0$. Equating $s_2^2 = (u - s_1)^2$ gives the solutions $\pm s_2 + s_1$ for u, which yields the bounds

$$u \in \big[|s_1 - s_2|, s_1 + s_2\big]$$

for the standard deviation u of B_1. Solving the equation at the beginning for the correlation gives $\rho = (s_2^2 - s_1^2 - u^2)/(2\,s_1 u)$, which ranges over the whole interval $[-1, 1]$ if $s_2 > s_1$. If $s_1 \geq s_2$, the correlation must be negative, and maximizing the above expression for ρ over u yields $u = \sqrt{s_1^2 - s_2^2}$, and finally the upper bound in (4.3). □

Proof of Proposition 4.2. It is enough to show that all mixed moments of order n are identified from $n + 1$ support points, the claim then follows by induction. It is

$$\mathbb{E}\big[Y^n \,\big|\, W_1 = w\big] = \mathbb{E}\big[(B_0 + w\,B_1)^n\big] = \sum_{k=0}^{n} \binom{n}{k} w^k\, \mathbb{E}\big[B_0^{n-k}\, B_1^k\big]$$

by model (4.2). If W_1 has distinct support points $w_1, \dots, w_{n+1}$, we obtain a linear system for the moments $\mathbb{E}\big[B_0^{n-k}\, B_1^k\big]$, $k = 0, \dots, n$. Its design matrix

$$\mathbb{X} = \left(\binom{n}{k-1} w_j^{k-1}\right)_{j,k \in \{1,\dots,n+1\}}$$

satisfies

$$\begin{aligned}\det\left(\mathbb{X}\right) &= \prod_{l=0}^{n} \binom{n}{l} \ \det\left(w_j^{k-1}\right)_{j,k\in\{1,\dots,n+1\}}\\ &= \prod_{l=0}^{n} \binom{n}{l} \prod_{1\leq j<k\leq n+1} (w_k - w_j)\\ &\neq 0\,,\end{aligned}$$

where we used in the last equation the determinant of the Vandermonde matrix. Hence the design matrix $\mathbb{X}$ is nonsingular and its kernel is trivial. As a consequence, the identity of the conditional moments of Y implies identical mixed moments $\mathbb{E}\big[B_0^{n-k}\,B_1^k\big]$, $k = 0, \dots, n$. □

Proof of Lemma 4.4. Equivalences are clear by the definition of vectors in general position. □

Proof of Proposition 4.5. The linear system $\mathbb{E}[Y \,|\, \mathbf{W} = \mathbf{w_j}] = \mathbb{E}[B_0] + \mathbf{w_j}^\top \mathbb{E}[\mathbf{B}]$, $j = 1, \dots, p$, has design matrix

$$\begin{bmatrix} 1 & \mathbf{w_1}^\top \\ \vdots & \vdots \\ 1 & \mathbf{w_p}^\top \end{bmatrix}.$$

The matrix is of full rank, since its rank is the same as the one of the matrix

$$\begin{bmatrix} 1 & \mathbf{w_1}^\top \\ 0 & \mathbf{w_2}^\top - \mathbf{w_1}^\top \\ \vdots & \vdots \\ 0 & \mathbf{w_p}^\top - \mathbf{w_1}^\top \end{bmatrix},$$

which is invertibe by Lemma 4.4. Hence the means $\mu = \mathbb{E}[\mathbf{A}]$ are identified, see proof of Proposition 4.2 for further detailed arguments. Similarly, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big(Y \,\big|\, \mathbf{W} = \bar{\mathbf{w}}_\mathbf{j}\big) &= \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \,|\, \mathbf{W} = \bar{\mathbf{w}}_\mathbf{j}]\big)^2 \,\Big|\, \mathbf{W} = \bar{\mathbf{w}}_\mathbf{j}\Big]\\ &= \mathbb{E}\Big[\big(B_0 + \bar{\mathbf{w}}_\mathbf{j}^\top \mathbf{B} - \mathbb{E}[B_0] - \bar{\mathbf{w}}_\mathbf{j}^\top \mathbb{E}[\mathbf{B}]\big)^2\Big]\\ &= \sigma_1^2 + (\bar{\mathbf{w}}_\mathbf{j}{}^2)^\top \left(\sigma_2^2, \dots, \sigma_p^2\right)^\top\end{aligned}$$

for $j = 1, \dots, p$, because the covariances of the coefficients are assumed to be equal to zero. The linear system has design matrix

$$\begin{bmatrix} 1 & (\bar{\mathbf{w}}_\mathbf{1}{}^2)^\top \\ \vdots & \vdots \\ 1 & (\bar{\mathbf{w}}_\mathbf{p}{}^2)^\top \end{bmatrix},$$

which is also of full rank by Lemma 4.4. Hence also the variances $\sigma_1^2, \dots, \sigma_p^2$ are identified. □

4.3.2. Proofs for Section 4.2

Proof of Theorem 4.7. Suppose that S is of full rank. Since S contains the matrix

$$\begin{bmatrix} 1 & 2\mathbf{w_1}^\top \\ \vdots & \vdots \\ 1 & 2\mathbf{w}^\top_{\mathbf{\frac{p(p+1)}{2}}} \end{bmatrix} \in \mathbb{R}^{\frac{p(p+1)}{2}\times p}$$

as a submatrix, in order for S to have full rank, it is necessary that this submatrix has rank p. This implies that there are p points among the support points $\mathbf{w_1}, \ldots, \mathbf{w_{\frac{p(p+1)}{2}}}$ in general position, thus, identifying the means by Proposition 4.5. Then the linear system which determines $\mathbb{V}\mathrm{ar}(Y \,|\, \mathbf{W} = \mathbf{w_j})$, $j = 1, \ldots, p(p+1)/2$, in terms of the entries of Σ has design matrix S, see (4.4). Thus, identification of Σ from the conditional variances follows by the nonsingularity of S.

Conversely, let $m = p(p+1)/2$ and suppose that the condition of Theorem 4.7 is not satisfied. Then all support points $\mathbf{w}$ of $\mathbf{W}$ are such that the vectors $\mathrm{v}\big((1, \mathbf{w}^\top)^\top\big)$ are contained in a $(m-1)$-dimensional linear subspace V of $\mathbb{R}^m$. The $(p \times p)$-dimensional positive semi-definite matrices form a convex cone in the space of all $(p \times p)$-dimensional symmetric matrices. Its interior consists of all positive definite matrices and hence the image under the map vec is a convex cone $\mathcal{C} \subset \mathbb{R}^m$ with non-empty interior.

Let $\mathbf{z}$ be a unit vector orthogonal to V, and let Z be the $(p \times p)$-dimensional symmetric matrix for which $\mathrm{vec}(Z) = \mathbf{z}$. Since the positive definite matrices are open in the space of all $(p \times p)$-dimensional symmetric matrices, given a positive definite matrix Σ, for small $\epsilon > 0$ the matrix $\Sigma_1 = \Sigma + \epsilon Z$ will still be positive definite, and hence a covariance matrix. Moreover, it is $\mathrm{vec}(\Sigma_1) = \mathrm{vec}(\Sigma) + \epsilon\, \mathrm{vec}(Z) = \mathrm{vec}(\Sigma) + \epsilon\, \mathbf{z}$ and $(1, \mathbf{w}^\top)\, Z\, (1, \mathbf{w}^\top)^\top = \mathrm{v}\big((1, \mathbf{w}^\top)^\top\big)^\top \mathbf{z} = 0$ for $\mathbf{w}$ in the support of $\mathbf{W}$ by construction. Hence the conditional variances $(1, \mathbf{w}^\top)\, \Sigma\, (1, \mathbf{w}^\top)^\top$ and $(1, \mathbf{w}^\top)\, \Sigma_1\, (1, \mathbf{w}^\top)^\top$ will be the same over the support of $\mathbf{W}$. Thus, for normally distributed $\mathbf{A} \sim \mathcal{N}_p(\mathbf{0}_p, \Sigma)$ or $\mathbf{A} \sim \mathcal{N}_p(\mathbf{0}_p, \Sigma_1)$, the conditional normal distributions of $Y \,|\, \mathbf{W} = \mathbf{w}$ will coincide, showing nonidentifiability of the covariance matrix. □

For the proof of Proposition 4.10 we require two lemmas.

Lemma 4.12. *Suppose that $\Sigma \in \mathbb{R}^{p \times p}$ is a symmetric matrix and $\mathbf{v_1}, \ldots, \mathbf{v_p} \in \mathbb{R}^p$ form a known basis of $\mathbb{R}^p$. If $\mathbf{v} \in \mathbb{R}^p$ and $\mathbf{v}^\top \Sigma\, \mathbf{v_j}$, $1 \le j \le p$, is identified, then $\mathbf{v}^\top \Sigma\, \mathbf{u}$ is identified for any vector $\mathbf{u} \in \mathbb{R}^p$. In particular, Σ is identified from the values $\mathbf{v_j}^\top \Sigma\, \mathbf{v_k}$, $1 \le j \le k \le p$.*

Proof of Lemma 4.12. Given $\mathbf{u} \in \mathbb{R}^p$ we may write $\mathbf{u} = \sum_{j=1}^p \lambda_j \mathbf{v_j}$ with $\lambda_1, \ldots, \lambda_p \in \mathbb{R}$. Then

$$\mathbf{v}^\top \Sigma\, \mathbf{u} = \sum_{j=1}^p \lambda_j \mathbf{v}^\top \Sigma\, \mathbf{v_j}\,,$$

showing the first claim. For the second, let e_k denote the k^{th} unit vector in $\mathbb{R}^p$. By assumption, one may write $e_k = \sum_{j=1}^p \lambda_{k,j}\mathbf{v_j}$, where $\lambda_{j,k} \in \mathbb{R}$. Then

$$\Sigma_{kl} = e_k^\top \Sigma\, e_l = \sum_{j_1,j_2=1}^{p} \lambda_{k,j_1}\lambda_{l,j_2}\mathbf{v}_{\mathbf{j_1}}^\top \Sigma\, \mathbf{v}_{\mathbf{j_2}}\,.$$

The result follows from the assumptions and the symmetry of Σ. □

Lemma 4.13. *Let $\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3} \in \mathbb{R}^p$ be such that each pair is linearly independent, but all three are linearly dependent, so that $\mathbf{v_3} = \lambda_1\mathbf{v_1} + \lambda_2\mathbf{v_2}$, where $\lambda_1, \lambda_2 \in \mathbb{R}\setminus\{0\}$. Then for a symmetric matrix $\Sigma \in \mathbb{R}^{p\times p}$ it holds that*

$$\mathbf{v_1}^\top \Sigma\, \mathbf{v_2} = \frac{1}{2\,\lambda_1\lambda_2}\Big(\mathbf{v_3}^\top \Sigma\, \mathbf{v_3} - \lambda_1^2\, \mathbf{v_1}^\top \Sigma\, \mathbf{v_1} - \lambda_2^2\, \mathbf{v_2}^\top \Sigma\, \mathbf{v_2}\Big)\,.$$

Proof of Lemma 4.13. Plug in the expression for $\mathbf{v_3}$ and compute the right side of the equation. □

Proof of Proposition 4.10. By Proposition 4.5 and the first assumption the means μ are identified. Hence we obtain the equation (4.4) with $\mathbf{w}$ ranging over the support points mentioned in the statement of the proposition. To show that the design matrix S in (4.6) formed from all the points $\mathbf{w_j}, \mathbf{w_{j,k}}$ and $\mathbf{z_{j,k}}$ has full rank $p(p+1)/2$, it suffices to show that from these equations one can uniquely solve for σ. To this end, from the second assumption, for $j \in \{1,\dots,p\}$ and $k \in \{1,\dots,p-1\}$, letting $\mathbf{v_1} = (1, \mathbf{w_j}^\top)^\top$, $\mathbf{v_2} = (1, \mathbf{w_{j,k}}^\top)^\top$ and $\mathbf{v_3} = (1, \mathbf{z_{j,k}}^\top)^\top$ in Lemma 4.13 we identify $(1, \mathbf{w_j}^\top)\,\Sigma\,(1, \mathbf{w_{j,k}}^\top)^\top$. Since $(1, \mathbf{w_j}^\top)\,\Sigma\,(1, \mathbf{w_j})$ is also identified, from the first part in Lemma 4.12 we identify $(1, \mathbf{w_j}^\top)\,\Sigma\,(1, \mathbf{w_l}^\top)^\top$, $j, l \in \{1,\dots,p\}$. Hence from the second part of that lemma and the first assumption of Proposition 4.10 together with Lemma 4.4 the entries σ of the covariance matrix Σ itself are identified. □

Proof of Theorem 4.11. For the sufficiency, suppose that the support of W_j contains $\{w_{j,k},\ k = 1,2,3\}$, $j = 1,\dots,p-1$. We apply Proposition 4.10 with

- $\mathbf{w_j} = (w_{1,1},\dots,w_{j-1,1}, w_{j,2}, w_{j+1,1},\dots,w_{p-1,1})^\top$, $j \in \{1,\dots,p-1\}$, and $\mathbf{w_p} = (w_{1,1},\dots,w_{p-1,1})^\top$,
- for $j \in \{1,\dots,p-1\}$ let $\mathbf{w_{j,k}}$, $k \in \{1,\dots,p-1\}$, $k \neq j$, enumerate the points having k^{th} coordinate $w_{k,2}$ and j^{th} coordinate $w_{j,2}$, otherwise coordinates $w_{i,1}$, the corresponding $\mathbf{z_{j,k}}$ having k^{th} coordinate $w_{k,3}$, j^{th} coordinate $w_{j,2}$, otherwise coordinates $w_{i,1}$. Furthermore, let $\mathbf{w_{j,j}} = \mathbf{w_p}$ and let $\mathbf{z_{j,j}}$ have j^{th} coordinate $w_{j,3}$, otherwise $w_{i,1}$,
- let $\mathbf{w_{p,k}} = \mathbf{w_k}$, $k \in \{1,\dots,p-1\}$, and $\mathbf{z_{p,k}} = (w_{1,1},\dots,w_{j-1,1}, w_{j,3}, w_{j+1,1},\dots, w_{p-1,1})^\top$.

The requirements of the proposition are then easily checked by applying Lemma 4.4, 1. By Proposition 4.10 it is also obvious that we find $p(p+1)/2$ points among all the points $\mathbf{w_j}, \mathbf{w_{j,k}}$ and $\mathbf{z_{j,k}}$ so that the consequent design matrix S in (4.6) of these points has full rank. The necessity of at least three support points in each coordinate, if Σ has full rank, is clear from Example 4.9. □

5. Sign-consistency for a fixed number of coefficients

Our main goal in this chapter is variable selection and estimation of the first and second moments for a fixed number p of random coefficients in model (2.6). For this purpose we consider the ordinary LASSO and the adaptive LASSO, which are introduced in Section 2.2. These estimators are very common and well-studied, cf. Tibshirani (1996), Zhao and Yu (2006), Zou (2006) and Wainwright (2009b).

The regression model of the first moments of the coefficients is actually a linear regression model with independent and heteroscedastic errors. Hence the associated results for variable selection are also provided in Wagener and Dette (2012), where the authors give asymptotics for bridge estimators and the adaptive LASSO under heteroscedasticity and fixed designs. See also Knight and Fu (2000), Zou (2006) and Zhao and Yu (2006) for the homoscedastic case and a discussion about random designs. However, in the linear regression model of the variances and covariances of the coefficients the errors have a variant form since the appropriate response variable involves also the estimation error from the first stage mean regression. This leads for example to correlated errors.

Consider the linear regression model (2.6) with square-integrable coefficients. In the following we want to estimate their mean vector and covariance matrix, hence we denote by

$$\mu^* := \mathbb{E}[\mathbf{A}] \quad \in \mathbb{R}^p$$

and

$$\Sigma^* := \mathbb{C}\mathrm{ov}(\mathbf{A}) \quad \in \mathbb{R}^{p\times p}\,, \qquad \sigma^* := \mathrm{vec}(\Sigma^*) \quad \in \mathbb{R}^{\frac{p(p+1)}{2}} \tag{5.1}$$

the true moments. Moreover, let

$$\begin{aligned} S_\mu &:= \mathrm{supp}(\mu^*) = \Big\{k \in \{1,\dots,p\} \,\Big|\, \mu_k^* \neq 0\Big\}\,,\\ S_\sigma &:= \mathrm{supp}\big(\sigma^*\big) = \left\{k \in \left\{1,\dots,\frac{p(p+1)}{2}\right\} \,\middle|\, \sigma_k^* \neq 0\right\} \end{aligned}$$

be the supports of the mean vector μ^* and the half-vectorization σ^* of the covariance matrix. In addition, we denote the cardinalities of these sets by

$$s_\mu := |S_\mu|\,, \qquad s_\sigma := |S_\sigma|$$

and the relative complements by

$$S_\mu^c := \{1, \dots, p\} \setminus S_\mu \,, \qquad S_\sigma^c := \left\{1, \dots, \frac{p(p+1)}{2}\right\} \setminus S_\sigma \,.$$

Throughout this chapter we allow the mean vector and the covariance matrix to be sparse. This means that the number s_μ of coefficients with mean unequal to zero can be smaller than the number p of coefficients, and that the number s_σ of variances and covariances unequal to zero can be smaller than $p(p+1)/2$. The sparsity implies the following points.

- There can be deterministic coefficients in the linear random coefficient regression model (2.6). If $k \in \{1, \dots, p\}$ is an element of S_σ^c, then $\mathbb{V}\mathrm{ar}(A_k) = 0$ and hence A_k is constant (almost surely). Note that this coefficient is then also uncorrelated with the other coefficients A_l, $l \in \{1, \dots, p\}, l \neq k$. As a consequence, the k^{th} column and row of Σ^* and the corresponding entries of σ^* are equal to zero as well, and hence elements of S_σ^c.

- Some of the coefficients can be equal to zero. If $k \in \{1, \dots, p\}$ is an element of S_μ^c and S_σ^c, then $\mathbb{E}[A_k] = \mathbb{V}\mathrm{ar}(A_k) = 0$ and hence $A_k = 0$ (almost surely).

- There can be uncorrelated coefficients. If $k \in \{p+1, \dots, p(p+1)/2\}$ is an element of S_σ^c, then $\mathbb{C}\mathrm{ov}(A_l, A_{l'}) = 0$ for some $l, l' \in \{1, \dots, p\}$.

This chapter is structured as follows. In Section 5.1 we introduce the linear regression model of the first moments of the random coefficients and give asymptotic results for a growing number of observations. In particular, we provide sign-consistency and asymptotic normality of the adaptive LASSO. In Section 5.2 we proceed analogously for the variances and covariances of the coefficients. The main steps of the proofs are given in Section 5.4 and in Section 5.3 we report on results of appropriate numerical experiments. Finally, Section 5.5 provides the technical proofs.

5.1. First moments

5.1.1. Regression model and estimator

We observe independent random vectors $(Y_1, \mathbf{X_1}^\top)^\top, \dots, (Y_n, \mathbf{X_n}^\top)^\top$ distributed according to the linear random coefficient regression model (2.9), and write

$$Y_i = \mathbf{X_i}^\top \mu^* + \mathbf{X_i}^\top \big(\mathbf{A_i} - \mu^*\big) \,, \quad i = 1, \dots, n \,, \tag{5.2}$$

where $\mathbf{X_i} = (1, \mathbf{W_i}^\top)^\top \in \mathbb{R}^p$ with $\mathbf{W_i} \sim \mathbf{W} \in \mathbb{R}^{p-1}$ and $\mathbf{A_i} = (B_{i,0}, \mathbf{B_i}^\top)^\top \sim \mathbf{A} \in \mathbb{R}^p$ are independent random vectors. Here $\mathbf{X_i} = (X_{i,1}, \dots, X_{i,p})^\top$ represents the observed individual covariates and $\mathbf{A_i} = (A_{i,1}, \dots, A_{i,p})^\top$ the unobserved individual regression coefficients. Moreover, note that the deterministic parameter vector μ^* is s_μ-sparse and the errors are evidently heteroscedastic. We define

$$\begin{aligned} \mathbb{Y}_n^\mu &:= \big(Y_1, \dots, Y_n\big)^\top && \in \mathbb{R}^n\,, \\ \mathbb{X}_n^\mu &:= \big[\mathbf{X_1}, \dots, \mathbf{X_n}\big]^\top && \in \mathbb{R}^{n\times p}\,, \\ \varepsilon_n^\mu &:= \Big(\mathbf{X_1}^\top\big(\mathbf{A_1} - \mu^*\big), \dots, \mathbf{X_n}^\top\big(\mathbf{A_n} - \mu^*\big)\Big)^\top && \in \mathbb{R}^n\,, \end{aligned}$$

and hence model (5.2) can be written in matrix form as

$$\mathbb{Y}_n^\mu = \mathbb{X}_n^\mu\,\mu^* + \varepsilon_n^\mu = \mathbb{X}_{n,S_\mu}^\mu\,\mu_{S_\mu}^* + \varepsilon_n^\mu\,.$$

The entries of ε_n^μ are pairwise independent and identically distributed since $(\mathbf{X_1}^\top, \mathbf{A_1}^\top)^\top$, $\dots, (\mathbf{X_n}^\top, \mathbf{A_n}^\top)^\top$ are independent and identically distributed. The conditional mean of the error vector ε_n^μ regarding the regressors is given by

$$\mathbb{E}\big[\varepsilon_n^\mu \,\big|\, \mathbb{X}_n^\mu\big] = \Big(\mathbf{X_1}^\top\big(\mathbb{E}[\mathbf{A_1}] - \mu^*\big), \dots, \mathbf{X_n}^\top\big(\mathbb{E}[\mathbf{A_n}] - \mu^*\big)\Big)^\top = \mathbf{0}_n\,,$$

and the conditional variances are

$$\mathbb{V}\mathrm{ar}\Big(e_i^\top\varepsilon_n^\mu \,\Big|\, \mathbf{X_i}\Big) = \mathbb{V}\mathrm{ar}\Big(\mathbf{X_i}^\top\big(\mathbf{A_i} - \mu^*\big) \,\Big|\, \mathbf{X_i}\Big) = \sum_{k,l=1}^{p} X_{i,k}X_{i,l}\,\mathrm{Cov}\big(A_{i,k}, A_{i,l}\big) = \mathbf{X_i}^\top\Sigma^*\,\mathbf{X_i}\,.$$

Hence we obtain the conditional covariance matrix

$$\Omega_n^\mu := \mathbb{C}\mathrm{ov}\big(\varepsilon_n^\mu \,\big|\, \mathbb{X}_n^\mu\big) = \mathrm{diag}\big(\mathbf{X_1}^\top\Sigma^*\,\mathbf{X_1}, \dots, \mathbf{X_n}^\top\Sigma^*\,\mathbf{X_n}\big)\,. \tag{5.3}$$

Remark 5.1. The matrix Ω_n^μ is obviously positive definite if the values $\mathbf{X_1}^\top\Sigma^*\,\mathbf{X_1}, \dots,$ $\mathbf{X_n}^\top\Sigma^*\,\mathbf{X_n}$ are positive. That is generally given if the covariance matrix Σ^* of the random coefficients is positive definite and $\mathbf{X_1}, \dots, \mathbf{X_n} \neq \mathbf{0}_p$. However, we allow also sparsity of the second central moments of the coefficients, and in consequence the covariance matrix can be positive semi-definite. Especially, as we mentioned at the beginning of this chapter, if some of the coefficients are deterministic, the covariance matrix has rows and columns with zeros and in consequence no full rank. In that case the matrix can be expressed by

$$\Sigma^* = \begin{bmatrix} \Sigma_1^* & \mathbf{0}_{d\times(p-d)} \\ \mathbf{0}_{(p-d)\times d} & \mathbf{0}_{(p-d)\times(p-d)} \end{bmatrix}$$

with $\Sigma_1^* \in \mathbb{R}^{d\times d}$, $d \leq p$ (change the order of the coefficients and the associated explanatory variables to get the block form). So, if we assume Σ_1^* to be positive definite and the corresponding covariates $(X_{1,1}, \dots, X_{1,d})^\top, \dots, (X_{n,1}, \dots, X_{n,d})^\top \neq \mathbf{0}_d$, then the conditional covariance matrix Ω_n^μ of the errors ε_n^μ is still positive definite.

