Unterschrift der Betreuerin

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

D I P L O M A R B E I T

Confidence Sets Based on the Adaptive LASSO Estimator

Ausgeführt am Institut für

Stochastik und Wirtschaftsmathematik
der Technischen Universität Wien

unter der Anleitung von

Assoz. Prof. Mag. Ulrike Schneider, PhD

durch

Nicolai David Amann

Koflergasse 4/9A, 1120 Wien

Datum

Unterschrift

Abstract

This thesis deals examines the adaptive LASSO estimator in the setting of moving parameter in the low-dimensional case, while the tuning parameters may vary over the components. The main part deals with the construction of asymptotic confidence sets based on the adaptive LASSO estimator in the case where at least one component of the tuning parameter is tuned to perform consistent model selection. The asymptotic distribution of the appropriately scaled and centred adaptive LASSO estimator is derived implicitly as the minimizer of a stochastic function, which is used to create confidence sets with – asymptotically – infimal coverage probability of 1. Besides confidence sets of the partially consistent tuned adaptive LASSO estimator, a condition on the tuning parameters is shown to be equivalent to consistency in parameter estimation. Conditions concerning the consistency in model selection are also derived. In particular, obtaining consistency in model selection for the adaptive LASSO estimator requires consistency in parameter estimation.

Acknowledgements

First of all, I would like to thank my parents for your love and support you gave me. Furthermore, your encouragement to study the subject I enthuse most about resulted in mathematics, for which I am also very thankful.

Secondly, I give thanks to my math teacher in highschool, Martin Hölbling. Besides providing me fundamental skills in mathematics, you encouraged my passion in mathematics.

Thirdly, I thank my supervisor Ulrike Schneider for the many fruitful discussions, her numerous helpful comments, her support and time I occupied throughout the thesis as well as her patience.

Another big thanks goes to my colleagues, who brightened up my time during the studies, especially, but not restricted to, the intense bachelor studies. Besides this, I would like to point out Nathanael Skrepek for the helpful insights and discussions. Last, but not least, I thank Magdalena Hutze for the motivation to finish this work.

Contents

1 Introduction

> It is better to know some of the
> questions than all of the answers.
>
> *James Thurber*

The least absolute shrinkage and selection operator (LASSO), as proposed in R. Tibshirani (1996), has been a topic of research in statistics and econometrics as it satisfies several desirable properties. On the one hand it can be used in cases, where the number of explanatory variables exceeds those of observations. On the other hand, it performs model selection and parameter estimation simultaneously. Besides its theoretical properties, the LASSO can be – in terms of computational effort – efficiently calculated for all values of its tuning parameter concurrently via the LARS algorithm introduced by Efron et al. (2004), which supports its use in application. Heading back to theoretical properties, shrinkage operators may increase prediction accuracy in specific cases, but come at the cost of an increasing bias. A prominent example is the ridge estimator, which outperforms the least squares estimator in terms of mean square error in the standard linear regression model when the true parameter is located in the neighbourhood of 0. This estimator has been generalized in Frank and Friedman (1993) to the bridge estimator, where the parameter γ describes the type of the penalizing term's norm. Coinciding with the ridge estimator at $\gamma = 2$, it also includes the LASSO when using the ℓ_1-norm. As shown in Knight and Fu (2000), bridge estimators with $\gamma \leq 1$ may also achieve sparse solutions due to the singularity of the corresponding norms at the origin.

Fan and Li (2001) argued that penalized least squares estimators should yield the properties of sparsity, continuity and unbiasedness and introduced the smoothly clipped absolute deviation (SCAD) estimator. As the latter property was weakened to the restriction, that the bias should be near zero when the true parameter is sufficiently large, the SCAD estimator satisfies these conditions. They argued that continuity in terms of the given data may be a desirable property in order to produce stable solutions. Besides providing an interpretation concerning the component's influence on the model, sparsity may increase prediction accuracy when choosing the correct underlying model. Furthermore, Fan and Li argued that a good estimator should satisfy the so-called *oracle properties*, consisting of the optimal convergence rate (and asymptotic variance) as well as consistency in model selection.