Consider the linear regression model (5.2) of the first moments. The appropriate empirical quadratic loss function is given by

$$\mathcal{L}_{\mu,n}^{\mathrm{LS}}\big(\beta\big) := \frac{1}{n}\,\|\mathbb{Y}_n^{\mu} - \mathbb{X}_n^{\mu}\,\beta\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \mathbf{X_i}^{\top}\beta\big)^2\,, \qquad \beta \in \mathbb{R}^p\,.$$

Then we define the least squares estimator by

$$\widehat{\mu}_n^{\mathrm{LS}} \in \rho_{\mu,n}^{\mathrm{LS}} := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p}\ \mathcal{L}_{\mu,n}^{\mathrm{LS}}\big(\beta\big)\,, \tag{5.4}$$

the ordinary LASSO with regularization parameter $\lambda_n^{\mu} > 0$ by

$$\widehat{\mu}_n^{\mathrm{L}} \in \rho_{\mu,n,\lambda_n^{\mu}}^{\mathrm{L}} := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p}\ \Big(\mathcal{L}_{\mu,n}^{\mathrm{LS}}\big(\beta\big) + 2\lambda_n^{\mu}\,\|\beta\|_1\Big)\,, \tag{5.5}$$

and the adaptive LASSO by

$$\widehat{\mu}_n^{\mathrm{AL}} \in \rho_{\mu,n,\lambda_n^{\mu}}^{\mathrm{AL}} := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p}\ \left(\mathcal{L}_{\mu,n}^{\mathrm{LS}}\big(\beta\big) + 2\lambda_n^{\mu}\sum_{k=1}^{p}\frac{|\beta_k|}{\big|\widehat{\mu}_{n,k}^{\mathrm{init}}\big|}\right)\,, \tag{5.6}$$

where $\widehat{\mu}_n^{\mathrm{init}} = (\widehat{\mu}_{n,1}^{\mathrm{init}}, \dots, \widehat{\mu}_{n,p}^{\mathrm{init}})^{\top} \in \mathbb{R}^p$ is an initial estimator of μ^*. If $\widehat{\mu}_{n,k}^{\mathrm{init}} = 0$, we require $\beta_k = 0$, and, moreover, the sequence $\big(\widehat{\mu}_n^{\mathrm{init}}\big)_{n\in\mathbb{N}} \subset \mathbb{R}^p$ should be consistent with respect to μ^*.

5.1.2. Asymptotic results

We need some assumptions to prove asymptotic results for the LASSO and adaptive LASSO in the linear regression model (5.2) of the first moments.

Assumption 5.2. We assume that

(A1) the random coefficients $\mathbf{A}$ have finite second moments,

(A2) the covariates $\mathbf{X} = (1, \mathbf{W}^{\top})^{\top}$ (or rather $\mathbf{W}$) have finite fourth moments,

(A3) the symmetric second moment matrix

$$\mathrm{C}^{\mu} := \mathbb{E}\big[\mathbf{X}\,\mathbf{X}^{\top}\big] \quad \in \mathbb{R}^{p\times p}$$

of the covariates is positive definite.

The third assumption (A3) is connected to the identifiability of the first moments, discussed in Proposition 4.5. The proof of the following lemma is provided in Section 5.5.1.

Lemma 5.3. *If the support of* $\mathbf{W}$ *contains* p *points* $\mathbf{w_1}, \dots, \mathbf{w_p} \in \mathbb{R}^{p-1}$ *in general position, then* C^{μ} *is positive definite.*

Now we can give sufficient conditions for the sign-consistency of the LASSO. The proof is deferred to Section 5.4.

Theorem 5.4 (Asymptotics LASSO of the means)**.** *Consider the linear regression model* (5.2) *and suppose that Assumption 5.2 as well as the mutual incoherence condition*

$$\left\| \mathrm{C}^{\mu}_{S^c_\mu S_\mu} \left(\mathrm{C}^{\mu}_{S_\mu S_\mu}\right)^{-1} \right\|_{\mathrm{M},\infty} < 1 \tag{5.7}$$

are satisfied. If in addition $\lambda^\mu_n \to 0$ and $\sqrt{n}\,\lambda^\mu_n \to \infty$ hold, the LASSO $\widehat{\mu}^{\mathrm{L}}_n$ as a solution to (5.5) *is sign-consistent,*

$$\mathbb{P}\Big(\operatorname{sign}\big(\widehat{\mu}^{\mathrm{L}}_n\big) = \operatorname{sign}\big(\mu^*\big)\Big) \to 1\,,$$

and has estimation rate λ^μ_n on the support S_μ of μ^*, that is

$$\frac{1}{\lambda^\mu_n}\left(\widehat{\mu}^{\mathrm{L}}_{n,S_\mu} - \mu^*_{S_\mu}\right) = \mathcal{O}_{\mathbb{P}}\left(1\right).$$

Remark 5.5. The mutual incoherence condition (5.7) is crucial for the sign-consistency of the LASSO, cf. Zou (2006) and Wainwright (2009b). The assumption can be dropped if the adaptive LASSO is used instead. Moreover, this estimator enjoys additionally the oracle properties under homoscedasticity (Zou, 2006). That includes the selection of the true model by the estimator and an optimal estimation rate of $1/\sqrt{n}$. Theorem 5.4 shows that the LASSO satisfies only one of these properties, if the regularization parameter λ^μ_n is chosen as in the mentioned theorem, since λ^μ_n tends slower to zero than the square root of the number n of observations to infinity.

To achieve sign-consistency and an optimal estimation rate of the adaptive LASSO also assumptions on the regularization parameter λ^μ_n are necessary. However, they depend on the estimation rate of the initial estimator. Thus, we assume in the following that the initial estimator of the adaptive LASSO achieves already an optimal rate of $1/\sqrt{n}$. Furthermore, for the asymptotic normality on the true support S_μ we define in addition the matrix

$$\mathrm{B}^\mu := \mathbb{E}\Big[\big(\mathbf{X}^\top \Sigma^*\, \mathbf{X}\big)\, \mathbf{X}\, \mathbf{X}^\top\Big] \quad \in \mathbb{R}^{p\times p}\,. \tag{5.8}$$

It is evidently symmetric and positive semi-definite because of Assumption (A3) and the positive semi-definiteness of the covariance matrix Σ^*. The proof of the following theorem is deferred to Section 5.4 as well.

Theorem 5.6 (Asymptotics adaptive LASSO of the means)**.** *Consider the linear regression model* (5.2) *and let Assumption 5.2 be satisfied. In addition, suppose that the initial estimator $\widehat{\mu}^{\mathrm{init}}_n$ of the means μ^* of the random coefficients satisfies $\sqrt{n}\left(\widehat{\mu}^{\mathrm{init}}_n - \mu^*\right) = \mathcal{O}_{\mathbb{P}}\left(1\right)$, and that the regularization parameter λ^μ_n satisfies $\lambda^\mu_n \to 0$, $\sqrt{n}\,\lambda^\mu_n \to 0$ and $n\,\lambda^\mu_n \to \infty$. Then the adaptive LASSO $\widehat{\mu}^{\mathrm{AL}}_n$ as a solution to* (5.6) *is sign-consistent,*

$$\mathbb{P}\Big(\operatorname{sign}\big(\widehat{\mu}^{\mathrm{AL}}_n\big) = \operatorname{sign}\big(\mu^*\big)\Big) \to 1\,,$$

and satisfies the asymptotic normality

$$\sqrt{n}\left(\widehat{\mu}^{\mathrm{AL}}_{n,S_\mu} - \mu^*_{S_\mu}\right) \xrightarrow{d} \mathcal{N}_{s_\mu}\Big(\mathbf{0}_{s_\mu}, \left(\mathrm{C}^{\mu}_{S_\mu S_\mu}\right)^{-1} \mathrm{B}^{\mu}_{S_\mu S_\mu} \left(\mathrm{C}^{\mu}_{S_\mu S_\mu}\right)^{-1}\Big) \tag{5.9}$$

on the support S_μ of μ^*.

Remark 5.7. The asymptotic covariance matrix in (5.9) can be degenerate since we have no further assumption on B^{μ}. If $\mathrm{B}^{\mu}_{S_\mu S_\mu}$ is positive definite, then the asymptotic covariance matrix of the rescaled estimation error of the adaptive LASSO is positive definite as well. Moreover, we obtain also

$$\sqrt{n}\left(\widehat{\mu}_n^{\mathrm{LS}} - \mu^*\right) \xrightarrow{d} \mathcal{N}_p\Big(\mathbf{0}_p, \left(\mathrm{C}^{\mu}\right)^{-1}\mathrm{B}^{\mu}\left(\mathrm{C}^{\mu}\right)^{-1}\Big) \tag{5.10}$$

under Assumption 5.2, see Remark 5.28 in Section 5.4 for more details. This shows that the least squares estimator and the adaptive LASSO have on the true support S_μ the same asymptotic normality. Additionally, by (5.10) it follows that the least squares estimator can be used as initial estimator in Theorem 5.6. The LASSO $\widehat{\mu}_n^{\mathrm{L}}$ with an appropriate choice for its regularization parameter and the Ridge estimator,

$$\widehat{\mu}_n^{\mathrm{Ridge}} \in \rho^{\mathrm{Ridge}}_{\mu,n,\lambda_n^{\mu}} := \underset{\beta\in\mathbb{R}^p}{\arg\min}\left(\mathcal{L}^{\mathrm{LS}}_{\mu,n}(\beta) + \lambda_n^{\mu}\left\|\beta\right\|_2^2\right),$$

are also potential choices as initial estimator, cf. Knight and Fu (2000) and Wagener and Dette (2012).

5.2. Second central moments

5.2.1. Regression model and estimator

At first we want to point out that for the estimation of the variances and covariances of the random coefficients knowledge about their means μ^* is crucial. Hence we have to proceed in a two-step procedure. Firstly, we determine a consistent estimator $\widehat{\mu}_n$ of μ^* based on the observations $(Y_1, \mathbf{X_1}^\top)^\top, \ldots, (Y_n, \mathbf{X_n}^\top)^\top$ and with the help of the linear regression model (5.2). Then we consider the regression residuals

$$\widetilde{Y}_i := Y_i - \mathbf{X_i}^\top\widehat{\mu}_n\,, \quad i = 1, \ldots, n\,.$$

Note that these variables are also observable since they only depend on $(Y_1, \mathbf{X_1}^\top)^\top, \ldots,$ $(Y_n, \mathbf{X_n}^\top)^\top$. Moreover, the linear regression model (5.2) implies

$$\widetilde{Y}_i = \mathbf{X_i}^\top\mu^* + \mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right) - \mathbf{X_i}^\top\widehat{\mu}_n = \mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right) + \mathbf{X_i}^\top\left(\mu^* - \widehat{\mu}_n\right).$$

Henceforth we denote by

$$Y_i^{\sigma} := \widetilde{Y}_i^2 = \Big(\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\Big)^2 + \Big(\mathbf{X_i}^\top\left(\mu^* - \widehat{\mu}_n\right)\Big)^2 + 2\,\mathbf{X_i}^\top\left(\mathbf{A_i} - \mu^*\right)\mathbf{X_i}^\top\left(\mu^* - \widehat{\mu}_n\right),$$

$i = 1, \ldots, n$, the squared residuals. These include among other terms the products of the centered coefficients and hence they are the response variables in the linear regression model of the variances and covariances. Expansion of the products and rearranging leads to

$$\begin{aligned}
Y_i^\sigma &= \mathbf{X_i}^\top \Big(\big(\mathbf{A_i} - \mu^*\big)\big(\mathbf{A_i} - \mu^*\big)^\top + \big(\mu^* - \widehat{\mu}_n\big)\big(\mu^* - \widehat{\mu}_n\big)^\top + 2\,\big(\mathbf{A_i} - \mu^*\big)\big(\mu^* - \widehat{\mu}_n\big)^\top \Big)\mathbf{X_i} \\
&= \mathbf{X_i}^\top \Big(\big(\mathbf{A_i} - \mu^*\big)\big(\mathbf{A_i} - \mu^*\big)^\top + \big(\mu^* - \widehat{\mu}_n\big)\big(\mu^* - \widehat{\mu}_n\big)^\top \\
&\qquad\qquad + \big(\mathbf{A_i} - \mu^*\big)\big(\mu^* - \widehat{\mu}_n\big)^\top + \big(\mu^* - \widehat{\mu}_n\big)\big(\mathbf{A_i} - \mu^*\big)^\top \Big)\mathbf{X_i}\,.
\end{aligned}$$

Evidently, the dependent variables are a quadratic form, which is determined by the matrices in brackets, in the explanatory variables. We define the symmetric matrices

$$D_i := \big(\mathbf{A_i} - \mu^*\big)\big(\mathbf{A_i} - \mu^*\big)^\top \qquad \in \mathbb{R}^{p\times p}\,, \tag{5.11}$$

$$E_n := \big(\mu^* - \widehat{\mu}_n\big)\big(\mu^* - \widehat{\mu}_n\big)^\top \qquad \in \mathbb{R}^{p\times p}\,, \tag{5.12}$$

$$F_{n,i} := \big(\mathbf{A_i} - \mu^*\big)\big(\mu^* - \widehat{\mu}_n\big)^\top + \big(\mu^* - \widehat{\mu}_n\big)\big(\mathbf{A_i} - \mu^*\big)^\top \qquad \in \mathbb{R}^{p\times p} \tag{5.13}$$

for $i \in \{1,\dots,n\}$. The matrices $D_1,\dots,D_n$ contain the products of the centered individual coefficients. Hence $\mathbb{E}[D_i] = \Sigma^*$ holds since the coefficients $\mathbf{A_i}$ are identically distributed with covariance matrix Σ^*. Furthermore, E_n captures the (products of the) estimation error of the first stage mean regression and $F_{n,1},\dots,F_{n,n}$ contain the mixing products. With the above notation the response variables can be written as

$$Y_i^\sigma = \mathbf{X_i}^\top \Sigma^*\, \mathbf{X_i} + \mathbf{X_i}^\top \big(D_i - \Sigma^* + E_n + F_{n,i}\big)\mathbf{X_i}\,, \quad i = 1,\dots,n\,.$$

Now the first part of the sum on the right-hand side includes the variances and covariances in which we are interested. To get the common structure of a linear regression model we use the half-vectorization vec in (2.8) for symmetric matrices and the corresponding vector transformation v in (4.5). Then we obtain

$$Y_i^\sigma = \mathrm{v}\big(\mathbf{X_i}\big)^\top \sigma^* + \mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^* + E_n + F_{n,i}\big)\,, \quad i = 1,\dots,n\,, \tag{5.14}$$

where $\sigma^* = \mathrm{vec}(\Sigma^*)$. Note that the deterministic coefficient vector σ^* is s_σ-sparse. The errors are heteroscedastic and, moreover, they are not independent since they all depend on the estimate $\widehat{\mu}_n$ of the means μ^*. Let

$$\begin{aligned}
\mathbb{Y}_n^\sigma &:= \big(Y_1^\sigma,\dots,Y_n^\sigma\big)^\top && \in \mathbb{R}^n\,, \\
&= \Big(\big(Y_1 - \mathbf{X_1}^\top \widehat{\mu}_n\big)^2,\dots,\big(Y_n - \mathbf{X_n}^\top \widehat{\mu}_n\big)^2\Big)^\top\,, \\
\mathbb{X}_n^\sigma &:= \Big[\mathrm{v}\big(\mathbf{X_1}\big),\dots,\mathrm{v}\big(\mathbf{X_n}\big)\Big]^\top && \in \mathbb{R}^{n\times \frac{p(p+1)}{2}}\,, \\
\varepsilon_n^\sigma &:= \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(D_1 - \Sigma^* + E_n + F_{n,1}\big),\dots, && \qquad (5.15) \\
&\qquad\qquad \mathrm{v}\big(\mathbf{X_n}\big)^\top \mathrm{vec}\big(D_n - \Sigma^* + E_n + F_{n,n}\big)\Big)^\top && \in \mathbb{R}^n\,, \\
&= \Big(\mathbf{X_1}^\top \big(D_1 - \Sigma^* + E_n + F_{n,1}\big)\mathbf{X_1},\dots, \\
&\qquad\qquad \mathbf{X_n}^\top \big(D_n - \Sigma^* + E_n + F_{n,n}\big)\mathbf{X_n}\Big)^\top\,,
\end{aligned}$$

then the linear regression model in (5.14) can be written in matrix notation as

$$\mathbb{Y}_n^\sigma = \mathbb{X}_n^\sigma \, \sigma^* + \varepsilon_n^\sigma = \mathbb{X}_{n,S_\sigma}^\sigma \, \sigma_{S_\sigma}^* + \varepsilon_n^\sigma \,. \tag{5.16}$$

We define again the appropriate empirical quadratic loss function

$$\mathcal{L}_{\sigma,n}^{\mathrm{LS}}\big(\beta\big) := \frac{1}{n} \, \|\mathbb{Y}_n^\sigma - \mathbb{X}_n^\sigma \, \beta\|_2^2 = \frac{1}{n} \sum_{i=1}^n \Big(\big(Y_i - \mathbf{X_i}^\top \widehat{\mu}_n\big)^2 - \mathrm{v}\big(\mathbf{X_i}\big)^\top \beta \Big)^2 \,, \quad \beta \in \mathbb{R}^{\frac{p(p+1)}{2}} \,,$$

where $\widehat{\mu}_n$ is a consistent estimator of the first moments μ^* of the random coefficients based on the observations $(Y_1, \mathbf{X_1}^\top)^\top, \ldots, (Y_n, \mathbf{X_n}^\top)^\top$.

The least squares estimator of the variances and covariances is given analogously to Section 5.1.1 by

$$\widehat{\sigma}_n^{\mathrm{LS}} \in \rho_{\sigma,n}^{\mathrm{LS}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{\frac{p(p+1)}{2}}} \; \mathcal{L}_{\sigma,n}^{\mathrm{LS}}\big(\beta\big) \,, \tag{5.17}$$

the ordinary LASSO with regularization parameter $\lambda_n^\sigma > 0$ by

$$\widehat{\sigma}_n^{\mathrm{L}} \in \rho_{\sigma,n,\lambda_n^\sigma}^{\mathrm{L}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{\frac{p(p+1)}{2}}} \; \Big(\mathcal{L}_{\sigma,n}^{\mathrm{LS}}\big(\beta\big) + 2\lambda_n^\sigma \, \|\beta\|_1 \Big) \,, \tag{5.18}$$

and the adaptive LASSO by

$$\widehat{\sigma}_n^{\mathrm{AL}} \in \rho_{\sigma,n,\lambda_n^\sigma}^{\mathrm{AL}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{\frac{p(p+1)}{2}}} \; \left(\mathcal{L}_{\sigma,n}^{\mathrm{LS}}\big(\beta\big) + 2\lambda_n^\sigma \sum_{k=1}^{\frac{p(p+1)}{2}} \frac{|\beta_k|}{\big|\widehat{\sigma}_{n,k}^{\mathrm{init}}\big|} \right) \,, \tag{5.19}$$

where $\widehat{\sigma}_n^{\mathrm{init}} = (\widehat{\sigma}_{n,1}^{\mathrm{init}}, \ldots, \widehat{\sigma}_{n,p(p+1)/2}^{\mathrm{init}})^\top \in \mathbb{R}^{\frac{p(p+1)}{2}}$ is an initial estimator of the second central moments σ^*. If $\widehat{\sigma}_{n,k}^{\mathrm{init}} = 0$, we require again $\beta_k = 0$ in (5.19).

5.2.2. Structure of the regression errors

In this section we take a closer look at the errors in the linear regression model (5.14) of the variances and covariances of the random coefficients. In particular, we emphasize their structure if the first moments of the coefficients are known and do not have to be estimated in the first place.

Remark 5.8. The error vector ε_n^σ in (5.15) can be decomposed in three terms, namely

$$\varepsilon_n^\sigma = \delta_n + \zeta_n + \xi_n$$

with

$$\begin{aligned}
\delta_n &:= \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(D_1 - \Sigma^*\big), \ldots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \mathrm{vec}\big(D_n - \Sigma^*\big)\Big)^\top && \in \mathbb{R}^n\,, \qquad (5.20)\\
&= \Big(\mathbf{X_1}^\top (D_1 - \Sigma^*)\mathbf{X_1}, \ldots, \mathbf{X_n}^\top (D_n - \Sigma^*)\mathbf{X_n}\Big)^\top\,, \\
\zeta_n &:= \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(E_n\big), \ldots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \mathrm{vec}\big(E_n\big)\Big)^\top && \in \mathbb{R}^n\,, \qquad (5.21)\\
&= \Big(\mathbf{X_1}^\top E_1 \mathbf{X_1}, \ldots, \mathbf{X_n}^\top E_n \mathbf{X_n}\Big)^\top\,, \\
\xi_n &:= \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(F_{n,1}\big), \ldots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \mathrm{vec}\big(F_{n,n}\big)\Big)^\top && \in \mathbb{R}^n\,, \qquad (5.22)\\
&= \Big(\mathbf{X_1}^\top F_{n,1} \mathbf{X_1}, \ldots, \mathbf{X_n}^\top F_{n,n} \mathbf{X_n}\Big)^\top\,.
\end{aligned}$$

The matrices $D_1, \ldots, D_n, E_n, F_{n,1}, \ldots, F_{n,n}$ are defined in (5.11), (5.12) and (5.13). Note that the first term δ_n is unavoidable in the situation where we want to estimate the variances and covariances of the random coefficients. However, the error terms ζ_n and ξ_n occur because we have no knowledge about the first moments and have to estimate them simultaneously.

Remark 5.9. In the special case where the means μ^* of the random coefficients are known in advance, we can set $\widehat{\mu}_n = \mu^*$ in the linear regression model (5.14). As a consequence, the matrices E_n and $F_{1,n}, \ldots, F_{n,n}$ are equal to the null matrix and hence the model simplifies to

$$Y_i^\sigma = \mathrm{v}\big(\mathbf{X_i}\big)^\top \sigma^* + \mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^*\big)\,, \quad i = 1, \ldots, n\,,$$

respectively the error vector in (5.16) is $\varepsilon_n^\sigma = \delta_n$. In this setting the heteroscedastic errors are independent as well.

In the following lemma we provide the conditional first and second moments of the error vector δ_n which is not affected by the estimator error of the first stage mean regression. The corresponding proof is deferred to Section 5.5.2.

Lemma 5.10. *Assume that the random coefficients* $\mathbf{A}$ *have finite fourth moments. Then the vector* δ_n *in* (5.20) *satisfies* $\mathbb{E}\big[\delta_n \,\big|\, \mathbb{X}_n^\sigma\big] = \mathbf{0}_n$ *and*

$$\Omega_n^\sigma := \mathbb{C}\mathrm{ov}\big(\delta_n \,\big|\, \mathbb{X}_n^\sigma\big) = \mathrm{diag}\Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \Psi^*\, \mathrm{v}\big(\mathbf{X_1}\big), \ldots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \Psi^*\, \mathrm{v}\big(\mathbf{X_n}\big)\Big)\,, \qquad (5.23)$$

where

$$\begin{aligned}
\Psi^* := \Big[&\mathrm{vec}\big(\mathcal{M}^{11}\big), \ldots, \mathrm{vec}\big(\mathcal{M}^{pp}\big), \mathrm{vec}\big(\mathcal{M}^{12}\big), \ldots, \mathrm{vec}\big(\mathcal{M}^{1p}\big), \qquad (5.24)\\
&\mathrm{vec}\big(\mathcal{M}^{23}\big), \ldots, \mathrm{vec}\big(\mathcal{M}^{2p}\big), \ldots, \mathrm{vec}\big(\mathcal{M}^{(p-1)p}\big)\Big]^\top \quad \in \mathbb{R}^{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}}
\end{aligned}$$

with $\mathcal{M}^{kl} \in \mathbb{R}^{p\times p}$ and

$$\big(\mathcal{M}^{kl}\big)_{uv} = \mathbb{C}\mathrm{ov}\Big(\big(D_1\big)_{kl}, \big(D_1\big)_{uv}\Big) = \mathbb{C}\mathrm{ov}\Big(\big(A_k - \mu_k^*\big)\big(A_l - \mu_l^*\big), \big(A_u - \mu_u^*\big)\big(A_v - \mu_v^*\big)\Big) \tag{5.25}$$

holds. In particular, the entries of δ_n are independent as well.

Remark 5.11. The matrix Ψ^* is symmetric and contains the (mixed) fourth central moments of the random coefficients. Since we claimed in Lemma 5.10 the existence of the fourth moments, the Cauchy Schwarz inequality implies the well-definedness of $\mathcal{M}^{kl}$ for $k, l \in \{1, \dots, p\}$, and hence of Ψ^* as well. In addition, note that the matrix $\mathcal{M}^{kl}$ is symmetric, and that $\mathcal{M}^{kl}$ and $\mathcal{M}^{lk}$ are equal for $k, l \in \{1, \dots, p\}$. Hence no information about the fourth central moments is missing in the definition of Ψ^*. Moreover, let $\mathcal{K}^{kl}, \mathcal{K}^{k} \in \mathbb{R}^{p\times p}$ and $\kappa^{kl}, \kappa^k \in \mathbb{R}^p$ for $k, l \in \{1, \dots, p\}$ with

$$\begin{aligned}
\big(\mathcal{K}^{kl}\big)_{uv} &= \mathbb{C}\mathrm{ov}\big(A_k A_l, A_u A_v\big)\,,\\
\big(\mathcal{K}^{k}\big)_{uv} &= \mathbb{C}\mathrm{ov}\big(A_k, A_u A_v\big)\,,\\
\big(\kappa^{kl}\big)_{u} &= \mathbb{C}\mathrm{ov}\big(A_k A_l, A_u\big)\,,\\
\big(\kappa^{k}\big)_{u} &= \mathbb{C}\mathrm{ov}\big(A_k, A_u\big)\,,
\end{aligned}$$

then the properties of the covariance lead to

$$\begin{aligned}
\mathcal{M}^{kl} = \mathcal{K}^{kl} - \kappa^{kl}\big(\mu^*\big)^\top - \mu^*\big(\kappa^{kl}\big)^\top &+ \mu_l^*\Big(\kappa^k\big(\mu^*\big)^\top + \mu^*\big(\kappa^k\big)^\top - \mathcal{K}^k\Big)\\
&+ \mu_k^*\Big(\kappa^l\big(\mu^*\big)^\top + \mu^*\big(\kappa^l\big)^\top - \mathcal{K}^l\Big)\,.
\end{aligned}$$

In addition the proof of Lemma 5.10 shows that

$$\mathrm{v}\big(\mathbf{x}\big)^\top \Psi^* \,\mathrm{v}\big(\mathbf{x}\big) \geq 0 \tag{5.26}$$

is satisfied for all $\mathbf{x} \in \mathbb{R}^p$.

5.2.3. Asymptotic results

In the following we provide analogous results as in Theorem 5.4 and 5.6 in Section 5.1.2 for the half-vectorization σ^* of the covariance matrix Σ^* of the random coefficients. Note that the linear regression model (5.14) of the second central moments depends on the estimator $\widehat{\mu}_n$ of the first moments μ^* of the coefficients. For this purpose we choose throughout this section an estimator with estimation rate of $1/\sqrt{n}$, that is $\sqrt{n}\,\big(\widehat{\mu}_n - \mu^*\big) = \mathcal{O}_{\mathbb{P}}\,(1)$. Once again we need some assumptions for the results provided in this section.

Assumption 5.12. We assume that

(A4) the random coefficients $\mathbf{A}$ have finite fourth moments,

(A5) the covariates $\mathbf{X} = (1, \mathbf{W}^\top)^\top$ (or rather $\mathbf{W}$) have finite eighth moments, (A6) the symmetric matrix

$$\mathrm{C}^{\sigma} := \mathbb{E}\Big[\mathrm{v}\big(\mathbf{X}\big)\,\mathrm{v}\big(\mathbf{X}\big)^{\top}\Big] \quad \in \mathbb{R}^{\frac{p(p+1)}{2}\times\frac{p(p+1)}{2}}\,,$$

which contains the fourth moments of the covariates, is positive definite.

Remark 5.13. The matrix

$$\mathbb{E}\Big[(1, 2\,\mathbf{W}^{\top})^{\top}\,(1, 2\,\mathbf{W}^{\top})\Big] \quad \in \mathbb{R}^{p\times p}$$

is a submatrix of C^{σ} because of the definition of the vector transformation v in (4.5). If Assumption (A6) is satisfied, then also the above matrix has to be positive definite, which implies the positive definiteness of $\mathrm{C}^{\mu} = \mathbb{E}[\mathbf{X}\,\mathbf{X}^{\top}]$ as well. Hence, if Assumption 5.12 holds, also Assumption 5.2 is satisfied.

Furthermore, the critical Assumption (A6) is connected to the identification results in Section 4.2. The following lemma makes this precise and the corresponding proof is provided in Section 5.5.1.

Lemma 5.14. *Under the assumption of Theorem 4.11, that the support of the covariate vector* $\mathbf{W}$ *contains a Cartesian product with three distinct points in each coordinate, the matrix* C^{σ} *is positive definite.*

Now we give sufficient conditions, similar to Theorem 5.4, for the sign-consistency of the ordinary LASSO. The proof is deferred to Section 5.4.

Theorem 5.15 (Asymptotics LASSO of the variances and covariances)**.** *Consider the linear regression model* (5.14) *with an estimator* $\widehat{\mu}_n$ *of the first moments* μ^* *of the random coefficients that satisfies* $\sqrt{n}\,\big(\widehat{\mu}_n - \mu^*\big) = \mathcal{O}_{\mathbb{P}}\,(1)$. *In addition, suppose that Assumption 5.12 and the mutual incoherence condition*

$$\Big\|\mathrm{C}^{\sigma}_{S_{\sigma}^{c}S_{\sigma}}\,\big(\mathrm{C}^{\sigma}_{S_{\sigma}S_{\sigma}}\big)^{-1}\Big\|_{\mathrm{M},\infty} < 1 \tag{5.27}$$

are satisfied. Moreover, if $\lambda_n^{\sigma} \to 0$ *and* $\sqrt{n}\,\lambda_n^{\sigma} \to \infty$ *hold, the LASSO* $\widehat{\sigma}_n^{\,\mathrm{L}}$ *as a solution to* (5.18) *is sign-consistent,*

$$\mathbb{P}\Big(\mathrm{sign}\big(\widehat{\sigma}_n^{\,\mathrm{L}}\big) = \mathrm{sign}\big(\sigma^*\big)\Big) \to 1\,,$$

and has estimation rate λ_n^{σ} *on the support* S_{σ} *of* σ^*, *that is*

$$\frac{1}{\lambda_n^{\sigma}}\big(\widehat{\sigma}_{n,S_{\sigma}}^{\,\mathrm{L}} - \sigma^*_{S_{\sigma}}\big) = \mathcal{O}_{\mathbb{P}}\,(1)\,.$$

See also Remark 5.5 for a discussion about the disadvantages of the LASSO in comparison to the adaptive LASSO.

In the subsequent theorem we provide, similar to Theorem 5.6, a statement about the sign-consistency of the adaptive LASSO of the variances and covariances and its asymptotic normality on the true support S_σ. For this purpose we define

$$\mathrm{B}^\sigma := \mathbb{E}\Big[\Big(\mathrm{v}\big(\mathbf{X}\big)^\top \Psi^* \,\mathrm{v}\big(\mathbf{X}\big)\Big)\, \mathrm{v}\big(\mathbf{X}\big)\, \mathrm{v}\big(\mathbf{X}\big)^\top\Big] \quad \in \mathbb{R}^{\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}}\,, \tag{5.28}$$

where the matrix Ψ^* is given in (5.24). The above matrix is positive semi-definite because of Assumption (A6) and the inequality (5.26) in Remark 5.11. The proof of the following theorem is deferred to Section 5.4 as well.

Theorem 5.16 (Asymptotics adaptive LASSO of the variances and covariances)**.** *Consider the linear regression model* (5.14) *with an estimator $\widehat{\mu}_n$ of the first moments μ^* of the random coefficients that satisfies $\sqrt{n}\,\big(\widehat{\mu}_n - \mu^*\big) = \mathcal{O}_{\mathbb{P}}\,(1)$, and let Assumption 5.12 be satisfied. In addition, suppose that the initial estimator $\widehat{\sigma}_n^{\mathrm{init}}$ of the variances and covariances σ^* of the coefficients satisfies $\sqrt{n}\,\big(\widehat{\sigma}_n^{\mathrm{init}} - \sigma^*\big) = \mathcal{O}_{\mathbb{P}}\,(1)$, and that the regularization parameter λ_n^σ satisfies $\lambda_n^\sigma \to 0$, $\sqrt{n}\,\lambda_n^\sigma \to 0$ and $n\,\lambda_n^\sigma \to \infty$. Then the adaptive LASSO $\widehat{\sigma}_n^{\mathrm{AL}}$ as a solution to* (5.19) *is sign-consistent,*

$$\mathbb{P}\Big(\mathrm{sign}\big(\widehat{\sigma}_n^{\mathrm{AL}}\big) = \mathrm{sign}\big(\sigma^*\big)\Big) \to 1\,, \tag{5.29}$$

and satisfies the asymptotic normality

$$\sqrt{n}\,\big(\widehat{\sigma}_{n,S_\sigma}^{\mathrm{AL}} - \sigma_{S_\sigma}^*\big) \;\xrightarrow{d}\; \mathcal{N}_{s_\sigma}\Big(\mathbf{0}_{s_\sigma}, \big(\mathrm{C}_{S_\sigma S_\sigma}^\sigma\big)^{-1} \mathrm{B}_{S_\sigma S_\sigma}^\sigma \,\big(\mathrm{C}_{S_\sigma S_\sigma}^\sigma\big)^{-1}\Big) \tag{5.30}$$

on the support S_σ of σ^*.

Remark 5.17. Potential choices for the consistent estimator $\widehat{\mu}_n$ of the first moments μ^* of the random coefficients are the least squares estimator and the adaptive LASSO with an appropriate choice for its regularization parameter, cf. Theorem 5.6 and Remark 5.7. Moreover, for further comments on the choice of the initial estimator $\widehat{\sigma}_n^{\mathrm{init}}$ of the variances and covariances and the asymptotic covariance matrix in (5.30) see also Remarks 5.7 and 5.27.

In addition, the results in Section 5.4 show that the asymptotic normality of the adaptive Lasso $\widehat{\sigma}_n^{\mathrm{AL}}$ on the true support S_σ is independent of the fact whether we know the first moments μ^* in advance or whether we estimate them simultaneously with an appropriate estimator $\widehat{\mu}_n$, see Remark 5.27 for more details.

Remark 5.18. We take a closer look at the problem of this section, variable selection and estimation of the variances and covariances of the random coefficients. If we are only interested in the fact whether some coefficients are deterministic or uncorrelated, the adaptive LASSO $\widehat{\sigma}_n^{\mathrm{AL}}$ provides for a large number n of observations under the conditions of Theorem 5.16 the right solution with high probability. However, if we are interested in the whole covariance matrix Σ^*, a problem arises. Let

$$\begin{aligned} \mathrm{mat}\colon \mathbb{R}^{\frac{d(d+1)}{2}} &\to \mathbb{R}^{d\times d}\,, \\ \mathbf{x} &\mapsto \mathrm{vec}^{-1}(\mathbf{x}) \end{aligned}$$

be the inverse function of the half-vectorization vec, which is defined in (2.8). If we determine for a fixed number n the adaptive LASSO $\widehat{\sigma}_n^{\mathrm{AL}}$ of the half-vectorization of the covariance matrix, it is possible that the corresponding matrix $\widehat{\Sigma}_n^{\mathrm{AL}} = \mathrm{mat}\left(\widehat{\sigma}_n^{\mathrm{AL}}\right)$ is not positive semi-definite. Hence it would be preferable to optimize in the definition (5.19) of the adaptive LASSO only over the image

$$\mathbb{V}_p^+ = \left\{\mathrm{vec}(M) \,\middle|\, M \in \mathbb{S}_p^+\right\} \quad \subset \mathbb{R}^{\frac{p(p+1)}{2}}$$

of the half-vectorizations of the cone of the positive semi-definite matrices

$$\mathbb{S}_p^+ = \left\{M \in \mathbb{R}^{p\times p} \,\middle|\, M \text{ is symmetric and positive semi-definite}\right\} \quad \subset \mathbb{R}^{p\times p}\,.$$

This leads to the adaptive LASSO $\widehat{\sigma}_n^{\mathrm{AL,pos}}$ as a solution of the constrained optimization problem

$$\widehat{\sigma}_n^{\mathrm{AL,pos}} \in \rho_{\sigma,n,\lambda_n^\sigma}^{\mathrm{AL,pos}} := \operatorname*{arg\,min}_{\beta\in\mathbb{V}_p^+} \left(\mathcal{L}_{\sigma,n}^{\mathrm{LS}}(\beta) + 2\lambda_n^\sigma \sum_{k=1}^{\frac{p(p+1)}{2}} \frac{|\beta_k|}{\left|\widehat{\sigma}_{n,k}^{\mathrm{init}}\right|}\right),$$

with regularization parameter $\lambda_n^\sigma > 0$ and initial estimator $\widehat{\sigma}_n^{\mathrm{init}} \in \mathbb{R}^{\frac{p(p+1)}{2}}$. Evidently, we get for every number n of observations by $\widehat{\Sigma}_n^{\mathrm{AL,pos}} = \mathrm{mat}\left(\widehat{\sigma}_n^{\mathrm{AL,pos}}\right)$ a positive semi-definite matrix. However, technically it is hard to extend the primal-dual witness approach underlying the proof of Theorem 5.16 to this setting. Moreover, if the true covariance matrix Σ^* is positive definite, meaning it is in the interior of $\mathbb{S}_p^+$ respectively σ^* is in the interior of $\mathbb{V}_p^+$, then asymptotically the solutions of the constrained and unconstrained optimization problem are equal with probability tending to one, that is

$$\mathbb{P}\left(\widehat{\sigma}_n^{\mathrm{AL}} = \widehat{\sigma}_n^{\mathrm{AL,pos}}\right) \to 1\,,$$

because the optimality conditions are the same. Hence, if we assume for the true covariance matrix Σ^* the block form

$$\Sigma^* = \begin{bmatrix} \Sigma_1^* & \mathbf{0}_{d\times(p-d)} \\ \mathbf{0}_{(p-d)\times d} & \mathbf{0}_{(p-d)\times(p-d)} \end{bmatrix}$$

with $\Sigma_1^* \in \mathbb{R}^{d\times d}$, $d \leq p$, is positive definite, then under the conditions of Theorem 5.16 for a large number n of observations the image $\mathrm{mat}\left(\widehat{\sigma}_n^{\mathrm{AL}}\right)$ of the ordinary adaptive LASSO $\widehat{\sigma}_n^{\mathrm{AL}}$ is a covariance matrix of the above form with high probability. This is an implication of the sign-consistency and the consistency on the true support S_σ in Theorem 5.16. Note that the corresponding statement would not be true for the ordinary least squares estimator.

5.3. Simulations

We consider in our numerical study the linear random coefficient regression model (2.6) with normally distributed coefficients $\mathbf{A} = (B_0, B_1, \ldots, B_5)^\top \sim \mathcal{N}_6(\mu^*, \Sigma^*)$ with mean vector $\mu^* = (40, 15, 0, -10, 20, 0)^\top \in \mathbb{R}^6$ and covariance matrix

$$\Sigma^* = \begin{bmatrix} 10 & 15.65 & -5.20 & 0 & 0 & 0 \\ 15.65 & 50 & 0 & 12.65 & 0 & 0 \\ -5.20 & 0 & 30 & -12.25 & 0 & 0 \\ 0 & 12.65 & -12.25 & 20 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{6\times 6}\,.$$

The exact correlation coefficients are $\rho_{12} = 0.7$, $\rho_{13} = -0.3$, $\rho_{24} = 0.4$, $\rho_{34} = -0.5$ and evidently $\rho_{14} = \rho_{23} = 0$. Furthermore, the covariates $W_1, \ldots, W_5$ are assumed to be independent and identically distributed. We simulate n pairs $(Y_1, \mathbf{W_1}^\top)^\top, \ldots, (Y_n, \mathbf{W_n}^\top)^\top$ of data according to the specified model (2.6) and afterwards we estimate the means, variances and covariances of the random coefficients $\mathbf{A}$ with the adaptive LASSO based on these observations. The appropriate estimators $\widehat{\mu}_n^{\mathrm{AL}}$ and $\widehat{\sigma}_n^{\mathrm{AL}}$ with regularization parameters $\lambda_n^\mu, \lambda_n^\sigma > 0$ are given in (5.6) and (5.19). The initial estimator of the adaptive LASSO is always the least squares estimator, see (5.4) and (5.17), and in the second stage mean regression for the variances and covariances σ^* we use the adaptive LASSO $\widehat{\mu}_n^{\mathrm{AL}}$ as estimator for the first moments μ^*. Moreover, note that the mean and variance of the random intercept B_0 are not penalized in our simulation because we use the function `glmnet()`. This is plausible in terms of content since the random intercept includes the deterministic intercept as well as a random error which is not affected by the regressors. In each of the following scenarios we perform two Monte Carlo simulations with $m = 10.000$ iterations each to illustrate the sign-consistency and asymptotic normality of $\widehat{\mu}_n^{\mathrm{AL}}$ and $\widehat{\sigma}_n^{\mathrm{AL}}$, provided in the Theorems 5.6 and 5.16, for various sample sizes. In first simulation we consider always $n_1 = 10.000$ observations and in the second one $n_2 = 100.000$. The regularization parameters are chosen such that there is a satisfactory trade-off between a high sign-recovery rate and a small estimation error. For this purpose we use 1000 independent repetitions, run through a grid for the parameters λ in each data set and determine the parameters with a correct number of degrees of freedom and, in addition, among these the one with the smallest ℓ_2 norm of the respective estimator error. Based on this information we choose the regularization parameters for the adaptive LASSO estimators $\widehat{\mu}_n^{\mathrm{AL}}$ and $\widehat{\sigma}_n^{\mathrm{AL}}$ in the Monte Carlo simulations.

(a) **$\mathcal{U}[-1,1]$ distributed covariates.**
We consider $W_1, \dots, W_5 \sim \mathcal{U}[-1,1]$ independent and uniformly distributed on the interval $[-1,1]$, and the regularization parameters are chosen by $\lambda^{\mu}_{n_1} = 0.4, \lambda^{\mu}_{n_2} = 0.2, \lambda^{\sigma}_{n_1} = 16.5$ and $\lambda^{\sigma}_{n_2} = 12.0$. For the sign-consistency we obtain the empirical probabilities

$$\mathbb{P}\Big(\operatorname{sign}\big(\widehat{\mu}^{\mathrm{AL}}_{n_1}\big) = \operatorname{sign}\big(\mu^*\big)\Big) = 0.9427\,, \qquad \mathbb{P}\Big(\operatorname{sign}\big(\widehat{\mu}^{\mathrm{AL}}_{n_2}\big) = \operatorname{sign}\big(\mu^*\big)\Big) = 0.9641\,,$$
$$\mathbb{P}\Big(\operatorname{sign}\big(\widehat{\sigma}^{\mathrm{AL}}_{n_1}\big) = \operatorname{sign}\big(\sigma^*\big)\Big) = 0.7513\,, \qquad \mathbb{P}\Big(\operatorname{sign}\big(\widehat{\sigma}^{\mathrm{AL}}_{n_2}\big) = \operatorname{sign}\big(\sigma^*\big)\Big) = 0.9127\,.$$

A comparison of the empirical and asymptotic densities of the rescaled estimation errors for the means, variances and covariances of the random coefficients, which are unequal to zero, are shown in the following figures. In particular, the larger sample size n_2 increases the sign-recovery rate for the second central moments and the empirical densities lie closer to the asymptotic ones as well.

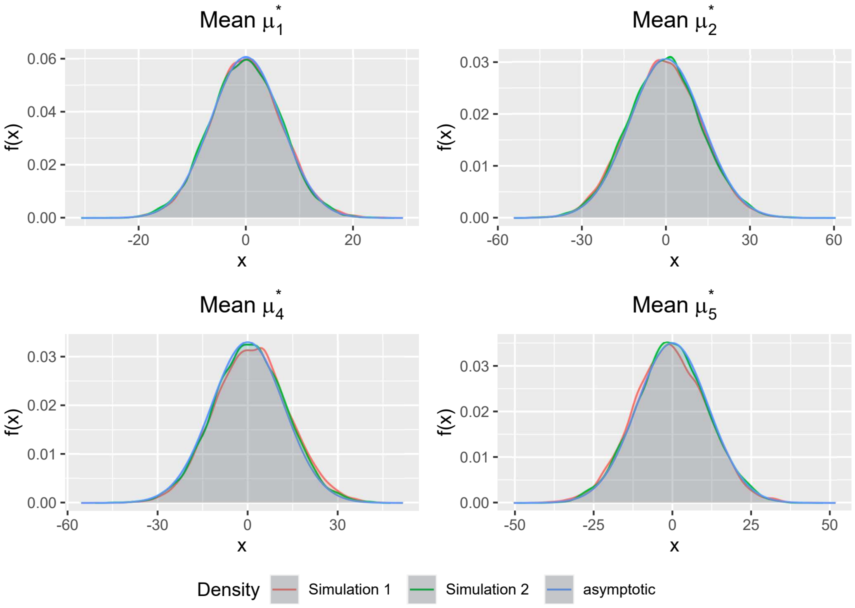

Figure 5.1.: comparison of the densities of the rescaled estimation error for the means with $\mathcal{U}[-1,1]$ distributed covariates.

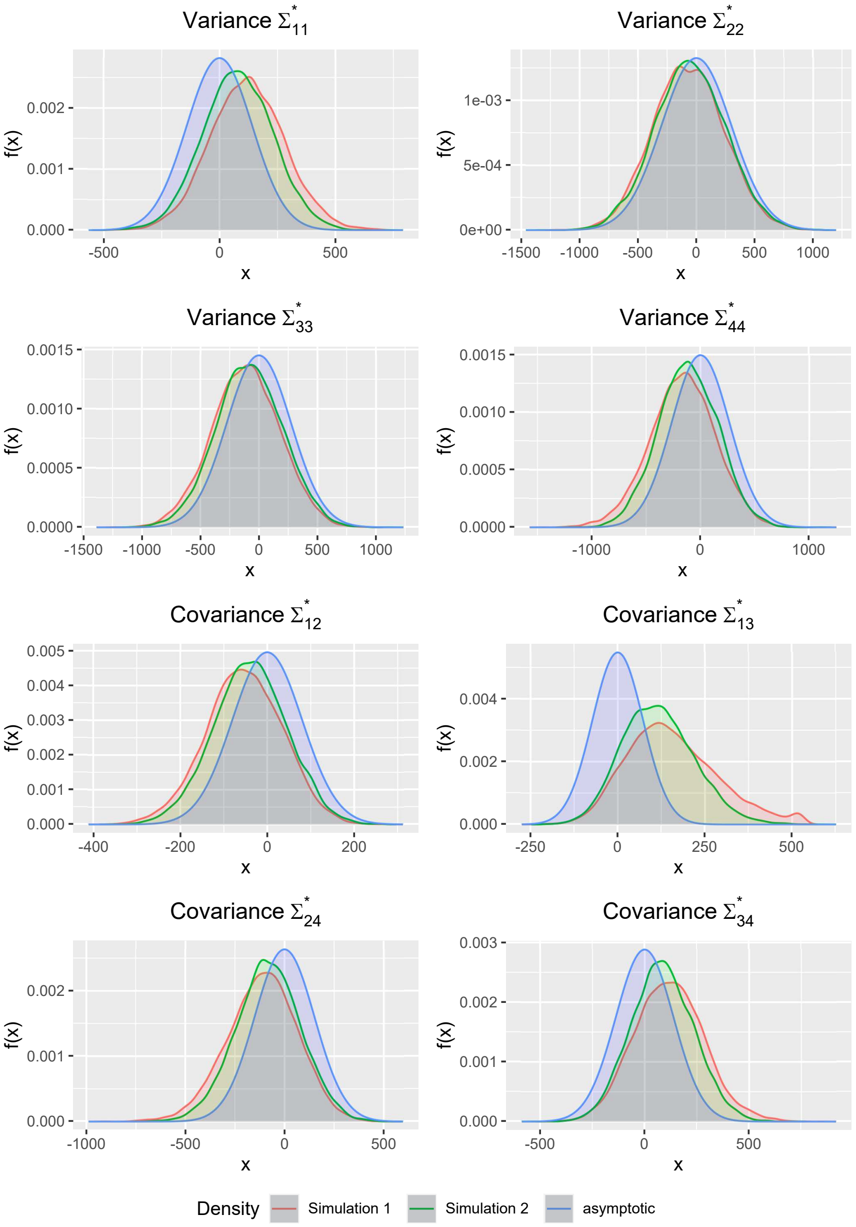

Figure 5.2.: comparison of the densities of the rescaled estimation error for the variances and covariances with $\mathcal{U}[-1, 1]$ distributed covariates.