Zou (2006) proved that the LASSO may either converge with optimal rate of root-n or perform consistent model selection, but it cannot be tuned to possess both properties at a time. Moreover, even when the convergence rate is sacrificed to obtain consistent model selection, that property still depends on the structure of the underlying model and cannot be guaranteed either. In the same paper, Zou introduced the adaptive LASSO estimator, which is closely related to the nonnegative garotte of Breiman (1995), and proved its oracle properties when tuned suitably. However, it has been argued that the asymptotic distribution, when considering fixed parameters over the sample size, may be highly misleading (Leeb and Pötscher, 2005; Leeb and Pötscher, 2008; Pötscher and Leeb, 2009), especially in the case of confidence sets (Pötscher, 2009; Pötscher and Schneider, 2010). When allowing the true parameter to vary over sample size, Pötscher and Schneider (2009) showed that in the case of orthogonal regressors the adaptive LASSO estimator's uniform convergence rate is actually slower than root-n when tuned to perform consistent model selection. Besides this, Pötscher and Schneider provided an impossibility result for estimating the distribution of the adaptive LASSO estimator. Considering the case of orthogonal regressors, Pötscher and Schneider (2010) examined confidence sets based on the adaptive LASSO estimator, the hard- as well as the soft-thresholding estimator proposed by Donoho and I. M. Johnstone (1994), which coincides with the LASSO

in that specific framework.

In this thesis, we study the adaptive LASSO estimator's consistency in model selection and parameter estimation in a *moving-parameter framework* for general regressor matrices with full column rank and allow the tuning parameter to vary over its components. Our results generalize the findings of Pötscher and Schneider (2009), which were derived for orthogonal regressors and uniform tuning. While most of the required conditions carry over to the general setting in a natural way, componentwise tuning may need another restriction depending on the regressor's structure to ensure consistency in model selection. Besides the consistency properties, we examine the asymptotic distribution of the adaptive LASSO estimator in the framework of partial consistent tuning, where at least one component has to be tuned to perform consistent model selection, while the other components may be tuned arbitrarily. We then use our findings to construct asymptotic confidence sets with infimal coverage probability 1, which generalizes the results of Pötscher and Schneider (2010) concerning the adaptive LASSO estimator. The reason for the counterintuitive coverage probability of 1 is based on the existence of a compact set, which covers the asymptotic distribution regardless of the underlying true parameter. Hence, confidence sets containing that set may possess coverage probability of 1 (and actually do so if they include an open superset of the asymptotic distributions' support). The boundedness of the asymptotic distribution can easily be understood in the case of uniform tuning, where the stochastic parts vanish asymptotically. In that case, the confidence set covers all possible deterministic residuals and hence may possess asymptotic infimal coverage probability of 1.

This thesis is organized as follows: the model and its framework are described in Section 2. Conditions concerning consistency in model selection, pointwise parameter estimation and its uniform equivalent are derived in Section 3. Section 4 deals with the asymptotic behaviour of the appropriately scaled and centred adaptive LASSO estimator. These results were used in Section 5 to obtain the 0–1 confidence sets based on the adaptive LASSO estimator described above. Section 6 summarizes the results and conclusions. The appendix in Section A gives a short overview on random functions and repeats a specific result concerning the asymptotic behaviour of their minimizers used in this thesis.

2 Setting

> Everything is vague to a degree you do not realize till you have tried to make it precise.
>
> *Bertrand Russell* (1872 - 1970)

2.1 Model assumptions

During this work we postulate the following assumptions.

i. The underlying model is of the form $y = \mathbf{X}_n\beta_n + \varepsilon$.

ii. The true parameter $\beta_n \in \mathbb{R}^k$ is non-stochastic.

iii. The error term $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ consists of independent and identically distributed components $\varepsilon_i \underset{iid}{\sim} (0, \sigma^2), i = 1, \ldots, n$ with finite variance $\sigma^2 > 0$.

iv. The regressor matrix $\mathbf{X}_n = (x_1', \dots, x_n')' \in \mathbb{R}^{n \times k}$ is non-stochastic.

v. $\mathbf{X}_n$ has full rank k for every fixed n.

vi. There is a positive definite matrix $\mathbf{C} \in \mathbb{R}^{k \times k}$, such that $\lim_{n \to \infty} \frac{\mathbf{X}_n' \mathbf{X}_n}{n} = \mathbf{C}$.

Denote $\hat{\beta}_n^{\mathrm{LS}} = (\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{X}_n' y$ the least squares estimator for β_n. Assumption v guarantees the uniqueness of the least squares estimator for every fixed n. Furthermore, the assumptions ensure asymptotic normality for the centred and scaled least squares estimator and, hence, imply its consistency in parameter estimation.

2.2 Definition of the adaptive LASSO estimator

We define the weights as follows

$$\hat{w}_{n,j} = \begin{cases} 1/|\hat{\beta}_{n,j}^{\mathrm{LS}}| & \text{if } \hat{\beta}_{n,j}^{\mathrm{LS}} \neq 0 \\ 0 & \text{else.} \end{cases}$$

Furthermore, $L_n : \mathbb{R}^k \to \mathbb{R}$ denotes the objective function

$$b \mapsto (y_n - \mathbf{X}_n b)'(y_n - \mathbf{X}_n b) + 2 \sum_{j=1}^{k} \lambda_{n,j} \hat{w}_{n,j} |b_j|.$$

The components of the tuning parameter fulfil $\lambda_{n,j} \geq 0$. Throughout this work we assume the tuning parameter to be non-stochastic and, in particular, data-independent. The objective function's properties $\lim_{\|b\| \to \infty} L_n(b) = \infty$ and $L_n(b) \geq 0$ for all $b \in \mathbb{R}^k$ imply the existence of a minimizer. Furthermore, it is strictly convex even in the case where $\lambda_{n,j} \hat{w}_{n,j} = 0$ for all j due to assumption v. Defining the adaptive LASSO estimator as the minimizer of L_n, i.e., $\hat{\beta}_n^{\mathrm{A}} = \operatorname*{arg\,min}_{b \in \mathbb{R}^k} L_n(b)$, we immediately conclude its uniqueness.

2.3 Notation

$\overline{\mathbb{R}}$ denotes the real numbers extended by $\{-\infty, \infty\}$. For a given $n \in \mathbb{N}$ we define λ_n^* as the largest component of the tuning parameter, i.e., $\lambda_n^* = \max_{j=1,\dots,k} \lambda_{n,j}$. Let e_i be the i-th canonical unit vector

$$e_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

The bold $\mathbf{0}$ denotes the zero-vector, i.e., $\mathbf{0}_i = 0$ for all i. Furthermore $\operatorname{sgn} : \mathbb{R} \to \{-1, 0, 1\}$ stands for the sign-function

$$x \mapsto \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0. \end{cases}$$

Denote $\mathbb{1}_A(x) : \mathbb{R}^l \to \{0, 1\}$ the indicator function of the set $A \subseteq \mathbb{R}^l$ for some $l \in \mathbb{N}$

$$x \mapsto \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

The one-sided directional derivatives of a function f with respect to the direction $z \neq \mathbf{0}$ at x will be denoted as $\mathcal{D}_z^+ f(x)$ and $\mathcal{D}_z^- f(x)$, where

$$\mathcal{D}_z^+ f(x) = \lim_{h \searrow 0} \frac{f(x+hz) - f(x)}{h}, \mathcal{D}_z^- f(x) = \lim_{h \searrow 0} \frac{f(x-hz) - f(x)}{h}.$$

Following this notation, the one-sided partial derivatives of a function f with respect to the j-th component at x will be denoted as $\mathcal{D}_{e_j}^+ f(x)$ and $\mathcal{D}_{e_j}^- f(x)$, respectively.

3 Properties of the adaptive LASSO estimator

> Education's purpose is to replace an empty mind with an open one.
>
> *Malcom Forbes*

In this section we study the behaviour of the adaptive LASSO estimator in finite samples as well as its consistency in parameter estimation and model selection. This section's results mainly base on the following lemma, which limits the deviation of the adaptive LASSO estimator from the least squares estimator.

Lemma 1. The term $\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}$ is contained in the following set

$$\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}} \in \left\{ z \in \mathbb{R}^k : z_j \left(\frac{\mathbf{X}_n' \mathbf{X}_n}{n} z \right)_j \leq \frac{\lambda_{n,j}}{n} \text{ for all } j = 1, \ldots, k \right\}$$

Proof Consider the function $W_n : \mathbb{R}^k \to \mathbb{R}$

$$v \mapsto L_n(v + \hat{\beta}_n^{\mathrm{LS}}) - L_n(\hat{\beta}_n^{\mathrm{LS}}).$$