(b) **$\mathcal{U}\{-1,0,1\}$ distributed covariates.**
We consider $W_1,\dots,W_5 \sim \mathcal{U}\{-1,0,1\}$ independent and uniformly distributed on the set $\{-1,0,1\}$ and the regularization parameters are chosen by $\lambda^{\mu}_{n_1} = 0.5, \lambda^{\mu}_{n_2} = 0.3, \lambda^{\sigma}_{n_1} = 23.8$ and $\lambda^{\sigma}_{n_2} = 17.4$. The following results are very similar to the ones in scenario (a), which confirms numerically that three distinct support points for each regressor are sufficient for the variable selection and estimation of the means, variances and covariances of the random coefficients.

$$\mathbb{P}\Big(\mathrm{sign}\big(\widehat{\mu}^{\mathrm{AL}}_{n_1}\big) = \mathrm{sign}\big(\mu^*\big)\Big) = 0.9398\,, \qquad \mathbb{P}\Big(\mathrm{sign}\big(\widehat{\mu}^{\mathrm{AL}}_{n_2}\big) = \mathrm{sign}\big(\mu^*\big)\Big) = 0.9688\,,$$
$$\mathbb{P}\Big(\mathrm{sign}\big(\widehat{\sigma}^{\mathrm{AL}}_{n_1}\big) = \mathrm{sign}\big(\sigma^*\big)\Big) = 0.7045\,, \qquad \mathbb{P}\Big(\mathrm{sign}\big(\widehat{\sigma}^{\mathrm{AL}}_{n_2}\big) = \mathrm{sign}\big(\sigma^*\big)\Big) = 0.898\,.$$

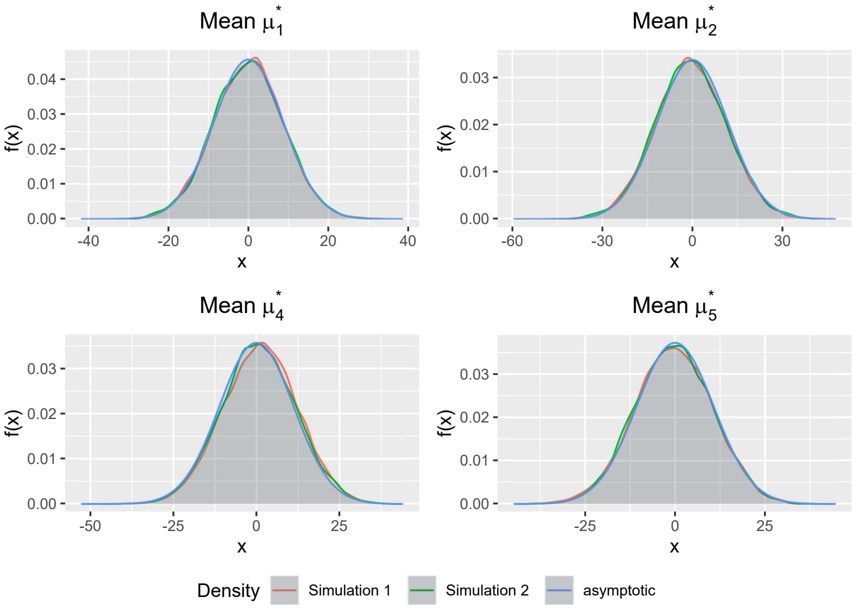

Figure 5.3.: comparison of the densities of the rescaled estimation error for the means with $\mathcal{U}\{-1,0,1\}$ distributed covariates.

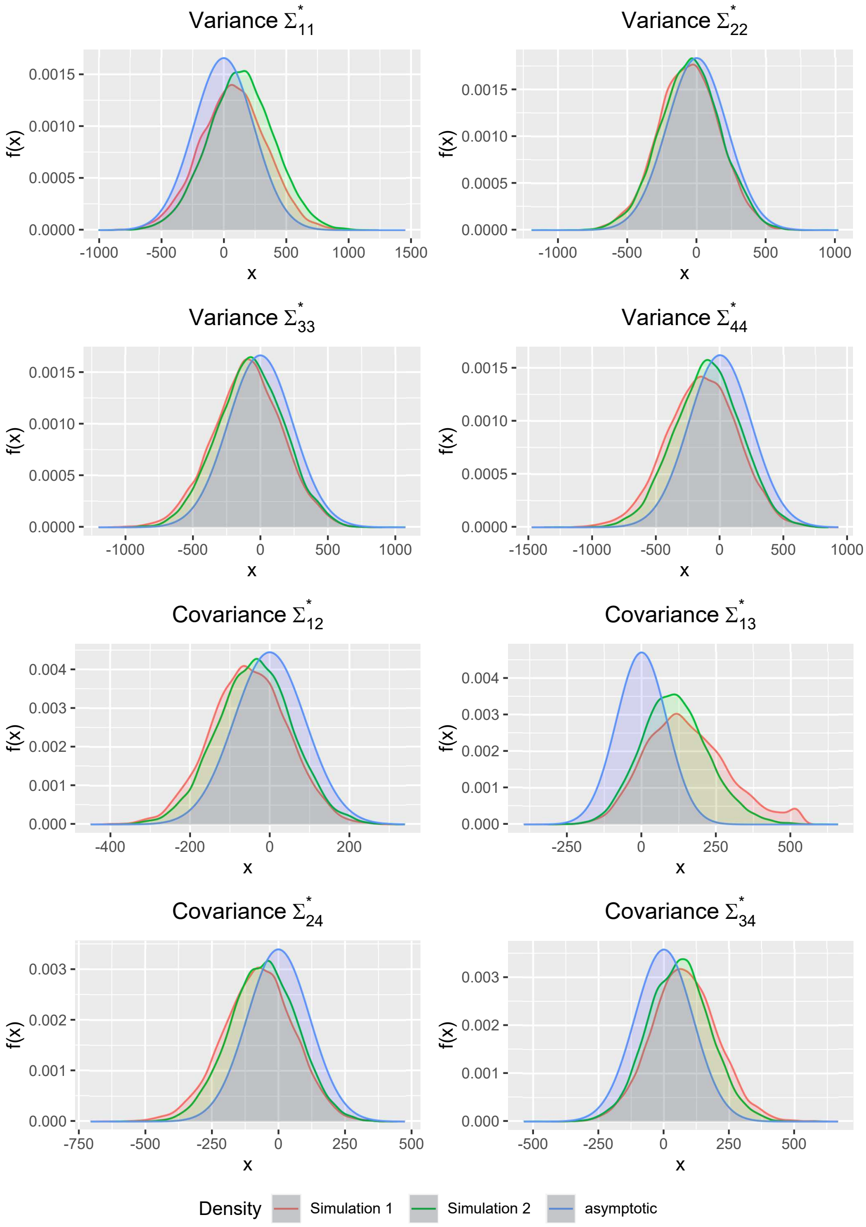

Figure 5.4.: comparison of the densities of the rescaled estimation error for the variances and covariances with $\mathcal{U}\{-1, 0, 1\}$ distributed covariates.

(c) **$\boldsymbol{\mathcal{U}}[\mathbf{0},\mathbf{1}]$ distributed covariates.**
We consider $W_1, \ldots, W_5 \sim \mathcal{U}[0,1]$ independent and uniformly distributed on the interval $[0,1]$, and the regularization parameters are chosen by $\lambda^{\mu}_{n_1} = 0.2, \lambda^{\mu}_{n_2} = 0.1, \lambda^{\sigma}_{n_1} = 2.1$ and $\lambda^{\sigma}_{n_2} = 1.0$. Here the parameter for the variance estimation is chosen such that we obtain the highest possible probability for sign-consistency. However, these results are not comparable to the ones in the scenarios (a) and (b). The plots in Figure 5.6 suggest that the variances Σ^*_{33} and Σ^*_{44} and all non-zero covariances are set to zero in some iterations of the first simulation. In the second one problems persist only for the variance Σ^*_{33} and the covariance Σ^*_{13}, while the empirical densities of the covariances $\Sigma^*_{12}, \Sigma^*_{24}$ and Σ^*_{34} look much better. Apparently, positive and negative values for the regressors are crucial to perform satisfying variable selection for the second central moments if the initial estimator is the least squares estimator. Maybe one achieves better results if for that purpose the empirical quadratic loss is combined with the Ridge or elastic net penalty. This stabilizes the regularization paths even for regressors that are highly correlated.

$$\mathbb{P}\Big(\operatorname{sign}\big(\widehat{\mu}^{\,\mathrm{AL}}_{n_1}\big) = \operatorname{sign}\big(\mu^*\big)\Big) = 0.861\,, \qquad \mathbb{P}\Big(\operatorname{sign}\big(\widehat{\mu}^{\,\mathrm{AL}}_{n_2}\big) = \operatorname{sign}\big(\mu^*\big)\Big) = 0.917\,,$$

$$\mathbb{P}\Big(\operatorname{sign}\big(\widehat{\sigma}^{\,\mathrm{AL}}_{n_1}\big) = \operatorname{sign}\big(\sigma^*\big)\Big) = 0.0021\,, \qquad \mathbb{P}\Big(\operatorname{sign}\big(\widehat{\sigma}^{\,\mathrm{AL}}_{n_2}\big) = \operatorname{sign}\big(\sigma^*\big)\Big) = 0.1024\,.$$

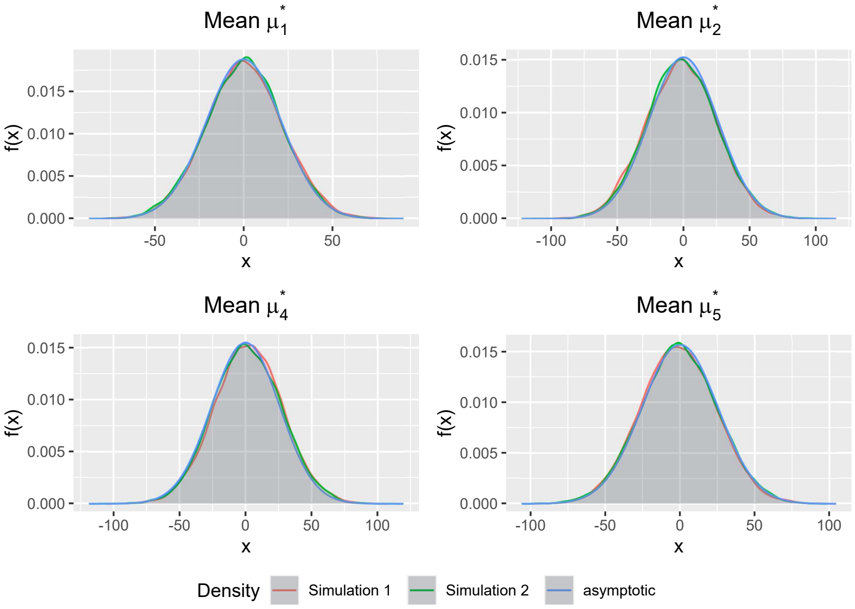

Figure 5.5.: comparison of the densities of the rescaled estimation error for the means with $\mathcal{U}[0,1]$ distributed covariates.

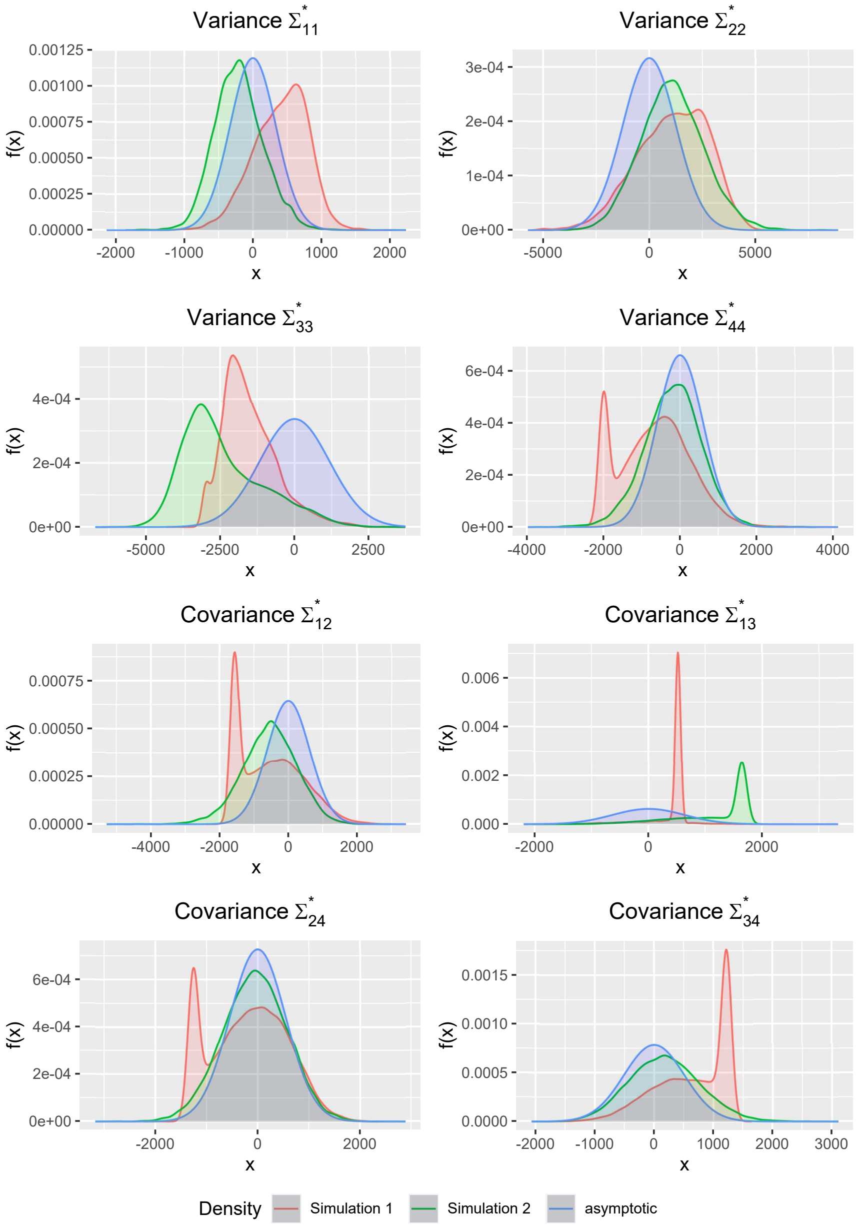

Figure 5.6.: comparison of the densities of the rescaled estimation error for the variances and covariances with $\mathcal{U}[0,1]$ distributed covariates.

5.4. Main steps of the proofs

The proofs of the results for the LASSO in the Theorems 5.4 and 5.15 as well as the ones for the adaptive LASSO in the Theorems 5.6 and 5.16 are very similar. Thus, we mainly prove the statements for the estimation of the variances and covariances of the random coefficients in Section 5.2.3 in detail, and merely outline the ones for the mean regression in Section 5.1.2.

At first we consider the equivalence of the convergence of a sequence of matrices and the convergence of their entries. The proof is deferred to Section 5.5.3.

Lemma 5.19.

1. *A sequence of matrices $(Q_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{d_1\times d_2}$ converges to $Q \in \mathbb{R}^{d_1\times d_2}$ with respect to the ℓ_∞ operator norm if and only if the entries $\big((Q_n)_{kl}\big)_{n\in\mathbb{N}}$ converge to Q_{kl} for all $k \in \{1,\dots,d_1\}$ and $l \in \{1,\dots,d_2\}$.*
2. *A sequence of random matrices $(Q_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{d_1\times d_2}$ converges almost surely to $Q \in \mathbb{R}^{d_1\times d_2}$ with respect to the ℓ_∞ operator norm if and only if the entries $\big((Q_n)_{kl}\big)_{n\in\mathbb{N}}$ converge almost surely to Q_{kl} for all $k \in \{1,\dots,d_1\}$ and $l \in \{1,\dots,d_2\}$.*

In the following proofs the rescaled Gram matrices $\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma$ and the matrices $\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \Omega_n^\sigma\, \mathbb{X}_n^\sigma$, where the conditional covariance matrix Ω_n^σ is defined in (5.23), play an important role. Thus, we discuss in the subsequent remark their expected value and convergence.

Remark 5.20. The rescaled gram matrices can also be written as

$$\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma = \frac{1}{n}\sum_{i=1}^n \mathrm{v}\big(\mathbf{X_i}\big)\,\mathrm{v}\big(\mathbf{X_i}\big)^\top .$$

Under Assumption (A5) the entries of the matrices are integrable and we obtain the expected value

$$\mathbb{E}\left[\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\mathrm{v}\big(\mathbf{X_i}\big)\,\mathrm{v}\big(\mathbf{X_i}\big)^\top\Big] = \mathbb{E}\Big[\mathrm{v}\big(\mathbf{X}\big)\,\mathrm{v}\big(\mathbf{X}\big)^\top\Big] = \mathrm{C}^\sigma$$

since the regressors $\mathbf{X_1},\dots,\mathbf{X_n}$ are independent and identically distributed. Furthermore, the strong law of large numbers implies

$$e_q^\top\left(\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma\right)e_r = \frac{1}{n}\sum_{i=1}^n e_q^\top \mathrm{v}\big(\mathbf{X_i}\big)\,\mathrm{v}\big(\mathbf{X_i}\big)^\top e_r \xrightarrow{a.s.} \mathbb{E}\big[e_q^\top \mathrm{v}\big(\mathbf{X}\big)\,\mathrm{v}\big(\mathbf{X}\big)^\top e_r\big] = e_q^\top\,\mathrm{C}^\sigma\, e_r$$

for $q, r \in \{1,\dots,p(p+1)/2\}$, and together with Lemma 5.19 the almost sure convergence

$$\left\|\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma - \mathrm{C}^\sigma\right\|_{\mathrm{M},\infty} \xrightarrow{a.s.} 0 \tag{5.31}$$

follows. If in addition Assumption (A4) is satisfied, the matrix Ψ^* with the fourth moments of the random coefficients exists, cf. Lemma 5.10 and Remark 5.11, and, moreover, we obtain

$$\frac{1}{n}\,\big(\mathbb{X}_n^\sigma\big)^\top \Omega_n^\sigma\,\mathbb{X}_n^\sigma = \frac{1}{n}\sum_{i=1}^{n}\Big(\mathrm{v}\big(\mathbf{X_i}\big)^\top \Psi^*\,\mathrm{v}\big(\mathbf{X_i}\big)\Big)\,\mathrm{v}\big(\mathbf{X_i}\big)\,\mathrm{v}\big(\mathbf{X_i}\big)^\top$$

since Ω_n^σ is a diagonal matrix with entries $\mathrm{v}(\mathbf{X_1})^\top\Psi^*\,\mathrm{v}(\mathbf{X_1}),\dots,\mathrm{v}(\mathbf{X_n})^\top\Psi^*\,\mathrm{v}(\mathbf{X_n})$. This leads to

$$\mathbb{E}\bigg[\frac{1}{n}\,\big(\mathbb{X}_n^\sigma\big)^\top \Omega_n^\sigma\,\mathbb{X}_n^\sigma\bigg] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\bigg[\Big(\mathrm{v}\big(\mathbf{X_i}\big)^\top \Psi^*\,\mathrm{v}\big(\mathbf{X_i}\big)\Big)\,\mathrm{v}\big(\mathbf{X_i}\big)\,\mathrm{v}\big(\mathbf{X_i}\big)^\top\bigg] = \mathrm{B}^\sigma \tag{5.32}$$

under the Assumptions (A4) and (A5), where B^σ is given in (5.28). Furthermore, we obtain also

$$\bigg\|\frac{1}{n}\,\big(\mathbb{X}_n^\sigma\big)^\top \Omega_n^\sigma\,\mathbb{X}_n^\sigma - \mathrm{B}^\sigma\bigg\|_{\mathrm{M},\infty} \xrightarrow{a.s.} 0 \tag{5.33}$$

by the strong law of large numbers and Lemma 5.19. Note that the almost sure convergences in (5.31) and (5.33) hold with respect to ℓ_2 operator norm as well since $\|M\|_{\mathrm{M},2} \le \|M\|_{\mathrm{M},\infty}$ for a symmetric matrix $M \in \mathbb{R}^{d\times d}$. Additionally, the above averages can be used as consistent estimators for the matrices B^σ and C^σ, and hence also for the asymptotic covariance matrix in (5.30) in Theorem 5.16. In doing so one can in principle construct a consistent estimator for the matrix Ψ^*, which can be obtained from linear regression equations for higher-order moments of the random coefficients similarly to those for the variances and covariances.

Remark 5.21. Similar to Remark 5.20 we obtain under the Assumptions (A1) and (A2) the expected values

$$\mathbb{E}\bigg[\frac{1}{n}\,\big(\mathbb{X}_n^\mu\big)^\top \mathbb{X}_n^\mu\bigg] = \mathrm{C}^\mu\,, \qquad \mathbb{E}\bigg[\frac{1}{n}\,\big(\mathbb{X}_n^\mu\big)^\top \Omega_n^\mu\,\mathbb{X}_n^\mu\bigg] = \mathrm{B}^\mu\,,$$

and the almost sure convergences

$$\bigg\|\frac{1}{n}\,\big(\mathbb{X}_n^\mu\big)^\top \mathbb{X}_n^\mu - \mathrm{C}^\mu\bigg\|_{\mathrm{M},\infty} \xrightarrow{a.s.} 0\,, \qquad \bigg\|\frac{1}{n}\,\big(\mathbb{X}_n^\mu\big)^\top \Omega_n^\mu\,\mathbb{X}_n^\mu - \mathrm{B}^\mu\bigg\|_{\mathrm{M},\infty} \xrightarrow{a.s.} 0\,,$$

where the second moment matrix C^μ of the covariates is given in (A3), the conditional covariance matrix Ω_n^μ of the first stage mean regression errors in (5.3) and B^μ in (5.8).

A further crucial object in the following proofs is the gradient

$$\nabla\mathcal{L}_{\sigma,n}^{\mathrm{LS}}\big(\sigma^*\big) = -\frac{2}{n}\,\big(\mathbb{X}_n^\sigma\big)^\top\big(\mathbb{Y}_n^\sigma - \mathbb{X}_n^\sigma\,\sigma^*\big) = -\frac{2}{n}\,\big(\mathbb{X}_n^\sigma\big)^\top\varepsilon_n^\sigma$$

of the empirical quadratic loss function $\mathcal{L}^{\mathrm{LS}}_{\sigma,n}$ with argument σ^*. We will see that stochastic boundedness of the sequence

$$Z^{\sigma}_n := -\frac{\sqrt{n}}{2}\,\nabla\mathcal{L}^{\mathrm{LS}}_{\sigma,n}\big(\sigma^*\big) = \frac{1}{\sqrt{n}}\,\big(\mathbb{X}^{\sigma}_n\big)^{\top}\varepsilon^{\sigma}_n \qquad \in \mathbb{R}^{\frac{p(p+1)}{2}}$$

of random vectors is required for the sign-consistency of the (adaptive) LASSO. Furthermore, the asymptotic covariance matrix in the normality (5.30) in Theorem 5.16 is essentially determined by the asymptotic covariance matrix of this sequence.

Remark 5.22. Remember the decomposition of the error vector ε^{σ}_n in Remark 5.8, then we can write

$$Z^{\sigma}_n = \frac{1}{\sqrt{n}}\,\big(\mathbb{X}^{\sigma}_n\big)^{\top}\big(\delta_n + \zeta_n + \xi_n\big) = Z^{\sigma,1}_n + Z^{\sigma,2}_n$$

with

$$Z^{\sigma,1}_n := \frac{1}{\sqrt{n}}\,\big(\mathbb{X}^{\sigma}_n\big)^{\top}\delta_n\,, \qquad Z^{\sigma,2}_n := \frac{1}{\sqrt{n}}\,\big(\mathbb{X}^{\sigma}_n\big)^{\top}\big(\zeta_n + \xi_n\big)\,. \tag{5.34}$$

The random vectors ζ_n and ξ_n, which are given in (5.21) and (5.22), are affected by the first stage mean regression, whereas the random vector δ_n, given in (5.20), is independent of this estimation error and exists also if the means of the random coefficients are known in advance. We can show in the following lemma that the random vectors $Z^{\sigma,2}_n$ converge in probability to zero if we assume that the estimator $\widehat{\mu}_n$ of the means μ^* has rate $1/\sqrt{n}$. Hence, in this case, the stochastic boundedness and asymptotic covariance matrix of Z^{σ}_n is only determined by the random vectors $Z^{\sigma,1}_n$.

Lemma 5.23. *Suppose that the Assumptions* (A4) *and* (A5) *hold. Then the random vectors $Z^{\sigma,2}_n$ in* (5.34) *converge in probability to zero,*

$$Z^{\sigma,2}_n = \mathrm{o}_{\mathbb{P}}\,(1)\,,$$

if $\sqrt{n}\,\big(\widehat{\mu}_n - \mu^*\big) = \mathcal{O}_{\mathbb{P}}\,(1)$ is satisfied.

The technical proof is deferred to Section 5.5.4. Now we make a statement about the remaining part of the rescaled gradient Z^{σ}_n.