Denote $\mathcal{S} = \{j : \hat{\beta}_{n,j}^{\mathrm{LS}} \neq 0\} \subseteq \{1, \ldots, k\}$ the set of all non-zero components of the least squares estimator. Rearranging the terms, $W_n(v)$ can be written as

$$v'(\mathbf{X}_n' \mathbf{X}_n) v + 2(y - \mathbf{X}_n \hat{\beta}_n^{\mathrm{LS}})' \mathbf{X}_n v + 2 \sum_{j \in \mathcal{S}} \frac{|v_j + \hat{\beta}_{n,j}^{\mathrm{LS}}| - |\hat{\beta}_{n,j}^{\mathrm{LS}}|}{|\hat{\beta}_{n,j}^{\mathrm{LS}}|} \lambda_{n,j}.$$

The second term vanishes due to the normal equations of the least squares estimator. The function W_n achieves its minimum at $(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})$, where all of its one-sided partial derivatives are nonnegative. Let $j \notin \mathcal{S}$, then the desired result immediately follows from $\lambda_{n,j} \geq 0$ and

$$0 = \frac{\partial W_n}{\partial v_j}(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}) = 2 \left(\mathbf{X}_n' \mathbf{X}_n (\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}) \right)_j.$$

For the remaining part of the proof we assume $j \in \mathcal{S}$, for which we have

$$0 \leq \mathcal{D}_{e_j}^+ W_n(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}) = 2 \left(\mathbf{X}_n' \mathbf{X}_n (\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}) \right)_j + 2\lambda_{n,j} \frac{\mathbb{1}_{[0,\infty)}(\hat{\beta}_{n,j}^{\mathrm{A}}) - \mathbb{1}_{(-\infty,0)}(\hat{\beta}_{n,j}^{\mathrm{A}})}{|\hat{\beta}_{n,j}^{\mathrm{LS}}|}$$

$$0 \leq \mathcal{D}_{e_j}^- W_n(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}) = -2 \left(\mathbf{X}_n' \mathbf{X}_n (\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}) \right)_j - 2\lambda_{n,j} \frac{\mathbb{1}_{(0,\infty)}(\hat{\beta}_{n,j}^{\mathrm{A}}) - \mathbb{1}_{(-\infty,0]}(\hat{\beta}_{n,j}^{\mathrm{A}})}{|\hat{\beta}_{n,j}^{\mathrm{LS}}|}.$$

Hence, we conclude

$$\left(\mathbf{X}_n'\mathbf{X}_n(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})\right)_j = -\frac{\lambda_{n,j}\mathrm{sgn}(\hat{\beta}_{n,j}^{\mathrm{A}})}{|\hat{\beta}_{n,j}^{\mathrm{LS}}|} \tag{1}$$

for $\hat{\beta}_{n,j}^{\mathrm{A}} \neq 0$ and

$$\left|\left(\mathbf{X}_n'\mathbf{X}_n(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})\right)_j\right| \leq \frac{\lambda_{n,j}}{|\hat{\beta}_{n,j}^{\mathrm{LS}}|}.$$

for $\hat{\beta}_{n,j}^{\mathrm{A}} = 0$. Considering the case $|(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})_j| \leq |\hat{\beta}_{n,j}^{\mathrm{LS}}|$, it follows

$$\left|(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})_j \left(\mathbf{X}_n'\mathbf{X}_n(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})\right)_j\right| \leq \lambda_{n,j}.$$

Otherwise we have $\mathrm{sgn}(\hat{\beta}_{n,j}^{\mathrm{A}}) = \mathrm{sgn}(\hat{\beta}_{n,j}^{\mathrm{A}} - \hat{\beta}_{n,j}^{\mathrm{LS}}) \neq 0$. Thus, the following equation holds true

$$(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})_j \left(\mathbf{X}'\mathbf{X}(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})\right)_j = -\lambda_{n,j}\frac{|(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})_j|}{|\hat{\beta}_{n,j}^{\mathrm{LS}}|} \leq 0,$$

which completes the proof.

❑

Inspection of the foregoing proof shows that no assumption but the normal-equations have been used. Therefore the statement still remains true in the case of stochastic covariates, tuning parameters or even stochastic β_n, in the sense that $\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}^{\mathrm{LS}}$ is surely contained in the set on the right-hand side. The full rank condition of $\mathbf{X}_n$ is not needed either. However, neither $\hat{\beta}_n^{\mathrm{A}}$ nor $\hat{\beta}_n^{\mathrm{LS}}$ need not to be well defined in this case. Therefore the statement can be read as follows.

Lemma 2. Let $\tilde{\beta}_n^A$ be a minimizer of L_n and $\tilde{\beta}_n^{LS}$ fulfil the normal equations, i.e., $\mathbf{X}_n'y = \mathbf{X}_n'\mathbf{X}_n\tilde{\beta}_n^{LS}$. If the weighting vector $\hat{w}_n$ is defined by using that $\tilde{\beta}_n^{LS}$, then the following statement is true for every event of the σ-algebra of the corresponding measure space.

$$(\tilde{\beta}_n^A - \tilde{\beta}_n^{LS})_j \left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}(\tilde{\beta}_n^A - \tilde{\beta}_n^{LS})\right)_j \leq \frac{\lambda_{n,j}}{n} \text{ for all } j = 1, \ldots, k$$

However, we can be obtain another interesting conclusion if we head back to our assumptions and suppose additionally that $\lim_{n\to\infty} \lambda_n^* = 0$. Then, Lemma 1 shows that the bias is governed by the stochastic noise and therefore the adaptive LASSO estimator is asymptotically equivalent to the least squares estimator.

Lemma 3. Suppose $\lim_{n\to\infty} \lambda_n^* = 0$. Then the scaled and centred adaptive LASSO estimator converges in law to a normally distributed random vector with expectation $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{C}^{-1}$.

Proof In the decomposition

$$\sqrt{n}\left(\hat{\beta}_n^{\mathrm{A}} - \beta_n\right) = \sqrt{n}\left(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}\right) + \sqrt{n}\left(\hat{\beta}_n^{\mathrm{LS}} - \beta_n\right)$$

the bias vanishes asymptotically while the second term converges in distribution to $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{C}^{-1})$.

❑

The lemma above is a well known result in literature (Pötscher and Schneider, 2009). We would like to stress that the true parameter vector β_n may vary over the sample size and the asymptotic normality extends to all components regardless of the active set. Furthermore, the statement is valid for componentwise tuning as well as non-orthogonal covariates.

3.1 Consistency in parameter estimation

First we start with a general statement on the asymptotic behaviour of the adaptive LASSO estimator using weak assumptions.

Lemma 4. If $\left(\min(\frac{\lambda_{n,j}}{n}, |\beta_{n,j}|)\right)_{n\in\mathbb{N}}$ is bounded for all $j = 1, \dots, k$, then the sequence of the centred adaptive LASSO estimator $\left(\hat{\beta}_n^{\mathrm{A}} - \beta_n\right)_{n\in\mathbb{N}}$ is tight.

Proof Consider the function $\frac{1}{n}L_n(\cdot + \beta_n) : \mathbb{R}^k \to \mathbb{R}$. Rewriting $\frac{1}{n}L_n(u + \beta_n)$ gives

$$\frac{\varepsilon'\varepsilon}{n} + u'\frac{\mathbf{X}_n'\mathbf{X}_n}{n}u - \frac{2}{n}\varepsilon'\mathbf{X}_n u + 2\sum_{j=1}^{k}\frac{\lambda_{n,j}}{n}\hat{w}_{n,j}|u_j + \beta_{n,j}|$$

As $u'\frac{\mathbf{X}_n'\mathbf{X}_n}{n}u$ governs the third term, every sequence of random vectors $(x_n)_{n\in\mathbb{N}}$ not being tight implies that the sequence $(\frac{1}{n}L_n(x_n + \beta_n))_{n\in\mathbb{N}}$ is not bounded in probability either. If we find a sequence γ_n inducing $(\frac{1}{n}L_n(\gamma_n+\beta_n))_{n\in\mathbb{N}}$ being tight, then $\min_{u\in\mathbb{R}^k} L_n(u+\beta_n)$ and, subsequently, $\hat{\beta}^{\mathrm{A}}-\beta_n = \arg\min_{u\in\mathbb{R}^k} L_n(u+\beta_n)$ are bounded in probability as well. Now we define γ_n as follows:

$$\gamma_{n,j} = \begin{cases} -\beta_{n,j} & \text{if } |\beta_{n,j}| \leq \max(1, \frac{\lambda_{n,j}}{n}) \\ 0 & \text{else.} \end{cases}$$