Lemma 5.24. *Suppose that Assumption* (A4) *holds. Then the random vectors $Z^{\sigma,1}_n$ in* (5.34) *satisfy*

$$\mathbb{E}\big[Z^{\sigma,1}_n \,\big|\, \mathbb{X}^{\sigma}_n\big] = \mathbf{0}_{\frac{p(p+1)}{2}} \qquad \text{and} \qquad \mathbb{C}\mathrm{ov}\big(Z^{\sigma,1}_n \,\big|\, \mathbb{X}^{\sigma}_n\big) = \frac{1}{n}\,\big(\mathbb{X}^{\sigma}_n\big)^{\top}\Omega^{\sigma}_n\,\mathbb{X}^{\sigma}_n\,.$$

If in addition Assumption (A5) *holds, then*

$$\mathbb{C}\mathrm{ov}\big(Z^{\sigma,1}_n\big) = \mathrm{B}^{\sigma}\,,$$

which implies

$$Z^{\sigma,1}_n = \mathcal{O}_{\mathbb{P}}\,(1)\,.$$

Proof of Lemma 5.24. Consider the definition of the random variables $Z_n^{\sigma,1}$ in (5.34). Under Assumption (A4) we obtain

$$\mathbb{E}\big[Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big] = \frac{1}{\sqrt{n}} \left(\mathbb{X}_n^\sigma\right)^\top \mathbb{E}\big[\delta_n \,\big|\, \mathbb{X}_n^\sigma\big] = \mathbf{0}_{\frac{p(p+1)}{2}}$$

and

$$\mathbb{C}\mathrm{ov}\big(Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big) = \frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \mathbb{C}\mathrm{ov}\big(\delta_n \,\big|\, \mathbb{X}_n^\sigma\big)\, \mathbb{X}_n^\sigma = \frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \Omega_n^\sigma\, \mathbb{X}_n^\sigma$$

by Lemma 5.10. For real-valued, square-integrable random variables Q_1, Q_2 and Q_3 on the same probability space the law of total covariance implies the decomposition

$$\mathbb{C}\mathrm{ov}\big(Q_1, Q_2\big) = \mathbb{E}\Big[\mathbb{C}\mathrm{ov}\big(Q_1, Q_2 \,\big|\, Q_3\big)\Big] + \mathbb{C}\mathrm{ov}\Big(\mathbb{E}\big[Q_1 \,\big|\, Q_3\big], \mathbb{E}\big[Q_2 \,\big|\, Q_3\big]\Big)\,.$$

This can be extended to random vectors and covariance matrices, and hence we obtain under the Assumptions (A4) and (A5) the matrix

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\big(Z_n^{\sigma,1}\big) &= \mathbb{E}\Big[\mathbb{C}\mathrm{ov}\big(Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big)\Big] + \mathbb{C}\mathrm{ov}\Big(\mathbb{E}\big[Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big]\Big) \\
&= \mathbb{E}\left[\frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \Omega_n^\sigma\, \mathbb{X}_n^\sigma\right] + \mathbb{C}\mathrm{ov}\Big(\mathbf{0}_{\frac{p(p+1)}{2}}\Big) \\
&= \mathbb{E}\left[\frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \Omega_n^\sigma\, \mathbb{X}_n^\sigma\right].
\end{aligned}$$

Moreover, the equation (5.32) in Remark 5.20 leads to

$$\mathbb{C}\mathrm{ov}\big(Z_n^{\sigma,1}\big) = \mathbb{E}\left[\frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \Omega_n^\sigma\, \mathbb{X}_n^\sigma\right] = \mathrm{B}^\sigma\,.$$

□

In the subsequent lemma we formulate the analogous result for the rescaled gradient of the empirical quadratic loss function $\mathcal{L}_{\mu,n}^{\mathrm{LS}}$ of the first moments of the random coefficients.

Lemma 5.25. *Suppose that Assumption* (A1) *holds. Then the random vectors*

$$Z_n^\mu := -\frac{\sqrt{n}}{2}\, \nabla \mathcal{L}_{\mu,n}^{\mathrm{LS}}\big(\mu^*\big) = \frac{1}{\sqrt{n}} \left(\mathbb{X}_n^\mu\right)^\top \varepsilon_n^\mu \qquad \in \mathbb{R}^p$$

satisfy

$$\mathbb{E}\big[Z_n^\mu \,\big|\, \mathbb{X}_n^\mu\big] = \mathbf{0}_p \qquad \textit{and} \qquad \mathbb{C}\mathrm{ov}\big(Z_n^\mu \,\big|\, \mathbb{X}_n^\mu\big) = \frac{1}{n} \left(\mathbb{X}_n^\mu\right)^\top \Omega_n^\mu\, \mathbb{X}_n^\mu\,.$$

If in addition Assumption (A2) *holds, then*

$$\mathbb{C}\mathrm{ov}\big(Z_n^\mu\big) = \mathrm{B}^\mu\,,$$

which implies

$$Z_n^\mu = \mathcal{O}_{\mathbb{P}}\left(1\right).$$

Proof of Lemma 5.25. Proceed similarly to the proof of Lemma 5.24. In doing so we use the results in Remark 5.21. □

Now we can prove the Theorems 5.4 and 5.15.

Proof of Theorem 5.15. We use the primal-dual witness characterization of the LASSO in Lemma 2.3 in Section 2.2 to prove the desired results.
By Assumption (A6) and the almost sure convergence in (5.31) in Remark 5.20 there exists a $N \in \mathbb{N}$ so that the matrices $\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma$ are invertible almost surely for all $n \geq N$. Thus, henceforth we assume that the number n of observations is at least N. By Loh and Wainwright (2017, Lemma 11) the inequality

$$\left\| \left(\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma\right)^{-1} - \left(\mathrm{C}^\sigma\right)^{-1} \right\|_{\mathrm{M},\infty} \leq \frac{\left\|\left(\mathrm{C}^\sigma\right)^{-1}\right\|_{\mathrm{M},\infty}^2 \left\|\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma - \mathrm{C}^\sigma\right\|_{\mathrm{M},\infty}}{1 - \left\|\left(\mathrm{C}^\sigma\right)^{-1}\right\|_{\mathrm{M},\infty} \left\|\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma - \mathrm{C}^\sigma\right\|_{\mathrm{M},\infty}}$$

follows, and hence by Assumption (A6) and the aforementioned convergence in (5.31) we obtain

$$\left\| \left(\frac{1}{n}\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma\right)^{-1} - \left(\mathrm{C}^\sigma\right)^{-1} \right\|_{\mathrm{M},\infty} \xrightarrow{a.s.} 0\,. \tag{5.35}$$

Furthermore, basic properties of the ℓ_∞ operator norm lead to

$$\begin{aligned}
&\left\| \left(\mathbb{X}_{n,S_\sigma^c}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma \left(\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma\right)^{-1} - \mathrm{C}_{S_\sigma^c S_\sigma}^\sigma \left(\mathrm{C}_{S_\sigma S_\sigma}^\sigma\right)^{-1} \right\|_{\mathrm{M},\infty} \\
&\quad = \left\| \frac{1}{n}\left(\mathbb{X}_{n,S_\sigma^c}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma \left(\frac{1}{n}\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma\right)^{-1} - \mathrm{C}_{S_\sigma^c S_\sigma}^\sigma \left(\mathrm{C}_{S_\sigma S_\sigma}^\sigma\right)^{-1} \right\|_{\mathrm{M},\infty} \\
&\quad \leq \left(\left\|\mathrm{C}_{S_\sigma^c S_\sigma}^\sigma\right\|_{\mathrm{M},\infty} + \left\| \frac{1}{n}\left(\mathbb{X}_{n,S_\sigma^c}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma - \mathrm{C}_{S_\sigma^c S_\sigma}^\sigma \right\|_{\mathrm{M},\infty} \right) \\
&\qquad\qquad \cdot \left\| \left(\frac{1}{n}\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma\right)^{-1} - \left(\mathrm{C}_{S_\sigma S_\sigma}^\sigma\right)^{-1} \right\|_{\mathrm{M},\infty} \\
&\qquad + \left\| \frac{1}{n}\left(\mathbb{X}_{n,S_\sigma^c}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma - \mathrm{C}_{S_\sigma^c S_\sigma}^\sigma \right\|_{\mathrm{M},\infty} \left\| \left(\mathrm{C}_{S_\sigma S_\sigma}^\sigma\right)^{-1} \right\|_{\mathrm{M},\infty},
\end{aligned}$$

and hence by the almost sure convergences in (5.31) and (5.35) it follows that

$$\left\| \left(\mathbb{X}_{n,S_\sigma^c}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma \left(\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma\right)^{-1} - \mathrm{C}_{S_\sigma^c S_\sigma}^\sigma \left(\mathrm{C}_{S_\sigma S_\sigma}^\sigma\right)^{-1} \right\|_{\mathrm{M},\infty} \xrightarrow{a.s.} 0\,. \tag{5.36}$$

Consequently, we obtain

$$\begin{aligned}
&\lim_{n\to\infty} \left\| \big(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \operatorname{sign}\big(\sigma^*_{S_\sigma}\big) \right\|_{\infty} \\
&\quad \leq \lim_{n\to\infty} \left\| \big(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \right\|_{\mathrm{M},\infty} \big\| \operatorname{sign}\big(\sigma^*_{S_\sigma}\big) \big\|_{\infty} \\
&\quad \leq \lim_{n\to\infty} \left\| \big(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \right\|_{\mathrm{M},\infty} = \left\| \mathrm{C}^{\sigma}_{S^c_\sigma S_\sigma} \big(\mathrm{C}^{\sigma}_{S_\sigma S_\sigma}\big)^{-1} \right\|_{\mathrm{M},\infty} < 1
\end{aligned}$$

almost surely by the mutual incoherence condition in (5.27). Hence there exists a $\eta > 0$ so that for large sample sizes n it holds that

$$\mathbb{P}\left(\left\| \big(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \operatorname{sign}\big(\sigma^*_{S_\sigma}\big) \right\|_{\infty} \leq 1 - \eta \right) = 1\,.$$

Moreover, by Lemmas 5.23 and 5.24 together with Remark 5.22 and the convergence in (5.36) it follows that

$$\begin{aligned}
&\frac{1}{n\,\lambda^{\sigma}_n} \left(\big(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\big)^{\top} \varepsilon^{\sigma}_n - \big(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \varepsilon^{\sigma}_n \right) \\
&\quad = \frac{1}{\sqrt{n}\,\lambda^{\sigma}_n} \left(\frac{1}{\sqrt{n}} \big(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\big)^{\top} \varepsilon^{\sigma}_n - \big(\mathbb{X}^{\sigma}_{n,S^c_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \left(\big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \frac{1}{\sqrt{n}} \big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \varepsilon^{\sigma}_n \right) \\
&\quad = \frac{1}{\sqrt{n}\,\lambda^{\sigma}_n} \mathcal{O}_{\mathbb{P}}(1) = \mathrm{o}_{\mathbb{P}}(1) \qquad (5.37)
\end{aligned}$$

since $\sqrt{n}\,\lambda^{\sigma}_n \to \infty$ holds by assumption. With the help of the triangle inequality the first condition (2.3) of Lemma 2.3 follows with high probability for a sufficient large number n of observations. Furthermore, let

$$\widetilde{\sigma}_{n,S_\sigma} = \sigma^*_{S_\sigma} + \left(\frac{1}{n} \big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \left(\frac{1}{n} \big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \varepsilon^{\sigma}_n - \lambda^{\sigma}_n \operatorname{sign}\big(\sigma^*_{S_\sigma}\big) \right).$$

Then we obtain

$$\begin{aligned}
\frac{1}{\lambda^{\sigma}_n} \big(\widetilde{\sigma}_{n,S_\sigma} - \sigma^*_{S_\sigma}\big) &= \frac{1}{\lambda^{\sigma}_n} \left(\frac{1}{n} \big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \left(\frac{1}{n} \big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \varepsilon^{\sigma}_n - \lambda^{\sigma}_n \operatorname{sign}\big(\sigma^*_{S_\sigma}\big) \right) \\
&= \left(\frac{1}{n} \big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \mathbb{X}^{\sigma}_{n,S_\sigma} \right)^{-1} \left(\frac{1}{\sqrt{n}\,\lambda^{\sigma}_n} \left(\frac{1}{\sqrt{n}} \big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \varepsilon^{\sigma}_n \right) - \operatorname{sign}\big(\sigma^*_{S_\sigma}\big) \right).
\end{aligned}$$

The vector in brackets satisfies

$$\frac{1}{\sqrt{n}\,\lambda^{\sigma}_n} \left(\frac{1}{\sqrt{n}} \big(\mathbb{X}^{\sigma}_{n,S_\sigma}\big)^{\top} \varepsilon^{\sigma}_n \right) - \operatorname{sign}\big(\sigma^*_{S_\sigma}\big) = \frac{1}{\sqrt{n}\,\lambda^{\sigma}_n} \mathcal{O}_{\mathbb{P}}(1) + \mathcal{O}(1) = \mathrm{o}_{\mathbb{P}}(1) + \mathcal{O}(1) = \mathcal{O}_{\mathbb{P}}(1)$$

because of the results in the Lemmas 5.23 and 5.24 together with the assumption $\sqrt{n}\,\lambda_n^\sigma \to \infty$. Hence the convergence in (5.35) implies

$$\frac{1}{\lambda_n^\sigma}\left(\widetilde{\sigma}_{n,S_\sigma} - \sigma^*_{S_\sigma}\right) = \mathcal{O}_{\mathbb{P}}\left(1\right), \tag{5.38}$$

and the condition $\lambda_n^\sigma \to 0$ leads to

$$\widetilde{\sigma}_{n,S_\sigma} - \sigma^*_{S_\sigma} = \mathrm{o}_{\mathbb{P}}\left(1\right).$$

As a consequence, the second condition, $\operatorname{sign}\left(\widetilde{\sigma}_{n,S_\sigma}\right) = \operatorname{sign}\left(\sigma^*_{S_\sigma}\right)$, of Lemma 2.3 is satisfied with high probability for large sample sizes n. The mentioned lemma and equation (5.38) imply the assertions of Theorem 5.15. □

Proof of Theorem 5.4. Proceed similarly to the proof of Theorem 5.15. In doing so we use the results in Remark 5.21 and Lemma 5.25. □

The asymptotic normalities in the Theorems 5.6 and 5.16 are based on the following central limit theorem of random vectors.

Proposition 5.26 (Lindeberg-Feller central limit theorem)**.**
For each $n \in \mathbb{N}$ let $Q_{n,1}, \dots, Q_{n,k_n} \in \mathbb{R}^d$, $k_n \in \mathbb{N}$, be independent random vectors with finite variances such that

$$\lim_{n\to\infty} \left\| \sum_{i=1}^{k_n} \mathbb{C}\mathrm{ov}\left(Q_{n,i}\right) - \Pi \right\|_{\mathrm{M},\infty} = 0$$

with $\Pi \in \mathbb{R}^{d\times d}$, and for every $\delta > 0$

$$\lim_{n\to\infty} \sum_{i=1}^{k_n} \mathbb{E}\left[\left\|Q_{n,i}\right\|_2^2 \, \mathbb{1}\left\{\left\|Q_{n,i}\right\|_2 > \delta\right\} \right] = 0\,. \tag{5.39}$$

Then the sequence $\sum_{i=1}^{k_n}\left(Q_{n,i} - \mathbb{E}[Q_{n,i}]\right)$ converges in distribution to a normal $\mathcal{N}_d(\mathbf{0}_d, \Pi)$ distribution.

Proof. Cf. van der Vaart (1998, Proposition 2.27). □

Finally, we prove the results of the adaptive LASSO in the Theorems 5.6 and 5.16.

Proof of Theorem 5.16. We use the primal-dual witness characterization of the adaptive LASSO in Lemma 2.4 in Section 2.2 to prove the sign-consistency in (5.29), and the Lindeberg-Feller central limit theorem 5.26 to prove the asymptotic normality in (5.30). The assumptions $\sqrt{n}\,\lambda_n^\sigma \to 0$ and $\sqrt{n}\,(\widehat{\sigma}_n^{\,\mathrm{init}} - \sigma^*) = \mathcal{O}_{\mathbb{P}}\left(1\right)$ in Theorem 5.16 imply

$$0 \le \frac{\sqrt{n}\,\lambda_n^\sigma}{\left|\widehat{\sigma}_{n,k}^{\,\mathrm{init}}\right|} \le \frac{\sqrt{n}\,\lambda_n^\sigma}{\left|\left|\sigma_k^*\right| - \left|\widehat{\sigma}_{n,k}^{\,\mathrm{init}} - \sigma_k^*\right|\right|} \xrightarrow{\mathbb{P}} 0 \tag{5.40}$$

for all $k \in S_\sigma$ since $|\sigma_k^*| > 0$ for these k. Hence it follows by (5.36) and (5.37) in the proof of Theorem 5.15 that

$$\begin{aligned}\sqrt{n}\,\Bigg[&\big(\mathbb{X}^\sigma_{n,S_\sigma^c}\big)^\top \mathbb{X}^\sigma_{n,S_\sigma}\Big(\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \mathbb{X}^\sigma_{n,S_\sigma}\Big)^{-1}\lambda_n^\sigma\left(\frac{1}{|\widehat{\sigma}^{\,\text{init}}_{n,S_\sigma}|}\odot \operatorname{sign}\big(\sigma^*_{S_\sigma}\big)\right)\\ &\qquad\qquad\qquad\qquad + \frac{1}{n}\big(\mathbb{X}^\sigma_{n,S_\sigma^c}\big)^\top \mathrm{P}_{\mathbb{X}^\sigma_{\mathrm{n,S_\sigma}}}\,\varepsilon_n^\sigma\Bigg]\\ &= \mathcal{O}_{\mathbb{P}}(1)\,\mathrm{o}_{\mathbb{P}}(1) + \sqrt{n}\,\frac{1}{\sqrt{n}}\,\mathcal{O}_{\mathbb{P}}(1)\\ &= \mathcal{O}_{\mathbb{P}}(1)\,, \end{aligned} \tag{5.41}$$

where

$$\mathrm{P}_{\mathbb{X}^\sigma_{\mathrm{n,S_\sigma}}} = \mathrm{I}_n - \mathbb{X}^\sigma_{n,S_\sigma}\Big(\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \mathbb{X}^\sigma_{n,S_\sigma}\Big)^{-1}\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top .$$

Note that $(\mathbb{X}^\sigma_{n,S_\sigma})^\top \mathbb{X}^\sigma_{n,S_\sigma}$ is invertible for large sample sizes n with probability equal to one. Moreover, $\sqrt{n}\,(\widehat{\sigma}^{\,\text{init}}_n - \sigma^*) = \mathcal{O}_{\mathbb{P}}(1)$ leads also to $\sqrt{n}\,\widehat{\sigma}^{\,\text{init}}_{n,k} = \mathcal{O}_{\mathbb{P}}(1)$ for all $k \in S_\sigma^c$ since $\sigma_k^* = 0$ for these k. Thus, by the third requirement $n\,\lambda_n^\sigma \to \infty$ on the regularization parameter it follows that

$$\frac{\sqrt{n}\,\lambda_n^\sigma}{\big|\widehat{\sigma}^{\,\text{init}}_{n,k}\big|} = \frac{n\,\lambda_n^\sigma}{\sqrt{n}\,\big|\widehat{\sigma}^{\,\text{init}}_{n,k}\big|} \xrightarrow{\mathbb{P}} \to \infty$$

for all $k \in S_\sigma^c$. Together with (5.41) this implies the first condition (2.4) of Lemma 2.4 with high probability for a sufficient large number n of observations. Furthermore, let

$$\widetilde{\sigma}_{n,S_\sigma} = \sigma^*_{S_\sigma} + \left(\frac{1}{n}\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \mathbb{X}^\sigma_{n,S_\sigma}\right)^{-1}\left(\frac{1}{n}\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \varepsilon_n^\sigma - \lambda_n^\sigma\left(\frac{1}{|\widehat{\sigma}^{\,\text{init}}_{n,S_\sigma}|}\odot \operatorname{sign}\big(\sigma^*_{S_\sigma}\big)\right)\right).$$

Then we obtain

$$\sqrt{n}\,\big(\widetilde{\sigma}_{n,S_\sigma} - \sigma^*_{S_\sigma}\big) = \left(\frac{1}{n}\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \mathbb{X}^\sigma_{n,S_\sigma}\right)^{-1}\frac{1}{\sqrt{n}}\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \varepsilon_n^\sigma + \mathrm{o}_{\mathbb{P}}(1) \tag{5.42}$$

by the convergences in (5.35) and (5.40). Moreover, by Remark 5.22 and the Lemmas 5.23 and 5.24 it follows that

$$\begin{aligned}\sqrt{n}\,\big(\widetilde{\sigma}_{n,S_\sigma} - \sigma^*_{S_\sigma}\big) &= \left(\frac{1}{n}\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \mathbb{X}^\sigma_{n,S_\sigma}\right)^{-1}\frac{1}{\sqrt{n}}\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \delta_n + \mathrm{o}_{\mathbb{P}}(1) \qquad (5.43)\\ &= \mathcal{O}_{\mathbb{P}}(1) + \mathrm{o}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1)\,,\end{aligned}$$

where δ_n is given in (5.20), which in turn leads to

$$\widetilde{\sigma}_{n,S_\sigma} - \sigma^*_{S_\sigma} = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) = \mathrm{o}_{\mathbb{P}}(1)\,.$$

Hence the second condition, $\operatorname{sign}(\widetilde{\sigma}_{n,S_\sigma}) = \operatorname{sign}(\sigma^*_{S_\sigma})$, of Lemma 2.4 is satisfied with high probability for large sample sizes n as well. Sign-consistency of the adaptive LASSO and $\widehat{\sigma}^{\mathrm{AL}}_{n,S_\sigma} = \widetilde{\sigma}_{n,S_\sigma}$ is the consequence.