Denoting $\{n \in \mathbb{N} : \gamma_{n,j} + \beta_{n,j} \neq 0\}$ with $\mathcal{L}_j$, we conclude the boundedness of $(\frac{\lambda_{n,j}}{n})_{n\in\mathcal{L}_j}$ as well as the tightness of $(\beta_{n,j}\hat{w}_{n,j})_{n\in\mathcal{L}_j}$. Thus, the sequence $\left(|\gamma_{n,j} + \beta_{n,j}|\hat{w}_{n,j}\frac{\lambda_{n,j}}{n}\right)_{n\in\mathbb{N}}$ is tight. Together with the fact, that the boundedness of $\min(\frac{\lambda_{n,j}}{n}, |\beta_{n,j}|)$ carries over to $\gamma_{n,j}$, we infer the tightness of $(\frac{1}{n}L_n(\gamma_n + \beta_n))_{n\in\mathbb{N}}$.

❑

The following proposition is a generalization of Theorem 2 in Pötscher and Schneider (2009), as the regressor matrix need not be orthogonal.

Proposition 5. *Let $a_n = \min(\sqrt{n}, \sqrt{\frac{n}{\lambda_n^*}})$. Then, for every $\epsilon > 0$, there exists a real number M such that*

$$\sup_{n\geq k}\sup_{\beta\in\mathbb{R}^k} \mathbb{P}_{n,\beta}\left(a_n\|\hat{\beta}_n^A - \beta\| > M\right) < \epsilon \tag{2}$$

holds. If $\lim_{n\to\infty}\frac{\lambda_n^*}{n} = 0$, then the adaptive LASSO estimator is uniformly a_n-consistent in parameter estimation.

Proof Denote $\alpha_{n,j}$ the j-th eigenvalue of $\frac{\mathbf{X}_n'\mathbf{X}_n}{n}$ and $\alpha_{0,j}$ the j-th eigenvalue of $\mathbf{C}$. As all these matrices are positive definite and $\lim_{n\to\infty}\alpha_{n,j} = \alpha_{0,j} > 0$ for all $j = 1, \dots, k$, the infimum L of the set $\{\alpha_{n,j} : n \geq k, j = 1, \dots, k\}$ is strictly positive. Let $M \geq \sqrt{\frac{4k}{L}}$, then by Lemma 1 and the fact, that $\frac{a_n^2\lambda_{n,j}}{n} \leq 1$,

$$a_n^2\|\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}\|_2^2 \leq \frac{a_n^2}{\min_{j=1\dots k}\alpha_{n,j}}(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})'\frac{\mathbf{X}_n'\mathbf{X}_n}{n}(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}) \leq \frac{a_n^2}{\min_{j=1\dots k}\alpha_{n,j}}\sum_{j=1}^{k}\frac{\lambda_{n,j}}{n} \leq \frac{M^2}{4}$$

holds true surely. Thus, it follows that

$$\begin{aligned}
&\inf_{\beta\in\mathbb{R}^k} \mathbb{P}_{n,\beta}\left(a_n\|\hat{\beta}_n^{\mathrm{A}}-\beta\|\leq M\right) \geq \inf_{\beta\in\mathbb{R}^k} \mathbb{P}_{n,\beta}\left(a_n\|\hat{\beta}_n^{\mathrm{A}}-\hat{\beta}^{\mathrm{LS}}\|\leq \frac{M}{2}, a_n\|\hat{\beta}_n^{\mathrm{LS}}-\beta\|\leq \frac{M}{2}\right)\\
&= \inf_{\beta\in\mathbb{R}^k} \mathbb{P}_{n,\beta}\left(a_n\|\hat{\beta}_n^{\mathrm{A}}-\hat{\beta}^{\mathrm{LS}}\|\leq \frac{M}{2} \;\middle|\; a_n\|\hat{\beta}_n^{\mathrm{LS}}-\beta\|\leq \frac{M}{2}\right)\mathbb{P}_{n,\beta}\left(a_n\|\hat{\beta}_n^{\mathrm{LS}}-\beta\|\leq \frac{M}{2}\right)\\
&= \inf_{\beta\in\mathbb{R}^k} \mathbb{P}_{n,\beta}\left(a_n\|\hat{\beta}_n^{\mathrm{LS}}-\beta\|\leq \frac{M}{2}\right).
\end{aligned}$$

Now, Equation (2) directly follows from the uniform root-n-consistency of the least squares estimator.

❑

Proposition 5 allows us to create confidence sets based on the adaptive LASSO estimator. However, they are conservative in the sense that their *actual* coverage probability is not smaller than their nominal coverage probability. Another reason against the use of these confidence sets is the fact, that – for a given coverage probability – they are proper supersets of confidence sets based on the least squares estimator. Nevertheless, the Proposition above precisely describes the condition for performing consistent parameter estimation. The property $\lim_{n\to\infty}\frac{\lambda_n^*}{n}=0$ guarantees consistency in parameter estimation. However, that condition is not only a sufficient, but also a necessary one, as the following result shows.

Theorem 6. *The adaptive LASSO estimator $\hat{\beta}^A$ is consistent in parameter estimation if and only if $\lim_{n\to\infty}\frac{\lambda_{n,j}}{n}\to 0$ for all $j=1,\dots,k$.*

Proof The first implication follows directly from Proposition 5. For the second direction, assume that there exists a j^* such that $\limsup_{n\to\infty}\frac{\lambda_{n,j^*}}{n}=c\in(0,\infty]$. With the linear functions $f_n:\mathbb{R}^k\to\mathbb{R}$, $x\mapsto\left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}x\right)_{j^*}$ denote $\bar{L}=\sup_{n\in\mathbb{N}}\|f_n\|$ the supremum of the operator norms of f_n. As $\lim_{n\to\infty}\frac{\mathbf{X}_n'\mathbf{X}_n}{n}\to\mathbf{C}$ and $\mathbf{C}$ has full rank, we conclude $0<\bar{L}<\infty$. For a given $\epsilon>0$, let $\beta_{n,j^*}=2\epsilon$ and consider the event where $|\hat{\beta}_{n,j^*}^{\mathrm{A}}-\beta_{n,j^*}|<\epsilon$ and $\hat{\beta}_{n,j^*}^{\mathrm{LS}}\neq 0$. On this event we have $\hat{\beta}_{n,j^*}^{\mathrm{A}}\neq 0$ and from Equation (1)

$$\frac{\lambda_{n,j^*}}{n|\hat{\beta}_{n,j^*}^{\mathrm{LS}}|}=\left|\left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}\left(\hat{\beta}_n^{\mathrm{A}}-\hat{\beta}_n^{\mathrm{LS}}\right)\right)_{j^*}\right|\leq\bar{L}\|\hat{\beta}_n^{\mathrm{A}}-\hat{\beta}_n^{\mathrm{LS}}\|\leq\bar{L}\|\hat{\beta}_n^{\mathrm{LS}}-\beta_n\|+\bar{L}\epsilon.$$

Altogether we have

$$\begin{aligned}
&\liminf_{n\to\infty}\mathbb{P}\left(|\hat{\beta}_{n,j^*}^{\mathrm{A}}-\beta_{n,j^*}|<\epsilon\right)\leq\\
&\liminf_{n\to\infty}\mathbb{P}\left(\hat{\beta}_{n,j^*}^{\mathrm{LS}}=0\right)+\liminf_{n\to\infty}\mathbb{P}\left(\frac{\lambda_{n,j^*}}{n}\frac{1}{\bar{L}|\hat{\beta}_{n,j^*}^{\mathrm{LS}}|}-\epsilon\leq\|\hat{\beta}_n^{\mathrm{LS}}-\beta_n\|\right) \qquad (3)
\end{aligned}$$

Due to the consistency of the least squares estimator and the fact, that $\beta_{n,j^*}=2\epsilon$, the first term in (3) vanishes. On the other hand, $\limsup_{n\to\infty}\frac{\lambda_{n,j^*}}{n}=c>0$, implies the existence of a subsequence n_l, such that

$$\frac{\lambda_{n_l,j^*}}{n_l}\frac{1}{\bar{L}|\hat{\beta}_{n_l,j^*}^{\mathrm{LS}}|}-\epsilon\xrightarrow[l\to\infty]{p}\frac{c}{2\bar{L}\epsilon}-\epsilon,$$

which is strictly positive for sufficiently small ϵ. Note that the limit on the right-hand side is located in $(0, \infty]$, as c can take the value ∞. On the other hand $\|\hat{\beta}^{\text{LS}}_{n_l} - \beta_{n_l}\| \xrightarrow[l\to\infty]{p} 0$. Therefore the last term in (3) equals 0 as well.