Note that for the asymptotic normality of the rescaled estimation error in (5.30) only the first term in (5.43) is crucial. Hence we consider the random vectors

$$Z_n^{\sigma,1} = \frac{1}{\sqrt{n}} \left(\mathbb{X}_n^\sigma\right)^\top \delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(e_i^\top \delta_n\right) \mathrm{v}\big(\mathbf{X_i}\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big(\mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^*\big)\Big) \,\mathrm{v}\big(\mathbf{X_i}\big)\,,$$

where $D_i = \big(\mathbf{A_i} - \mu^*\big)\big(\mathbf{A_i} - \mu^*\big)^\top$. Now we want to apply the Lindeberg-Feller central limit theorem 5.26 for the array

$$Q_{n,i} = \frac{1}{\sqrt{n}} \Big(\mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^*\big)\Big)\,\mathrm{v}\big(\mathbf{X_i}\big) \qquad \in \mathbb{R}^{\frac{p(p+1)}{2}}$$

with $i \in \{1, \ldots, n\}$. The random vectors are independent and identically distributed in each row (for fixed n) since $(\mathbf{X_1}^\top, \mathbf{A_1}^\top)^\top, \ldots, (\mathbf{X_n}^\top, \mathbf{A_n}^\top)^\top$ enjoy this property. Furthermore, they are centered,

$$\mathbb{E}\big[Q_{n,i}\big] = \frac{1}{\sqrt{n}}\,\mathbb{E}\Big[\mathbb{E}\Big[\mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^*\big)\,\Big|\,\mathbb{X}_n^\sigma\Big]\,\mathrm{v}\big(\mathbf{X_i}\big)\Big] = \frac{1}{\sqrt{n}}\,\mathbb{E}\Big[0 \cdot \mathrm{v}\big(\mathbf{X_i}\big)\Big] = \mathbf{0}_{\frac{p(p+1)}{2}}\,,$$

and for the sum of the covariance matrices

$$\sum_{i=1}^n \mathbb{C}\mathrm{ov}\big(Q_{n,i}\big) = \mathbb{C}\mathrm{ov}\bigg(\sum_{i=1}^n Q_{n,i}\bigg) = \mathbb{C}\mathrm{ov}\big(Z_n^{\sigma,1}\big)$$

we get by Lemma 5.24 the equation

$$\sum_{i=1}^n \mathbb{C}\mathrm{ov}\big(Q_{n,i}\big) = \mathrm{B}^\sigma$$

under the Assumptions (A4) and (A5). Moreover, we obtain for arbitrary $\delta > 0$ the equation

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\Big[\|Q_{n,i}\|_2^2 \,\mathbb{1}\big\{\|Q_{n,i}\|_2 > \delta\big\}\Big] \\
&\quad = \mathbb{E}\Big[\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(D_1 - \Sigma^*\big)\,\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(D_1 - \Sigma^*\big)\,\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{v}\big(\mathbf{X_1}\big) \\
&\qquad\qquad \cdot \mathbb{1}\big\{\mathrm{v}(\mathbf{X_1})^\top \mathrm{vec}(D_1 - \Sigma^*)\,\mathrm{v}(\mathbf{X_1})^\top \mathrm{vec}(D_1 - \Sigma^*)\,\mathrm{v}(\mathbf{X_1})^\top \mathrm{v}(\mathbf{X_1}) > \delta^2 n\big\}\Big]\,.
\end{aligned}$$

The expected value $\mathbb{E}\big[\mathrm{v}(\mathbf{X_1})^\top \mathrm{vec}(D_1 - \Sigma^*)\,\mathrm{v}(\mathbf{X_1})^\top \mathrm{vec}(D_1 - \Sigma^*)\,\mathrm{v}(\mathbf{X_1})^\top \mathrm{v}(\mathbf{X_1})\big]$ exists because of Assumption 5.12 and the Cauchy Schwarz inequality. Thus, we get

$$\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}\Big[\|Q_{n,i}\|_2^2 \,\mathbb{1}\big\{\|Q_{n,i}\|_2 > \delta\big\}\Big] = 0$$

by Lebesgue's dominated convergence theorem, which coincides with Lindeberg's condition (5.39) in Proposition 5.26. Hence the proposition implies the weak convergence

$$Z_n^{\sigma,1} = \frac{1}{\sqrt{n}} \left(\mathbb{X}_n^\sigma\right)^\top \delta_n = \sum_{i=1}^n Q_{n,i} \ \xrightarrow{d}\ Q \sim \mathcal{N}_{\frac{p(p+1)}{2}} \left(\mathbf{0}_{\frac{p(p+1)}{2}}, \mathrm{B}^\sigma\right),$$

respectively

$$\frac{1}{\sqrt{n}} \left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \delta_n \ \xrightarrow{d}\ Q_{S_\sigma} \sim \mathcal{N}_{s_\sigma} \left(\mathbf{0}_{s_\sigma}, \mathrm{B}_{S_\sigma S_\sigma}^\sigma\right).$$

So, all in all a multivariate version of Slutsky's theorem, cf. van der Vaart (1998, Theorem 2.7, Lemma 2.8), together with equation (5.43) and the almost sure convergence in (5.35) in the proof of Theorem 5.15 leads to

$$\sqrt{n} \left(\widehat{\sigma}_{n,S_\sigma}^{\mathrm{AL}} - \sigma_{S_\sigma}^*\right) \ \xrightarrow{d}\ \left(\mathrm{C}_{\mathrm{S}_\sigma \mathrm{S}_\sigma}^\sigma\right)^{-1} Q_{S_\sigma}.$$

In addition, it follows that

$$\left(\mathrm{C}_{\mathrm{S}_\sigma \mathrm{S}_\sigma}^\sigma\right)^{-1} Q_{S_\sigma} \sim \mathcal{N}_{s_\sigma} \left(\mathbf{0}_{s_\sigma}, \left(\mathrm{C}_{\mathrm{S}_\sigma \mathrm{S}_\sigma}^\sigma\right)^{-1} \mathrm{B}_{S_\sigma S_\sigma}^\sigma \left(\mathrm{C}_{\mathrm{S}_\sigma \mathrm{S}_\sigma}^\sigma\right)^{-1}\right)$$

by the symmetry of $\mathrm{C}_{\mathrm{S}_\sigma \mathrm{S}_\sigma}^\sigma$ and the properties of the multivariate normal distribution, and hence the asserted asymptotic normality in (5.30). □

Proof of Theorem 5.6. Proceed similarly to the proof of Theorem 5.16. In doing so we use the results in Remark 5.21 and Lemma 5.25 □

Remark 5.27. Keep the considerations in the Remarks 5.8 and 5.9 in mind. Then it is obvious that we would obtain equation (5.43) in the proof of Theorem 5.16 as well, even if we know the first moments μ^* of the random coefficients in advance and $\varepsilon_n^\sigma = \delta_n$ holds. Hence the asymptotic normality of the estimator $\widehat{\sigma}_n^{\mathrm{AL}}$ of the variances and covariances σ^* of the coefficients on their support S_σ is independent of the fact whether we know the first moments μ^* in advance or whether we estimate them simultaneously with an appropriate estimator $\widehat{\mu}_n$.

Moreover, by Assumption 5.12 the Gram matrices $\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma$ are invertible with probability equal to one for large sample sizes n because of the almost sure convergence in (5.35) in the proof of Theorem 5.15. In this case the least squares estimator of the variances and covariances σ^* in (5.17) is unique and given by

$$\begin{aligned} \widehat{\sigma}_n^{\mathrm{LS}} &= \left(\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma\right)^{-1} \left(\mathbb{X}_n^\sigma\right)^\top \mathbb{Y}_n^\sigma = \left(\left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma\right)^{-1} \left(\mathbb{X}_n^\sigma\right)^\top \left(\mathbb{X}_n^\sigma \sigma^* + \varepsilon_n^\sigma\right) \\ &= \sigma^* + \left(\frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma\right)^{-1} \frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \varepsilon_n^\sigma. \end{aligned}$$

Hence under Assumption 5.12 and the assumption that the estimator $\widehat{\mu}_n$ of the first moments μ^* of the random coefficients satisfies $\sqrt{n}\left(\widehat{\mu}_n - \mu^*\right) = \mathcal{O}_{\mathbb{P}}(1)$, it follows that

$$\sqrt{n} \left(\widehat{\sigma}_n^{\mathrm{LS}} - \sigma^*\right) \ \xrightarrow{d}\ \mathcal{N}_{\frac{p(p+1)}{2}} \left(\mathbf{0}_{\frac{p(p+1)}{2}}, \left(\mathrm{C}^\sigma\right)^{-1} \mathrm{B}^\sigma \left(\mathrm{C}^\sigma\right)^{-1}\right)$$

by the considerations, especially equation (5.42), in the proof of Theorem 5.16.

Remark 5.28. Similar to Remark 5.27 for a large number n of observations the least squares estimator of the first moments μ^* of the random coefficients in (5.4) is unique and given by

$$\widehat{\mu}_n^{\mathrm{LS}} = \mu^* + \left(\frac{1}{n}\left(\mathbb{X}_n^\mu\right)^\top \mathbb{X}_n^\mu\right)^{-1} \frac{1}{n}\left(\mathbb{X}_n^\mu\right)^\top \varepsilon_n^\mu$$

with probability equal to one under Assumption 5.2. Moreover, the (more detailed) proof of Theorem 5.6 implies

$$\sqrt{n}\left(\widehat{\mu}_n^{\mathrm{LS}} - \mu^*\right) \xrightarrow{d} \mathcal{N}_p\Big(\mathbf{0}_p, \left(\mathrm{C}^\mu\right)^{-1} \mathrm{B}^\mu \left(\mathrm{C}^\mu\right)^{-1}\Big)$$

under Assumption 5.2

5.5. Technical proofs

5.5.1. Proofs of Lemmas 5.3 and 5.14

Proof of Lemma 5.3. Let $\mathbf{W_1}, \ldots, \mathbf{W_p} \sim \mathbf{W}$ be independent, then the matrix

$$\mathbb{X}_p^\mu = \left[\mathbf{X_1}, \ldots, \mathbf{X_p}\right]^\top = \begin{bmatrix} 1 & \mathbf{W_1^\top} \\ \vdots & \vdots \\ 1 & \mathbf{W_p^\top} \end{bmatrix}$$

is of full rank on a set A which has positive probability $\mathbb{P}(A) > 0$. If the points $\mathbf{w_1}, \ldots, \mathbf{w_p}$ have positive probability each, this is clear by the proof of Proposition 4.5. Otherwise, if $\mathbf{w_j}$ itself has probability equal to zero, every neighborhood of $\mathbf{w_j}$ must have positive probability. Moreover, for all points $\mathbf{z} \in \mathbb{R}^{p-1}$ in a very small open neighborhood of $\mathbf{w_j}$ we obtain that

$$\begin{bmatrix} 1 & \mathbf{w_1^\top} \\ \vdots & \vdots \\ 1 & \mathbf{w_{j-1}^\top} \\ 1 & \mathbf{z}^\top \\ 1 & \mathbf{w_{j+1}^\top} \\ \vdots & \vdots \\ 1 & \mathbf{w_p^\top} \end{bmatrix}$$

has full rank since the set of the full rank matrices is open and the coordinate projections are continuous. This can be done for each support point, and hence for each row of the above matrix. We conclude that

$$\left(\mathbb{X}_p^\mu\right)^\top \mathbb{X}_p^\mu = \sum_{i=1}^p \mathbf{X_i}\,\mathbf{X_i^\top}$$

is positive definite on A and otherwise positive semi-definite, which leads to

$$\begin{aligned} v^\top \mathrm{C}^\mu\, v &= \frac{1}{p}\, v^\top \mathbb{E}\Big[\sum_{i=1}^{p} \mathbf{X_i}\, \mathbf{X_i}^\top\Big]\, v = \frac{1}{p}\, \mathbb{E}\Big[v^\top \big(\mathbb{X}_p^\mu\big)^\top \mathbb{X}_p^\mu\, v\Big] \\ &= \frac{1}{p}\bigg(\int_A v^\top \big(\mathbb{X}_p^\mu\big)^\top \mathbb{X}_p^\mu\, v\, d\mathbb{P} + \int_{A^c} v^\top \big(\mathbb{X}_p^\mu\big)^\top \mathbb{X}_p^\mu\, v\, d\mathbb{P}\bigg) \\ &> 0 \end{aligned}$$

for all $v \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$, since the first integral is positive and the second non-negative. □

Proof of Lemma 5.14. Firstly, Theorem 4.11 implies that there exist $p(p+1)/2$ support points $\mathbf{w_1}, \dots, \mathbf{w_{\frac{p(p+1)}{2}}} \in \mathbb{R}^{p-1}$ of $\mathbf{W}$ such that the matrix S in (4.6) has full rank $p(p+1)/2$. Then we can argue similarly to the proof of Lemma 5.3 to show the positive definiteness of C^σ. □

5.5.2. Proof of Lemma 5.10

Proof of Lemma 5.10. Since $\mathbb{E}[D_i] = \Sigma^*$ holds, we get for the conditional mean vector of δ_n the equation

$$\begin{aligned} \mathbb{E}\big[\delta_n \,\big|\, \mathbb{X}_n^\sigma\big] &= \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathbb{E}\Big[\mathrm{vec}\big(D_1 - \Sigma^*\big)\Big], \dots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \mathbb{E}\Big[\mathrm{vec}\big(D_n - \Sigma^*\big)\Big]\Big)^\top \\ &= \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\Big(\mathbb{E}\big[D_1\big] - \Sigma^*\Big), \dots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \mathrm{vec}\Big(\mathbb{E}\big[D_n\big] - \Sigma^*\Big)\Big)^\top \\ &= \mathbf{0}_n\,. \end{aligned}$$

The entries

$$e_i^\top \delta_n = \mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^*\big) = \mathbf{X_i}^\top \big(D_i - \Sigma^*\big)\mathbf{X_i} = \mathbf{X_i}^\top \Big(\big(\mathbf{A_i} - \mu^*\big)\big(\mathbf{A_i} - \mu^*\big)^\top - \Sigma^*\Big)\, \mathbf{X_i}$$

of δ_n are pairwise independent because the random vectors $(\mathbf{X_1}^\top, \mathbf{A_1}^\top)^\top, \dots, (\mathbf{X_n}^\top, \mathbf{A_n}^\top)^\top$ are independent. Hence the conditional covariance matrix of δ_n is a diagonal matrix and it remains to determine the conditional variances of the errors. We get for $i \in \{1, \dots, n\}$ on the one hand

$$\begin{aligned} \mathbb{V}\mathrm{ar}\big(e_i^\top \delta_n \,\big|\, \mathbf{X_i}\big) &= \mathbb{V}\mathrm{ar}\Big(\mathbf{X_i}^\top \big(D_i - \Sigma^*\big)\mathbf{X_i} \,\Big|\, \mathbf{X_i}\Big) \\ &= \sum_{k,l,u,v=1}^{p} X_{i,k} X_{i,l} X_{i,u} X_{i,v}\, \mathbb{C}\mathrm{ov}\Big(\big(D_1\big)_{kl}, \big(D_1\big)_{uv}\Big) \\ &= \sum_{k,l,u,v=1}^{p} X_{i,k} X_{i,l} X_{i,u} X_{i,v}\, \big(\mathcal{M}^{kl}\big)_{uv} \end{aligned}$$

because of the definition of $\mathcal{M}^{kl}$ in (5.25). On the other hand it is

$$\begin{aligned}\mathbb{V}\mathrm{ar}\Big(e_i^\top \delta_n \,\Big|\, \mathrm{v}\big(\mathbf{X_i}\big)\Big) &= \mathbb{V}\mathrm{ar}\Big(\mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^*\big) \,\Big|\, \mathrm{v}\big(\mathbf{X_i}\big)\Big) \\ &= \sum_{q,r=1}^{\frac{p(p+1)}{2}} \mathrm{v}\big(\mathbf{X_i}\big)_q \mathrm{v}\big(\mathbf{X_i}\big)_r \,\mathbb{C}\mathrm{ov}\Big(\mathrm{vec}\big(D_1\big)_q, \mathrm{vec}\big(D_1\big)_r\Big) .\end{aligned}$$

Definition (2.8) of the half-vectorization vec and the last two equations imply

$$\mathbb{C}\mathrm{ov}\Big(\mathrm{vec}\big(D_1\big)_q, \mathrm{vec}\big(D_1\big)_r\Big) = \Psi^*_{qr}$$

for $q, r = \{1, \dots, p(p+1)/2\}$. Hence

$$\mathbb{V}\mathrm{ar}\Big(e_i^\top \delta_n \,\Big|\, \mathrm{v}\big(\mathbf{X_i}\big)\Big) = \mathrm{v}\big(\mathbf{X_i}\big)^\top \Psi^* \,\mathrm{v}\big(\mathbf{X_i}\big) ,$$

and in total the conditional covariance matrix of δ_n regarding $\mathrm{v}\big(\mathbf{X_1}\big), \dots, \mathrm{v}\big(\mathbf{X_n}\big)$ is given by

$$\mathbb{C}\mathrm{ov}\big(\delta_n \,\big|\, \mathbb{X}_n^\sigma\big) = \mathrm{diag}\Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \Psi^* \,\mathrm{v}\big(\mathbf{X_1}\big), \dots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \Psi^* \,\mathrm{v}\big(\mathbf{X_n}\big)\Big) .$$

□

5.5.3. Proof of Lemma 5.19

Proof of Lemma 5.19.

1. It is

$$\begin{aligned}\lim_{n\to\infty} \Big|\big(Q_n - Q\big)_{kl}\Big| &\le \lim_{n\to\infty} \max_{k\in\{1,\dots,d_1\}} \sum_{l=1}^{d_2} \Big|\big(Q_n - Q\big)_{kl}\Big| = \lim_{n\to\infty} \|Q_n - Q\|_{\mathrm{M},\infty} \\ &\le \lim_{n\to\infty} \sum_{k=1}^{d_1}\sum_{l=1}^{d_2} \Big|\big(Q_n - Q\big)_{kl}\Big| \\ &= \sum_{k=1}^{d_1}\sum_{l=1}^{d_2} \lim_{n\to\infty} \Big|\big(Q_n - Q\big)_{kl}\Big|\end{aligned}$$

   for $k \in \{1, \dots, d_1\}$ and $l \in \{1, \dots, d_2\}$. Hence the convergence of the matrices with respect to the ℓ_∞ operator norm is equivalent to the component-wise convergence of the matrices.

2. Suppose that $(Q_n)_{n\in\mathbb{N}}$ converges almost surely to Q with respect to the ℓ_∞ operator norm. This implies

$$\mathbb{P}\Big(\lim_{n\to\infty} \Big|\big(Q_n - Q\big)_{kl}\Big| \ne 0\Big) \le \mathbb{P}\Big(\lim_{n\to\infty} \|Q_n - Q\|_{\mathrm{M},\infty} \ne 0\Big) = 0$$

for $k \in \{1, \ldots, d_1\}$ and $l \in \{1, \ldots, d_2\}$. Conversely, we obtain

$$\begin{aligned}
\mathbb{P}\Big(\lim_{n\to\infty} \|Q_n - Q\|_{\mathrm{M},\infty} \neq 0 \Big) &\leq \mathbb{P}\Big(\sum_{k=1}^{d_1} \sum_{l=1}^{d_2} \lim_{n\to\infty} \Big| \big(Q_n - Q\big)_{kl} \Big| \neq 0 \Big) \\
&\leq \mathbb{P}\Big(\bigcup_{k=1}^{d_1} \bigcup_{l=1}^{d_2} \Big\{ \lim_{n\to\infty} \Big| \big(Q_n - Q\big)_{kl} \Big| \neq 0 \Big\} \Big) \\
&\leq \sum_{k=1}^{d_1} \sum_{l=1}^{d_2} \mathbb{P}\Big(\lim_{n\to\infty} \Big| \big(Q_n - Q\big)_{kl} \Big| \neq 0 \Big) \\
&= 0
\end{aligned}$$

if $(Q_n)_{n\in\mathbb{N}}$ converges component-wise almost surely to Q.

□

5.5.4. Proof of Lemma 5.23

For the proof of Lemma 5.23 we use in particular the consistency of the estimator $\widehat{\mu}_n$ of the first moments μ^* of the random coefficients with an estimation rate of $1/\sqrt{n}$. Lemma 5.23 is an immediate consequence of the following two lemmas.

Lemma 5.29. *Suppose that the Assumptions* (A4) *and* (A5) *hold. Then the random vectors*

$$Z_n^{\sigma,3} := \frac{1}{\sqrt{n}} \left(\mathbb{X}_n^{\sigma}\right)^{\top} \zeta_n \,,$$

where ζ_n is defined in (5.21)*, converge in probability to zero,*

$$Z_n^{\sigma,3} = \mathrm{o}_{\mathbb{P}}\,(1)\,,$$

if $\sqrt{n}\left(\widehat{\mu}_n - \mu^*\right) = \mathcal{O}_{\mathbb{P}}\,(1)$ is satisfied.

Proof of Lemma 5.29. The random vectors $Z_n^{\sigma,3}$ can be written with the definition of ζ_n in (5.21) as

$$\begin{aligned}
Z_n^{\sigma,3} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(e_i^{\top} \zeta_n\big) \mathrm{v}\big(\mathbf{X_i}\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\mathrm{v}\big(\mathbf{X_i}\big)^{\top} \mathrm{vec}\big(E_n\big) \Big) \mathrm{v}\big(\mathbf{X_i}\big) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg(\sum_{q=1}^{\frac{p(p+1)}{2}} \mathrm{v}\big(\mathbf{X_i}\big)_q \mathrm{vec}\big(E_n\big)_q \Bigg) \mathrm{v}\big(\mathbf{X_i}\big) \\
&= \sum_{q=1}^{\frac{p(p+1)}{2}} \sqrt{n}\, \mathrm{vec}\big(E_n\big)_q \Bigg(\frac{1}{n} \sum_{i=1}^{n} \mathrm{v}\big(\mathbf{X_i}\big)_q \mathrm{v}\big(\mathbf{X_i}\big) \Bigg) \,, \qquad (5.44)
\end{aligned}$$

where

$$E_n = \big(\mu^* - \widehat{\mu}_n\big)\big(\mu^* - \widehat{\mu}_n\big)^\top .$$

By the assumption $\sqrt{n}\left(\widehat{\mu}_n - \mu^*\right) = \mathcal{O}_{\mathbb{P}}(1)$ it follows that the entries of the matrix E_n satisfy

$$e_k^\top E_n\, e_l = \big(\widehat{\mu}_{n,k} - \mu_k^*\big)\big(\widehat{\mu}_{n,l} - \mu_l^*\big) = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n}\right)$$

for $k, l \in \{1, \ldots, p\}$, and hence also

$$\sqrt{n}\,\mathrm{vec}\big(E_n\big)_q = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) = \mathrm{o}_{\mathbb{P}}(1) \tag{5.45}$$

for all $q \in \{1, \ldots, p(p+1)/2\}$. Furthermore, the random vectors

$$Q_i^q = \mathrm{v}\big(\mathbf{X_i}\big)_q\,\mathrm{v}\big(\mathbf{X_i}\big)$$

are independent and identically distributed for $i \in \{1, \ldots, n\}$ since $\mathbf{X_1}, \ldots, \mathbf{X_n}$ enjoy this property. Additionally, the inequality $\|x\|_2 \leq \|x\|_1$ for $x \in \mathbb{R}^d$ implies

$$\mathbb{E}\Big[\|Q_i^q\|_2\Big] \leq \mathbb{E}\Big[\|Q_i^q\|_1\Big] = \mathbb{E}\Big[\Big\|\mathrm{v}\big(\mathbf{X_i}\big)_q\,\mathrm{v}\big(\mathbf{X_i}\big)\Big\|_1\Big] = \sum_{r=1}^{\frac{p(p+1)}{2}} \mathbb{E}\Big[\Big|\mathrm{v}\big(\mathbf{X_i}\big)_q\,\mathrm{v}\big(\mathbf{X_i}\big)_r\Big|\Big] < \infty$$

since the fourth moments of the regressors exist by Assumption (A5). Hence we get by the strong law of large numbers and Lemma 5.19 the convergence

$$\left\|\frac{1}{n}\sum_{i=1}^{n} \mathrm{v}\big(\mathbf{X_i}\big)_q\,\mathrm{v}\big(\mathbf{X_i}\big) - \mathbb{E}\big[Q_1^q\big]\right\|_\infty = \left\|\frac{1}{n}\sum_{i=1}^{n} Q_i^q - \mathbb{E}\big[Q_1^q\big]\right\|_\infty \xrightarrow{a.s.} 0\,,$$

which implies

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{v}\big(\mathbf{X_i}\big)_q\,\mathrm{v}\big(\mathbf{X_i}\big) = \mathcal{O}_{\mathbb{P}}(1) \tag{5.46}$$

for all $q \in \{1, \ldots, p(p+1)/2\}$. So all in all (5.44) - (5.46) lead to the assertion

$$Z_n^{\sigma,3} = \sum_{q=1}^{\frac{p(p+1)}{2}} \mathrm{o}_{\mathbb{P}}(1)\,\mathcal{O}_{\mathbb{P}}(1) = \mathrm{o}_{\mathbb{P}}(1)\,.$$

□

Lemma 5.30. *Suppose that the Assumptions* (A4) *and* (A5) *hold. Then the random vectors*

$$Z_n^{\sigma,4} := \frac{1}{\sqrt{n}}\left(\mathbb{X}_n^\sigma\right)^\top \xi_n\,,$$

where ξ_n *is defined in* (5.22)*, converge in probability to zero,*

$$Z_n^{\sigma,4} = \mathrm{o}_{\mathbb{P}}\left(1\right),$$

if $\sqrt{n}\left(\widehat{\mu}_n - \mu^*\right) = \mathcal{O}_{\mathbb{P}}\left(1\right)$ *is satisfied.*

Proof of Lemma 5.30. The random vectors $Z_n^{\sigma,4}$ can be written with the definition of ξ_n in (5.22) as

$$\begin{aligned}
Z_n^{\sigma,4} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(e_i^\top \xi_n\right) \mathrm{v}\big(\mathbf{X_i}\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(F_{n,i}\big)\Big) \mathrm{v}\big(\mathbf{X_i}\big) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg(\sum_{q=1}^{\frac{p(p+1)}{2}} \mathrm{v}\big(\mathbf{X_i}\big)_q \mathrm{vec}\big(F_{n,i}\big)_q\Bigg) \mathrm{v}\big(\mathbf{X_i}\big) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg(\sum_{k,l=1}^{p} X_{i,k} X_{i,l} \big(F_{n,i}\big)_{kl}\Bigg) \mathrm{v}\big(\mathbf{X_i}\big),
\end{aligned}$$

where

$$F_{n,i} = \big(\mathbf{A_i} - \mu^*\big)\big(\mu^* - \widehat{\mu}_n\big)^\top + \big(\mu^* - \widehat{\mu}_n\big)\big(\mathbf{A_i} - \mu^*\big)^\top.$$

Hence it is

$$\begin{aligned}
Z_n^{\sigma,4} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg(2 \sum_{k,l=1}^{p} X_{i,k} X_{i,l} \left(\mu_k^* - \widehat{\mu}_{n,k}\right)\big(A_{i,l} - \mu_l^*\big)\Bigg) \mathrm{v}\big(\mathbf{X_i}\big) \\
&= \sum_{k,l=1}^{p} \sqrt{n}\left(\widehat{\mu}_{n,k} - \mu_k^*\right) \Bigg(-\frac{2}{n} \sum_{i=1}^{n} X_{i,k} X_{i,l} \left(A_{i,l} - \mu_l^*\right) \mathrm{v}\big(\mathbf{X_i}\big)\Bigg). \qquad (5.47)
\end{aligned}$$

Similarly to the proof of Lemma 5.29 we consider the random vectors

$$Q_i^{k,l} = -2\, X_{i,k} X_{i,l} \left(A_{i,l} - \mu_l^*\right) \mathrm{v}\big(\mathbf{X_i}\big)$$

for $k, l \in \{1, \ldots, p\}$, which are independent and identically distributed since $(\mathbf{X_1}^\top, \mathbf{A_1}^\top)^\top$, $\ldots, (\mathbf{X_n}^\top, \mathbf{A_n}^\top)^\top$ enjoy this property. Furthermore, we obtain

$$\mathbb{E}\Big[\Big\|Q_i^{k,l}\Big\|_2\Big] \leq \mathbb{E}\Big[\Big\|Q_i^{k,l}\Big\|_1\Big] = 2 \sum_{r=1}^{\frac{p(p+1)}{2}} \mathbb{E}\Big[\Big|X_{i,k} X_{i,l}\, \mathrm{v}\big(\mathbf{X_i}\big)_r\Big|\Big]\, \mathbb{E}\Big[\big|A_{i,l} - \mu_l^*\big|\Big] < \infty$$

by the Assumptions (A4) and (A5), and hence the strong law of large numbers together with Lemma 5.19 implies

$$\left\| -\frac{2}{n} \sum_{i=1}^{n} X_{i,k} X_{i,l} \left(A_{i,l} - \mu_l^*\right) \mathrm{v}\big(\mathbf{X_i}\big) - \mathbb{E}\Big[Q_1^{k,l}\Big] \right\|_\infty = \left\| \frac{1}{n} \sum_{i=1}^{n} Q_i^{k,l} - \mathbb{E}\Big[Q_1^{k,l}\Big] \right\|_\infty \xrightarrow{a.s.} 0$$

with

$$\mathbb{E}\Big[Q_1^{k,l}\Big] = -2\,\mathbb{E}\Big[X_{1,k}X_{1,l}\,\mathrm{v}\big(\mathbf{X_1}\big)\Big]\,\mathbb{E}\big[A_{1,l}-\mu_l^*\big] = \mathbf{0}_{\frac{p(p+1)}{2}}$$

for all $k,l \in \{1,\dots,p\}$ since $\mathbf{X_1}$ and $\mathbf{A_1}$ are independent and $\mathbb{E}[\mathbf{A_1}] = \mu^*$ holds. In particular, this leads to

$$-\frac{2}{n}\sum_{i=1}^{n} X_{i,k}X_{i,l}\left(A_{i,l}-\mu_l^*\right)\mathrm{v}\big(\mathbf{X_i}\big) = \mathrm{o}_{\mathbb{P}}\left(1\right)$$

for all $k,l \in \{1,\dots,p\}$. Moreover, the remaining factors in the sum in (5.47) satisfy

$$\sqrt{n}\left(\widehat{\mu}_{n,k}-\mu_k^*\right) = \mathcal{O}_{\mathbb{P}}\left(1\right)$$

by assumption. So all in all we obtain the assertion

$$Z_n^{\sigma,4} = \sum_{k,l=1}^{p} \mathcal{O}_{\mathbb{P}}\left(1\right)\mathrm{o}_{\mathbb{P}}\left(1\right) = \mathrm{o}_{\mathbb{P}}\left(1\right).$$

□

6. High-dimensional variable selection in random coefficient regression models

In this chapter we consider the variable selection for the means, variances and covariances of the random coefficients in the linear regression model (2.6) in the high-dimensional framework, that means that the number p of coefficients is at least of the order of the sample size n. As we have seen in Section 5.1.1, the errors in the linear regression model of the first moments of the random coefficients are independent and heteroscedastic. Hence, we can immediately apply the theory for the adaptive LASSO Huber estimator, which is provided and discussed in Chapter 3, and is partially motivated by the problem of selecting the deterministic, random or correlated coefficients, to perform variable selection for their means. Since the response variables in the linear regression model of the variances and covariances of the coefficients, established in Section 5.2.1, include also the estimation error of the first stage mean regression and, moreover, the corresponding covariates consist of squares and products of the observed explanatory variables, we obtain a more complicated heteroscedastic mean regression model for the second central moments. Hence it is not possible to deduce results on variable selection for the variances and covariances of the coefficients immediately from the theory in Chapter 3 in general.

This chapter is structured as follows. In Section 6.1 we provide sign-consistency of the adaptive LASSO Huber estimator of the means of the coefficients, and also bounds for the ℓ_∞ norm of respective the estimation error. In Section 6.2 we discuss issues and extensions of the theory in Chapter 3 which are required to establish variable selection for the variances and covariances in a general high-dimensional framework as well. Moreover, we consider the special case where the first moments of the coefficients are known in advance and can be used for the estimation of the second central moments.

6.1. First moments

Remember the linear regression model (5.2) of the means μ^* of the random coefficients,

$$Y_i = \mathbf{X_i}^\top \mu^* + \mathbf{X_i}^\top \big(\mathbf{A_i} - \mu^*\big)\,, \quad i = 1, \dots, n\,, \tag{6.1}$$

established in Section 5.1.1. Note that $(\mathbf{X_1}^\top, \mathbf{A_1}^\top)^\top, \dots, (\mathbf{X_n}^\top, \mathbf{A_n}^\top)^\top$ are assumed to be independent and identically distributed, and, furthermore, that the errors are centered,

$$\mathbb{E}\Big[\mathbf{X_i}^\top \big(\mathbf{A_i} - \mu^*\big) \,\Big|\, \mathbf{X_i}\Big] = \mathbf{X_i}^\top \big(\mathbb{E}[\mathbf{A_i}] - \mu^*\big) = 0$$

since $\mathbf{X_i}$ and $\mathbf{A_i}$ are independent and $\mu^* = \mathbb{E}[\mathbf{A_i}]$ holds.

Firstly, we make similar assumptions in the above regression model (6.1), as in Assumption 3.1, to provide results for the adaptive LASSO Huber estimator of the means μ^* in the following.

Assumption 6.1.

(i) For $m = 2$ or $m = 3$ and $q > 1$ we have that

$$\mathbb{E}\Big[\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\mathbf{A_1} - \mu^*)\big|^m \,\Big|\, \mathbf{X_1}\Big]^q\Big] \le C_{\mu,\mathrm{m}} < \infty\,,$$

where $C_{\mu,\mathrm{m}} > 0$ is a positive constant.

(ii) For positive constants $0 < c_{\mathbf{X},\mathrm{l}} \le c_{\mathbf{X},\mathrm{u}}$ we have that $c_{\mathbf{X},\mathrm{l}} \le \lambda_{\min}\big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]\big) \le \lambda_{\max}\big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]\big) \le c_{\mathbf{X},\mathrm{u}} < \infty$, where $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimal and maximal eigenvalues of a symmetric matrix $M \in \mathbb{R}^{d\times d}$.

(iii) For any $v \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$ the variable $v^\top\mathbf{X_1}$ is sub-Gaussian with variance proxy at most $c^2_{\mathbf{X},\mathrm{sub}}\|v\|_2^2$, $c^2_{\mathbf{X},\mathrm{sub}} > 0$, that is $\mathbb{P}\big(|v^\top\mathbf{X_1}| \ge t\big) \le 2\,\exp\big(-t^2/(2\,c^2_{\mathbf{X},\mathrm{sub}}\|v\|_2^2)\big)$ for all $t \ge 0$.

(iv) We have the a-priori upper bound $\|\mu^*\|_2 \le C_\mu/2$, where $C_\mu \ge 1/8$ is a numerical constant.

Remark 6.2. If we consider $m = 2$ in the first assumption (i) we obtain

$$\begin{aligned}\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\mathbf{A_1} - \mu^*)\big|^2 \,\Big|\, \mathbf{X_1}\Big] &= \mathbb{E}\Big[\sum_{k,l=1}^p X_{1,k}X_{1,k}\big(A_{1,k} - \mu_k^*\big)\big(A_{1,k} - \mu_k^*\big) \,\Big|\, \mathbf{X_1}\Big]\\ &= \sum_{k,l=1}^p X_{1,k}X_{1,k}\,\mathbb{E}\Big[\big(A_{1,k} - \mu_k^*\big)\big(A_{1,k} - \mu_k^*\big)\Big]\\ &= \mathbf{X_1}^\top\Sigma^*\,\mathbf{X_1}\end{aligned}$$

with $\Sigma^* = \mathbb{C}\mathrm{ov}(\mathbf{A_1})$. Moreover, note that $\mathbf{X_1}^\top\Sigma^*\,\mathbf{X_1} = \mathrm{v}(\mathbf{X_1})^\top\sigma^*$, where the half-vectorization σ^* of the covariance matrix Σ^* of the random coefficients is given in (5.1) and the associated vector transformation v in (4.5), holds true. Hence it follows that

$$\mathbb{E}\Big[\mathbb{E}\Big[\big|\mathbf{X_1}^\top(\mathbf{A_1} - \mu^*)\big|^2 \,\Big|\, \mathbf{X_1}\Big]^q\Big] = \mathbb{E}\Big[\big(\mathrm{v}(\mathbf{X_1})^\top\sigma^*\big)^q\Big] = \mathbb{E}\Bigg[\bigg(\sum_{r\in S_\sigma}\mathrm{v}(\mathbf{X_1})_r\sigma_r^*\bigg)^q\Bigg]\,,$$

where S_σ denotes the support of σ^*. As a consequence, if we assume that the variances and covariances of the coefficients are uniformly bounded, then the above expected value depends on the cardinality s_σ of the set S_σ since the moments of the regressors $\mathbf{X_1}$ are bounded by (iii) of Assumption 6.1, cf. Rigollet and Hütter (2019, Lemma 1.4), as well. Thus, the first part (i) of Assumption 6.1 with $m = 2$ is satisfied if the number s_σ of second central moments unequal to zero does not grow with the number p of coefficients. Otherwise, the expected value may grow with rate s_σ^q.

We consider in the following the adaptive LASSO Huber estimator of the means,

$$\widehat{\mu}_n^{\mathrm{ALH}} \in \underset{\beta \in \mathbb{R}^p,\, \|\beta\|_2 \le C_\mu}{\arg\min} \left(\frac{1}{n} \sum_{i=1}^n l_{\alpha_n^\mu}\big(Y_i - \mathbf{X_i}^\top \beta\big) + \lambda_n^\mu \sum_{k=1}^p w_k \, |\beta_k| \right), \tag{6.2}$$

with regularization parameter $\lambda_n^\mu > 0$, robustification parameter $\alpha_n^\mu > 0$ and random weights

$$w_k = \max\Big\{1/\big|\widehat{\mu}_{n,k}^{\mathrm{init}}\big|, 1\Big\}, \qquad k = 1, \dots, p\,,$$

where $\widehat{\mu}_n^{\mathrm{init}} = (\widehat{\mu}_{n,1}^{\mathrm{init}}, \dots, \widehat{\mu}_{n,p}^{\mathrm{init}})^\top \in \mathbb{R}^p$ is a suitable initial estimator of μ^* and the pseudo Huber loss l_α is defined in (3.2). Here, if $\big|\widehat{\mu}_{n,k}^{\mathrm{init}}\big| = 0$, we require that $\beta_k = 0$ in (6.2). In the subsequent corollary we assume that the initial estimator $\widehat{\mu}_n^{\mathrm{init}}$ in the adaptive LASSO satisfies

$$\big\|\widehat{\mu}_n^{\mathrm{init}} - \mu^*\big\|_2 \le C_{\mathrm{init}}\, \lambda_n^{\mu,\mathrm{init}} \sqrt{s_\mu}\,, \qquad \big\|\widehat{\mu}_n^{\mathrm{init}} - \mu^*\big\|_1 \le C_{\mathrm{init}}\, \lambda_n^{\mu,\mathrm{init}}\, s_\mu \tag{6.3}$$

with

$$\lambda_n^{\mu,\mathrm{init}} \simeq \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}} \tag{6.4}$$

for a positive constant $C_{\mathrm{init}} \ge 1$. Under Assumption 6.1 the original LASSO Huber estimator given as a solution of

$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{n} \sum_{i=1}^n \tilde{l}_{\alpha_n}\big(Y_i - \mathbf{X_i}^\top \beta\big) + \lambda_n \sum_{k=1}^p |\beta_k| \right)$$

with Huber loss $\tilde{l}_\alpha$ defined in (2.5), achieves the upper bounds in (6.3) with probability at least $1 - 3/p$ if the orders of λ_n, α_n and n are chosen appropriately, see Section 3.2 for more details. Furthermore, we denote by

$$\mu^*_{\min} := \min_{k \in S_\mu} |\mu_k^*|$$

the smallest absolute value of the mean vector μ^* on its support S_μ. Now we can state our result on sign-consistency and convergence rates in the ℓ_∞ norm of the adaptive LASSO $\widehat{\mu}_n^{\mathrm{ALH}}$ of the first moments μ^* of the random coefficients. The following corollary is an immediate consequence of Theorem 3.2.

Corollary 6.3 (Sign-consistency and ℓ_∞ rate in the mean regression)**.** *In model* (6.1) *under Assumption 6.1, consider the adaptive LASSO estimator* $\widehat{\mu}_n^{\mathrm{ALH}}$ *with initial estimator* $\widehat{\mu}_n^{\mathrm{init}}$ *assumed to satisfy* (6.3)*. Further, suppose that*

$$\left\| \Big(\mathbb{E}\big[\mathbf{X_1}\mathbf{X_1}^\top\big]_{S_\mu S_\mu} \Big)^{-1} \right\|_{\mathrm{M},\infty} \le C_{S_\mu, \mathbf{X}}\,, \tag{6.5}$$

where $C_{S_\mu,\mathbf{X}} > 0$ is a positive constant, is also satisfied. Assume that the robustification parameter α_n^μ for the adaptive LASSO is chosen of the order

$$\alpha_n^\mu \simeq \left(\frac{\log(p)}{n}\right)^{\frac{1}{2}},$$

and that the regularization parameter λ_n^μ is chosen of order

$$\lambda_n^\mu \simeq \lambda_n^{\mu,\mathrm{init}} \left(\frac{|\overline{S}|\log(p)}{n}\right)^{\frac{1}{2}}, \qquad \textit{where} \quad \overline{S} = \left\{k \in \{1,\dots,p\} \,\middle|\, \left|\widehat{\mu}_{n,k}^{\mathrm{init}}\right| > \lambda_n^{\mu,\mathrm{init}}\right\}$$

and $\lambda_n^{\mu,\mathrm{init}} \simeq (\log(p)/n)^{\frac{1}{2}}$ is as in (6.4)*. If $n \gtrsim s_\mu^2 \log(p)$ and if μ^* satisfies a minimum condition of order $\mu_{\min}^* \gtrsim s_\mu\, \lambda_n^{\mu,\mathrm{init}}$, then with probability at least*

$$1 - c_1 \exp(-c_2 n) - \frac{c_3}{p^2},$$

where $c_1, c_2, c_3 > 0$ are suitable constants, the adaptive LASSO Huber estimator $\widehat{\mu}_n^{\mathrm{ALH}}$ as a solution to (6.2) *is unique and satisfies*

$$\operatorname{sign}\big(\widehat{\mu}_n^{\mathrm{ALH}}\big) = \operatorname{sign}\big(\mu^*\big) \qquad \text{and} \qquad \big\|\widehat{\mu}_n^{\mathrm{ALH}} - \mu^*\big\|_\infty \lesssim \lambda_n^{\mu,\mathrm{init}}. \tag{6.6}$$

If we drop assumption (6.5) *but instead have $s_\mu \le \log(p)$, then we retain the sign-consistency in* (6.6) *but only obtain a ℓ_∞-rate of order*

$$\big\|\widehat{\mu}_n^{\mathrm{ALH}} - \mu^*\big\|_\infty \lesssim \sqrt{s_\mu}\, \lambda_n^{\mu,\mathrm{init}}.$$

Remark 6.4. As previously discussed in Remark 6.2 the upper bound in the first part (i) of Assumption 6.1 may depend on s_σ, the number of variances and covariances of the coefficients unequal to zero. Then the orders in Corollary 6.3 depend on s_σ as well, and, in particular, this leads to a additional factor of $\sqrt{s_\sigma}$ in the orders of $\mu_{\min}^*$ and the ℓ_∞ norm of the estimation error. See also Remark 3.3 for further discussion on the conditions of Corollary 6.3.

6.2. Second central moments

Remember the linear regression model (5.14) of the covariance matrix Σ^*, respectively of its half-vectorization σ^*, of the random coefficients,

$$\big(Y_i - \mathbf{X_i}^\top \widehat{\mu}_n\big)^2 = Y_i^\sigma = \mathrm{v}\big(\mathbf{X_i}\big)^\top \sigma^* + \mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^* + E_n + F_{n,i}\big), \quad i = 1,\dots,n, \tag{6.7}$$

established in Section 5.2.1, with

$$\begin{aligned} D_i &= \big(\mathbf{A_i} - \mu^*\big)\big(\mathbf{A_i} - \mu^*\big)^\top, \qquad E_n = \big(\mu^* - \widehat{\mu}_n\big)\big(\mu^* - \widehat{\mu}_n\big)^\top, \\ F_{n,i} &= \big(\mathbf{A_i} - \mu^*\big)\big(\mu^* - \widehat{\mu}_n\big)^\top + \big(\mu^* - \widehat{\mu}_n\big)\big(\mathbf{A_i} - \mu^*\big)^\top, \end{aligned}$$

and $\widehat{\mu}_n$ is an estimator of the first moments μ^* of the coefficients based on the observations $(Y_1, \mathbf{X_1}^\top)^\top, \ldots, (Y_n, \mathbf{X_n}^\top)^\top$. Evidently, the heteroscedastic errors are in general mutually correlated since they all depend on the estimator $\widehat{\mu}_n$ of the means, see also Section 5.2.2 for further discussion on the error structure.

6.2.1. Special case: known means

Firstly, we consider in the following the special case where the means μ^* of the random coefficients are known in advance. Then we can set $\widehat{\mu}_n = \mu^*$ in model (6.7) and obtain the simplified linear regression model

$$\big(Y_i - \mathbf{X_i}^\top \mu^*\big)^2 = Y_i^\sigma = \mathrm{v}\big(\mathbf{X_i}\big)^\top \sigma^* + \mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^*\big)\,, \quad i = 1, \ldots, n\,. \tag{6.8}$$

Here the errors are obviously centered,

$$\mathbb{E}\Big[\mathrm{v}\big(\mathbf{X_i}\big)^\top \mathrm{vec}\big(D_i - \Sigma^*\big) \,\Big|\, \mathrm{v}\big(\mathbf{X_i}\big)\Big] = \mathrm{v}\big(\mathbf{X_i}\big)^\top \Big(\mathbb{E}\big[\mathrm{vec}(D_i)\big] - \sigma^*\Big) = 0\,,$$

since $\mathbf{X_i}$ and $\mathbf{A_i}$ are independent and $\sigma^* = \mathrm{vec}(\Sigma^*) = \mathbb{E}\big[\mathrm{vec}(D_i)\big]$ holds. Moreover, the response variables Y_i^σ are independent and identically distributed. Thus, we shall formulate an analogous result to Corollary 6.3 for the variable selection for the variances and covariances of the coefficients in this framework, which is an immediate consequence of Theorem 3.2 as well.

We make the following assumptions in the linear regression model (6.8).

Assumption 6.5.

(i) For $m = 2$ or $m = 3$ and $q > 1$ we have that

$$\mathbb{E}\bigg[\mathbb{E}\Big[\big|\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(D_1 - \Sigma^*\big)\big|^m \,\Big|\, \mathbf{X_1}\Big]^q\bigg] \le C_{\sigma,\mathrm{m}} < \infty\,,$$

where $C_{\sigma,\mathrm{m}} > 0$ is a positive constant.

(ii) For positive constants $0 < c_{\mathrm{v}(\mathbf{X}),\mathrm{l}} \le c_{\mathrm{v}(\mathbf{X}),\mathrm{u}}$ we have that

$$c_{\mathrm{v}(\mathbf{X}),\mathrm{l}} \le \lambda_{\min}\Big(\mathbb{E}\big[\mathrm{v}\big(\mathbf{X_1}\big)\,\mathrm{v}\big(\mathbf{X_1}\big)^\top\big]\Big) \le \lambda_{\max}\Big(\mathbb{E}\big[\mathrm{v}\big(\mathbf{X_1}\big)\,\mathrm{v}\big(\mathbf{X_1}\big)^\top\big]\Big) \le c_{\mathrm{v}(\mathbf{X}),\mathrm{u}} < \infty\,,$$

where $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimal and maximal eigenvalues of a symmetric matrix $M \in \mathbb{R}^{d\times d}$.

(iii) For any $v \in \mathbb{R}^{\frac{p(p+1)}{2}} \setminus \{\mathbf{0}_{\frac{p(p+1)}{2}}\}$ the variable $v^\top \mathrm{v}(\mathbf{X_1})$ is sub-Gaussian with variance proxy at most $c^2_{\mathrm{v}(\mathbf{X}),\mathrm{sub}}\|v\|_2^2$, $c^2_{\mathrm{v}(\mathbf{X}),\mathrm{sub}} > 0$, that is

$$\mathbb{P}\big(|v^\top \mathrm{v}(\mathbf{X_1})| \ge t\big) \le 2\,\exp\big(-t^2/(2\,c^2_{\mathrm{v}(\mathbf{X}),\mathrm{sub}}\|v\|_2^2)\big)$$

for all $t \ge 0$.

(iv) We have the a-priori upper bound $\|\sigma^*\|_2 \le C_\sigma/2$, where $C_\sigma \ge 1/8$ is a numerical constant.

Remark 6.6. The sub-Gaussian tail bound on the vector transformation $\mathrm{v}(\mathbf{X_1})$ of the regressors $\mathbf{X_1}$ in the third part (iii) of Assumption 6.5 is rather restrictive. For example, if the covariates are independent and normally distributed, the squares and mixed products, which are contained in $\mathrm{v}(\mathbf{X_1})$ have a sub-Exponential tail behavior. However, it should hold for independent and uniformly bounded explanatory variables. Take also notice of the discussion in Remark 6.2 about the first part (i) of Assumption 6.5.

Let

$$\widehat{\sigma}_n^{\mathrm{ALH}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{\frac{p(p+1)}{2}},\, \|\beta\|_2 \le C_\sigma} \left(\frac{1}{n} \sum_{i=1}^n l_{\alpha_n^\sigma}\big(Y_i^\sigma - \mathrm{v}(\mathbf{X_i})^\top \beta\big) + \lambda_n^\sigma \sum_{k=1}^{\frac{p(p+1)}{2}} w_k\,|\beta_k| \right) \tag{6.9}$$

be the adaptive LASSO Huber estimator of the variances and covariances with regularization parameter $\lambda_n^\sigma > 0$, robustification parameter $\alpha_n^\sigma > 0$ and random weights

$$w_k = \max\Big\{1/\big|\widehat{\sigma}_{n,k}^{\mathrm{init}}\big|, 1\Big\}, \qquad k = 1, \dots, p(p+1)/2\,,$$

where $\widehat{\sigma}_n^{\mathrm{init}} = (\widehat{\sigma}_{n,1}^{\mathrm{init}}, \dots, \widehat{\sigma}_{n,p(p+1)/2}^{\mathrm{init}})^\top \in \mathbb{R}^{\frac{p(p+1)}{2}}$ is a suitable initial estimator of σ^* and l_α is the pseudo Huber loss defined in (3.2). Here, if $\big|\widehat{\sigma}_{n,k}^{\mathrm{init}}\big| = 0$, we require that $\beta_k = 0$ in (6.9). In the subsequent corollary we assume that the initial estimator $\widehat{\sigma}_n^{\mathrm{init}}$ in the adaptive LASSO satisfies

$$\big\|\widehat{\sigma}_n^{\mathrm{init}} - \sigma^*\big\|_2 \le C_{\mathrm{init}}\, \lambda_n^{\sigma,\mathrm{init}} \sqrt{s_\sigma}\,, \qquad \big\|\widehat{\sigma}_n^{\mathrm{init}} - \sigma^*\big\|_1 \le C_{\mathrm{init}}\, \lambda_n^{\sigma,\mathrm{init}}\, s_\sigma \tag{6.10}$$

with

$$\lambda_n^{\sigma,\mathrm{init}} \simeq \left(\frac{\log\big(\frac{p(p+1)}{2}\big)}{n} \right)^{\frac{1}{2}} \tag{6.11}$$

for a positive constant $C_{\mathrm{init}} \ge 1$. Under Assumption 6.5 the original LASSO Huber estimator given as a solution of

$$\operatorname*{arg\,min}_{\beta \in \mathbb{R}^{\frac{p(p+1)}{2}}} \left(\frac{1}{n} \sum_{i=1}^n \tilde{l}_{\alpha_n}\big(Y_i^\sigma - \mathrm{v}(\mathbf{X_i})^\top \beta\big) + \lambda_n \sum_{k=1}^{\frac{p(p+1)}{2}} |\beta_k| \right)$$

with Huber loss $\tilde{l}_\alpha$ defined in (2.5), achieves the upper bounds in (6.10) in the linear regression model (6.8) with probability at least $1 - 6/\big(p(p+1)\big)$ if the orders of λ_n, α_n and n are chosen appropriately, see Section 3.2 for more details. Furthermore, we denote by

$$\sigma_{\min}^* := \min_{k \in S_\sigma} \big|\sigma_k^*\big|$$

the smallest absolute value of the covariance matrix Σ^* of the coefficients on its support S_σ. Now we state our result on sign-consistency and convergence rates in the ℓ_∞ norm of the adaptive LASSO $\widehat{\sigma}_n^{\mathrm{ALH}}$ of the second central moments σ^* of the random coefficients. The following corollary is an immediate consequence of Theorem 3.2 as well.

Corollary 6.7 (Sign-consistency and ℓ_∞ rate in the variance/covariances regression with known means)**.** *In model* (6.8) *under Assumption 6.5, consider the adaptive LASSO estimator $\widehat{\sigma}_n^{\mathrm{ALH}}$ with initial estimator $\widehat{\sigma}_n^{\mathrm{init}}$ assumed to satisfy* (6.10)*. Further, suppose that*

$$\left\| \left(\mathbb{E}\Big[\mathrm{v}\big(\mathbf{X_1}\big)\, \mathrm{v}\big(\mathbf{X_1}\big)^\top \Big]_{S_\sigma S_\sigma} \right)^{-1} \right\|_{\mathrm{M},\infty} \leq C_{S_\sigma,\mathrm{v}(\mathbf{X})}\,, \qquad (6.12)$$

where $C_{S_\sigma,\mathrm{v}(\mathbf{X})} > 0$ is a positive constant, is also satisfied. Assume that the robustification parameter α_n^σ for the adaptive LASSO is chosen of the order

$$\alpha_n^\sigma \simeq \left(\frac{\log\left(\frac{p(p+1)}{2}\right)}{n} \right)^{\frac{1}{2}},$$

and that the regularization parameter λ_n^σ is chosen of order

$$\lambda_n^\sigma \simeq \lambda_n^{\sigma,\mathrm{init}} \left(\frac{\big|\overline{S}\big| \log\left(\frac{p(p+1)}{2}\right)}{n} \right)^{\frac{1}{2}}, \quad \textit{where } \overline{S} = \left\{ k \in \left\{1, \ldots, \frac{p(p+1)}{2}\right\} \,\middle|\, \big|\widehat{\sigma}_{n,k}^{\mathrm{init}}\big| > \lambda_n^{\sigma,\mathrm{init}} \right\}$$

and $\lambda_n^{\sigma,\mathrm{init}} \simeq \big(\log\left(\frac{p(p+1)}{2}\right)/n\big)^{\frac{1}{2}}$ is as in (6.11)*. If $n \gtrsim s_\sigma^2 \log\left(\frac{p(p+1)}{2}\right)$ and if σ^* satisfies a minimum condition of order $\sigma^*_{\min} \gtrsim s_\sigma\, \lambda_n^{\sigma,\mathrm{init}}$, then with probability at least*

$$1 - c_1 \exp(-c_2 n) - \frac{c_3}{\big(p(p+1)\big)^2}\,,$$

where $c_1, c_2, c_3 > 0$ are suitable constants, the adaptive LASSO Huber estimator $\widehat{\sigma}_n^{\mathrm{ALH}}$ as a solution to (6.9) *is unique and satisfies*

$$\mathrm{sign}\big(\widehat{\sigma}_n^{\mathrm{ALH}}\big) = \mathrm{sign}\big(\sigma^*\big) \qquad \text{and} \qquad \big\|\widehat{\sigma}_n^{\mathrm{ALH}} - \sigma^*\big\|_\infty \lesssim \lambda_n^{\sigma,\mathrm{init}}\,. \qquad (6.13)$$

If we drop assumption (6.12) *but instead have $s_\sigma \leq \log\left(\frac{p(p+1)}{2}\right)$, then we retain the sign-consistency in* (6.13) *but only obtain a ℓ_∞-rate of order*

$$\big\|\widehat{\sigma}_n^{\mathrm{ALH}} - \sigma^*\big\|_\infty \lesssim \sqrt{s_\sigma}\, \lambda_n^{\sigma,\mathrm{init}}.$$

6.2.2. General case

Now we come back to the general linear regression model (6.7) of the covariance matrix of the random coefficients where the response variables additionally depend on the estimation error of the first stage mean regression. To get similar results as in Corollary 6.7

in this setting, some of the main steps in Chapter 3 have to be modified. This concerns all results considering the gradient of the empirical pseudo Huber loss, in particular, Lemma 3.6 for establishing the restricted strong convexity, and the Lemmas 3.8, 3.13 and 3.14 for bounding the ℓ_∞ norm of (a transformation of) the gradient. One has to pay attention to the specific error structure in model (6.7) to adapt the proofs appropriately. For this purpose also the results of the first stage mean regression in Corollary 6.3 are crucial.

A second challenge is to weaken the light tail assumption on the vector transformation of the regressors in the third part (iii) of Assumption 6.5. As mentioned in Remark 6.6, for instance for independent and normally distributed covariates this condition is not satisfied. Recent literature on this topic suggests robustification of the (potentially) heavy-tailed regressors by truncating them as well. For this purpose an additional tuning parameter $\omega > 0$ is used, see Fan et al. (2016, Section 3.1) and Sun et al. (2020, Section 4) for more details. Another approach is to use influence/weight functions for the covariates to shrink large values, cf. for example Loh (2017, Section 2.2).

To sum up, one could say that high-dimensional variable selection for the variances and covariances of the random coefficients in the general model (6.7) can not be solved completely with the results in this thesis yet. However, we provided and discussed essential approaches to tackle this problem in the future.

Bibliography

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 367–389.

Belloni, A. and V. Chernozhukov (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics 39*(1), 82–130.

Beran, R., A. Feuerverger, and P. Hall (1996). On nonparametric estimation of intercept and slope distributions in random coefficient regression. *The Annals of Statistics 24*(6), 2569–2592.

Beran, R. and P. Hall (1992). Estimating coefficient distributions in random coefficient regressions. *The Annals of Statistics 20*(4), 1970–1984.

Beran, R. and P. W. Millar (1994). Minimum distance estimation in random coefficient regression models. *The Annals of Statistics 22*(6), 1976–1992.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'IHP Probabilités et statistiques 48*(4), 1148–1185.

Charbonnier, P., L. Blanc-Feraud, G. Aubert, and M. Barlaud (1994). Two deterministic half-quadratic regularization algorithms for computed imaging. *Proceedings of 1st International Conference on Image Processing 2*, 168–172.

Dunker, F., K. Eckle, K. Proksch, and J. Schmidt-Hieber (2019). Tests for qualitative features in the random coefficients model. *Electronic Journal of Statistics 13*(2), 2257–2306.

Fan, J., Y. Fan, and E. Barut (2014). Adaptive robust variable selection. *The Annals of Statistics 42*(1), 324–351.

Fan, J., Q. Li, and Y. Wang (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 247–265.

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

Fan, J., H. Liu, Q. Sun, and T. Zhang (2018). I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics 46*(2), 814–841.

Fan, J., W. Wang, and Z. Zhu (2016). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315*.

Foucart, S. and H. Rauhut (2013). *A mathematical introduction to compressive sensing.* Springer.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gautier, E. and S. Hoderlein (2011). A triangular treatment effect model with random coefficients in the selection equation. *arXiv preprint arXiv:1109.0362*.

Gautier, E. and Y. Kitamura (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica 81*(2), 581–607.

Giraud, C. (2014). *Introduction to high-dimensional statistics*, Volume 138. CRC Press.

Gu, Y. and H. Zou (2016). High-dimensional generalizations of asymmetric least squares regression and their applications. *The Annals of Statistics 44*(6), 2661–2694.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations.* CRC Press.

Hermann, P. and H. Holzmann (2020). Support estimation in high-dimensional heteroscedastic mean regression. *arXiv preprint arXiv:2011.01591*.

Hildreth, C. and J. P. Houck (1968). Some estimators for a linear model with random coefficients. *Journal of the American Statistical Association 63*(322), 584–595.

Hoderlein, S., H. Holzmann, and A. Meister (2017). The triangular model with random coefficients. *Journal of Econometrics 201*(1), 144–169.

Hoderlein, S., J. Klemelä, and E. Mammen (2010). Analyzing the random coefficient model nonparametrically. *Econometric Theory 26*(3), 804–837.

Holzmann, H. and A. Meister (2020). Rate-optimal nonparametric estimation for random coefficient regression models. *Bernoulli 26*(4), 2790–2814.

Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica 18*(4), 1603–1618.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*(1), 73–101.

Ichimura, H. and T. S. Thompson (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics 86*(2), 269–295.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*(5), 1356–1378.

Lambert-Lacroix, S. and L. Zwald (2011). Robust regression through the huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics 5*, 1015–1053.

Lederer, J. and M. Vogt (2020). Estimating the lasso's effective noise. *arXiv preprint arXiv:2004.11554*.

Lewbel, A. and K. Pendakur (2017). Unobserved preference heterogeneity in demand using generalized random coefficients. *Journal of Political Economy 125*(4), 1100–1148.

Li, Y. and J. Zhu (2008). l_1-norm quantile regression. *Journal of Computational and Graphical Statistics 17*(1), 163–185.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics 45*(2), 866–896.

Loh, P.-L. and M. J. Wainwright (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics 45*(6), 2455–2482.

Massart, P. (2007). *Concentration inequalities and model selection*, Volume 6. Springer.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics 37*(1), 246–270.

Ndaoud, M. and A. B. Tsybakov (2020). Optimal variable selection and adaptive noisy compressed sensing. *IEEE Transactions on Information Theory 66*(4), 2517–2532.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science 27*(4), 538–557.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society 55*(4), 819–847.

Rigollet, P. and J.-C. Hütter (2019). High dimensional statistics. *Lecture notes for course 18S997*.

Ruszczynski, A. (2006). *Nonlinear Optimization.* Princeton University Press.

Sun, Q., W.-X. Zhou, and J. Fan (2020). Adaptive huber regression. *Journal of the American Statistical Association 115*(529), 254–265.

Swami, P. (1970). Efficient inference in a random coefficients model. *Econometrica 38*(2), 311–324.

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

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics 7*, 1456–1490.

van de Geer, S., P. Bühlmann, and S. Zhou (2011). The adaptive and the thresholded lasso for potentially misspecified models. *Electronic Journal of Statistics 5*, 688–749.

van de Geer, S. and P. Bühlmann (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics 3*, 1360–1392.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge University Press.

Wagener, J. and H. Dette (2012). Bridge estimators and the adaptive lasso under heteroscedasticity. *Mathematical Methods of Statistics 21*(2), 109–126.

Wagener, J. and H. Dette (2013). The adaptive lasso in high-dimensional sparse heteroscedastic models. *Mathematical Methods of Statistics 22*(2), 137–154.

Wainwright, M. J. (2009a). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory 55*(12), 5728–5741.

Wainwright, M. J. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press.

Wang, L. (2013). The l_1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis 120*, 135–151.

Wang, L., Y. Wu, and R. Li (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association 107*(497), 214–222.

Yi, C. and J. Huang (2017). Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics 26*(3), 547–557.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*(90), 2541–2563.

Zhou, S., S. van de Geer, and P. Bühlmann (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Zou, H. and M. Yuan (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics 36*(3), 1108–1126.

Zusammenfassung (deutsch)

Lineare Regressionsmodelle genießen seit einigen Jahrzehnten großes Interesse in der Statistik. Praktische Anwendungen kann man in einer Vielzahl von Bereichen, wie z.B. den Verhaltens- und Sozialwissenschaften, der Finanzwelt und der Ökonometrie, finden. Insbesondere bei der Analyse von Konsumentendaten und in der Medizin können die marginalen Effekte jedoch über die Individuen hinweg variieren. Regressionsmodelle mit zufälligen Koeffizienten sind sehr hilfreich, um diese unbeobachtete Heterogenität zu analysieren und modellieren. Hildreth und Houck (1968) und Swami (1970) betrachteten entsprechende lineare Modelle aus einem parametrischen Standpunkt heraus und schlugen unter der Annahme, dass die Kovarianzen verschwinden, konsistente Schätzer für die Erwartungswerte und Varianzen der zufälligen Regressionskoeffizienten vor. Darüber hinaus wurde in den letzten Jahrzehnten intensive Forschung in dem Gebiet der nichtparametrischen Identifikation und Schätzung der gemeinsamen Verteilung der Koeffizienten betrieben, wie z.B. in Beran und Hall (1992), Beran und Millar (1994), Beran et al. (1996), Hoderlein et al. (2010), Dunker et al. (2019) und Holzmann und Meister (2020). Des Weiteren untersuchten Lewbel und Pendakur (2017) nichtlineare und additive Modelle, Ichimura und Thompson (1998) und Gautier und Kitamura (2013) binäre Modelle, sowie Gautier und Hoderlein (2011) und Hoderlein et al. (2017) Dreiecksmodelle mit zufälligen Koeffizienten.

In der vorliegenden Arbeit betrachten wir das lineare Regressionsmodell mit zufälligen Koeffizienten und insbesondere deren Mittelwerte, Varianzen und Kovarianzen, die möglicherweise in vielen Anwendungen hauptsächlich von Interesse sind, in einem hochdimensionalen Rahmen mit Fokus auf Variablenwahl. Dies bedeutet, dass die Anzahl der Regressoren die Anzahl der Beobachtungen übersteigen kann, aber nur einige wenige erklärende Variablen tatsächlich einen Einfluss und/oder heterogene Effekte haben. Die hochdimensionale Statistik im Allgemeinen hat in den letzten Jahren viel Aufmerksamkeit erlangt, da in vielen industriellen und wissenschaftlichen Bereichen größere Datensätze mit einer großen Anzahl an Merkmalen gesammelt werden, wie z.B. in der funktionellen Magnetresonanztomographie oder auch der Analyse von Microarray- und Konsumentendaten. Ein umfassender Überblick über Methoden und zugehörige Theorie in diesem Themenbereich befindet sich unter anderem in Bühlmann und van de Geer (2011), Giraud (2014), Hastie et al. (2015), Vershynin (2018) und Wainwright (2019).

Eine weit verbreitete und effektive Methode für die Variablenwahl in hochdimensionalen Regressionsmodellen mit dünnbesetzten Parametervektoren sind Schätzer mit Penalisierungsfunktionen. Ein prominentes Beispiel ist der LASSO-Schätzer, welcher zuerst von Tibshirani (1996) vorgeschlagen wurde und die empirische quadratische Verlustfunktion mit der ℓ_1-Penalisierung kombiniert. Orakelungleichungen für den LASSO-Schätzer in linearen Regressionsmodellen mit unabhängigen und normalverteilten Fehlern wer-

den unter anderem in Bickel et al. (2009) und Meinshausen und Yu (2009) bewiesen. Dabei ist immer eine Annahme an die Datenmatrix notwendig; van de Geer und Bühlmann (2009) sowie Foucart und Rauhut (2013) diskutieren verschiedene Annahmen und ihre Beziehung zueinander. Darüber hinaus haben Zhao und Yu (2006) gezeigt, dass für die Vorzeichenkonsistenz des LASSO-Schätzers eine zusätzliche Annahme erforderlich ist, die meistens als wechselseitige Inkohärenzbedingung bezeichnet wird. Außerdem hat in diesem Kontext Wainwright (2009b) die Primal-Dual-Witness-Charakterisierung des LASSO-Schätzers eingeführt, sowie hinreichende und notwendige Bedingungen für die Vorzeichenkonsistenz unter unabhängigen sub-Gaußschen Fehlern gegeben. In einem nachfolgenden Forschungszweig, wie z.B. in Wainwright (2009a), wurden minimale Bedingungen diskutiert, unter denen für bestimmte Datenmatrizen, welche z.B. aus unabhängig und identisch normalverteilten Einträgen bestehen, Vorzeichenkonsistenz für verschiedene Konstellationen von Stichprobengröße, Anzahl der Regressoren, Anzahl von Koeffizienten ungleich null und betragsmäßig kleinstem Eintrag ungleich null des Koeffizientenvektors möglich ist. Umfangreiche Ergebnisse in dieser Richtung, die auch nicht-Gaußsche und endlastige Fehler einbeziehen, befinden sich in Ndaoud und Tsybakov (2020). In einem weiteren Literaturstrang wird versucht sich von der wechselseitigen Inkohärenzbedingung für die Variablenwahl zu befreien. In diesem Zusammenhang hat Zou (2006) den adaptiven LASSO-Schätzer vorgeschlagen. Für wachsende Dimension des Koeffizientenvektors liefern Huang et al. (2008) asymptotische Resultate, Wagener und Dette (2012) sowie Wagener und Dette (2013) erweitern diese Asymptotik für heteroskedastische Fehler. Darüber hinaus betrachten Zhou et al. (2009) und van de Geer et al. (2011) den adaptiven LASSO-Schätzer in hochdimensionalen linearen Regressionsmodellen mit unabhängig und identisch normalverteilten Fehlern. Außerdem bieten Loh und Wainwright (2017) eine hochdimensionale Analyse von nicht-konvexen Penalisierungen, wie z.B. der glatt abgeschnittenen absoluten Abweichung (Fan und Li, 2001, SCAD) oder der minimax-konkav-Penalisierung (Zhang, 2010, MCP), um die Bedingung der wechselseitigen Inkohärenz aufzuheben.

Falls die Anzahl der erklärenden Variablen die Anzahl der Beobachtungen übersteigt, dann basiert der Großteil der oben genannten Resultate auf einer sub-Gaußschen Abschätzung für die Ränder der unabhängigen Fehler im linearen Regressionsmodell. Das Fallenlassen dieser Annahme kann zu suboptimalen Raten für die ℓ_1-, ℓ_2- und ℓ_∞-Norm des Schätzfehlers führen. Aus den Resultaten in Lederer und Vogt (2020) geht jedoch hervor, dass der gewöhnliche LASSO-Schätzer die optimalen Raten aus dem Modell mit leichten Rändern für die Fehlerverteilung beibehält, vorausgesetzt die Regressoren sind gleichmäßig beschränkt und die Fehler haben etwas mehr als ein endliches viertes Moment. Insbesondere im linearen Regressionsmodell mit zufälligen Koeffizienten führt die Schätzung der zugehörigen Kovarianzmatrix zu einem heteroskedastischen Modell, in welchem die Fehler der Regression die zentrierten Quadrate und paarweisen Produkte der Koeffizienten enthalten. Falls wir also eine sub-Gaußsche Verteilung für die Koeffizienten annehmen, dann haben die eben genannten Fehler keine leichten Ränder mehr.

Ein aktueller Forschungszweig beschäftigt sich daher mit der Robustifizierung der verfügbaren Methodik in hochdimensionalen linearen Regressionsmodellen im Hinblick auf Abweichungen von der Annahme von leichten Rändern bei den Fehlern und manchmal auch bei den erklärenden Variablen. Ein gängiger Ansatz besteht darin, die quadratische Verlustfunktion durch eine andere, robuste Verlustfunktion, wie z.B. die Check-Funktion aus der Quantilsregression und insbesondere die absolute Abweichung für den Median (Li und Zhu, 2008; Zou und Yuan, 2008; Belloni und Chernozhukov, 2011; Wang, 2013; Fan et al., 2014), zu ersetzen. Dies führt jedoch vorzugsweise in linearen Regressionsmodellen mit potentiell heteroskedastischen, asymmetrischen Fehlern dazu, dass der Zielparameter geändert wird. Des Weiteren hat Loh (2017) robuste Verlustfunktionen, die den gewünschten Mittelwertparameter liefern, in homoskedastischen Modellen mit unabhängigen erklärenden Variablen und Fehlern analysiert.

Ein anderer Ansatz, welcher von Lambert-Lacroix und Zwald (2011), Fan et al. (2017) und Sun et al. (2020) verfolgt wurde, ist die Verwendung des Huber-Verlustes (Huber, 1964) mit einem zusätzlichen Tuning-Parameter. Der Huber-Verlust kombiniert einen quadratischen Verlust für kleine Werte und einen absoluten Verlust für große Werte. Der Tuning-Parameter, den wir als Robustifizierungsparameter bezeichnen, ist notwendig, um den Bias der Schätzung zu kontrollieren, da der Schätzfehler im Allgemeinen auch immer einen Approximationsfehler enthält. Lambert-Lacroix und Zwald (2011) liefern asymptotische Resultate für den entsprechenden adaptiven LASSO-Schätzer mit einer festen Wahl des Robustifizierungsparameters in linearen Regressionsmodellen mit symmetrischen Fehlern. Wenn der Tuning-Parameter mit einer angemessenen Rate, welche abhängig von der Stichprobengröße und der Dimension des Koeffizientenvektors ist, konvergiert, dann erreicht der LASSO-Schätzer mit Huber-Verlust in hochdimensionalen, heteroskedastischen linearen Regressionsmodellen mit sub-Gaußschen Regressoren und Fehlern mit endlichem zweitem Moment die gleichen Raten in der ℓ_1- und ℓ_2-Norm wie der gewöhnliche LASSO-Schätzer unter homoskedastischen Fehlern mit leichten Rändern. Entsprechende obere und untere Schranken wurden in Fan et al. (2017) und Sun et al. (2020) bewiesen. Allerdings wurde in diesen Modellen unseres Wissens nach die Variablenwahl und die ℓ_∞-Norm des Schätzfehlers noch nicht untersucht. Wir betrachten dazu im Folgenden eine strikt konvexe, glatte Variante des Huber-Verlustes und die adaptive LASSO-Penalisierung, um eine rechentechnische Effizienz zu gewährleisten. Im ersten Teil der vorliegenden Arbeit zeigen wir für den resultierenden Schätzer Vorzeichenkonsistenz und auch optimale Konvergenzraten in der ℓ_∞-Norm, welche aus linearen Regressionsmodellen mit homoskedastischen Fehlern mit leichten Rändern bekannt sind.

Die Arbeit ist wie folgt aufgebaut. Zu Beginn wird die grundlegende Notation eingeführt, welche in der Arbeit benötigt wird. In Kapitel 2 geben wir einen kurzen Überblick über die Vorzeichenkonsistenz in hochdimensionalen, homoskedastischen linearen Regressionsmodellen und motivieren die Notwendigkeit analoger Resultate für heteroskedastische Fehler, um Variablenwahl für die ersten und zweiten Momente in linearen Regressionsmodellen mit zufälligen Koeffizienten durchführen zu können. In Kapitel 3 führen wir den Pseudo-Huber-Verlust ein und zeigen Vorzeichenkonsistenz sowie optimale Raten in der ℓ_∞-Norm für den adaptiven LASSO-Schätzer in heteroskedastischen linearen Regressionsmodellen mit sub-Gaußschen Regressoren und Fehlern mit etwas mehr als einem endlichem zweiten Moment. Simulationen illustrieren die Vorteile der vorgeschlagenen Methodik im Vergleich zum gewöhnlichen adaptiven LASSO-Schätzer. Die Resultate von Kapitel 3 sind auch in Hermann und Holzmann (2020) enthalten. Im

zweiten Teil der vorliegenden Arbeit widmen wir uns dem linearen Regressionsmodell mit zufälligen Koeffizienten und insbesondere deren Erwartungswerten, Varianzen und Kovarianzen. Dazu geben wir in Kapitel 4 erst einmal hinreichende Bedingungen für die Identifizierbarkeit der ersten und zweiten Momente der Koeffizienten an. Dabei konzentrieren wir uns auf Situationen, in denen die Regressoren möglicherweise nur einen beschränkten oder sogar endlichen Träger haben. Dies steht im Gegensatz zu dem großflächigen Träger, der für die nichtparametrische Identifikation der gemeinsamen Verteilung der Koeffizienten notwendig ist. In Kapitel 5 stellen wir zunächst die heteroskedastischen linearen Regressionsmodelle für den dünnbesetzen Vektor der Erwartungswerte und die dünnbesetzte Kovarianzmatrix der zufälligen Koeffizienten auf. Anschließend beweisen wir asymptotische Resultate für die entsprechenden adaptiven LASSO-Schätzer, vorausgesetzt die Anzahl der Koeffizienten ist fest, wobei die Vorzeichenkonsistenz abermals unser Hauptziel ist. Schlussendlich wenden wir in Kapitel 6 die Methoden aus Kapitel 3 auf die hochdimensionalen Regressionsmodelle der Momente der Koeffizienten an und diskutieren ausstehende Probleme.