❑

Remark. Inspection of the proof of Theorem 6 gives us another result. In fact, the stricter condition of a component j^* fulfilling $\liminf_{n\to\infty} \frac{\lambda_{n,j^*}}{n} = c \in (0, \infty]$ implies the existence of β_{n,j^*} where $\limsup_{n\to\infty} \mathbb{P}\left(|\hat{\beta}^{\text{A}}_{n,j^*} - \beta_{n,j^*}| < \epsilon\right) = 0$ for every $\epsilon > 0$. The main idea is, again, based on Equation (1), which shows that, given $\hat{\beta}^{\text{A}}_{j^*} \neq 0$ and $\hat{\beta}^{\text{LS}}_{j^*} \neq 0$, $\hat{\beta}^{\text{A}}_n - \hat{\beta}^{\text{LS}}_n$ cannot converge to $\mathbf{0}$ in probability.

Remark. The equivalence of $\lim_{n\to\infty} \frac{\lambda_n}{n} = 0$ and the consistency in parameter estimation was already pointed out in Pötscher and Schneider (2009) for a special case. Theorem 6 confirm that the condition on the tuning parameter carries over in a natural way in the framework of componentwise tuning and non-orthogonal regressors.

Assuming $\lim_{n\to\infty} \lambda^*_n = \infty$, Proposition 5 shows that the convergence rate of $\sup_{\beta\in\mathbb{R}^k}(\hat{\beta}^{\text{A}}_n - \beta)$ towards 0 is at least of the order ${\lambda^*_n}^{\frac{1}{2}} n^{-\frac{1}{2}}$. The findings of Pötscher and Schneider (2009) imply that the convergence rate is indeed of that order in the case of orthogonal covariates and uniform tuning. Our results of Chapter 4 confirms that for the general case. Interestingly, the convergence rate is restricted by the cases, where the true parameter β_n converges to 0 while not being 0. The property, that the convergence rate of the adaptive LASSO estimator is restricted by parameters being in the neighbourhood of 0, was already pointed out in Pötscher and Schneider (2009). However, considering the fixed-parameter framework, i.e., $\beta_n = \beta$ for all n, the adaptive LASSO estimator may attain a faster convergence rate, which is a matter of interest when considering consistency in model selection.

Lemma 7. Denote $b_n = \min(\sqrt{n}, \frac{n}{\lambda^*_n})$. If $\beta_n = \beta$ for all $n \in \mathbb{N}$ and $\frac{\lambda^*_n}{n} \to 0$, then the adaptive LASSO estimator is b_n-consistent.

Proof We define the function $H_n(u) = \frac{b_n^2}{n}\left(L_n(\frac{u}{b_n} + \beta) - L_n(\beta)\right)$. Rewriting $H_n(u)$ gives

$$H_n(u) = u' \frac{\mathbf{X}'_n \mathbf{X}_n}{n} u - 2u' \mathbf{X}'_n \varepsilon \frac{b_n}{n} + 2 \sum_{j=1}^{k} \lambda_{n,j} \hat{w}_{n,j} \frac{b_n^2}{n} \left(|\frac{u_j}{b_n} + \beta_j| - |\beta_j| \right)$$

The minimum of H_n cannot be positive and is achieved at the point $b_n(\hat{\beta}^{\text{A}}_n - \beta)$. With the notation $\mathcal{A} = \{j : \beta_j \neq 0\}$, we conclude

$$b_n(\hat{\beta}^{\text{A}}_n - \beta)' \frac{\mathbf{X}'_n \mathbf{X}_n}{n} (\hat{\beta}^{\text{A}}_n - \beta) b_n \leq 2 \frac{\varepsilon' \mathbf{X}_n}{\sqrt{n}} (\hat{\beta}^{\text{A}}_n - \beta) b_n \frac{b_n}{\sqrt{n}} + 2 \sum_{j\in\mathcal{A}} \hat{w}_{n,j} |b_n(\hat{\beta}^{\text{A}}_n - \beta)_j| b_n \frac{\lambda_{n,j}}{n},$$

where we used $|\beta_j| - |\frac{u_j}{b_n} + \beta_j| \leq |\frac{u_j}{b_n}|$ for the latter term. Both $\frac{b_n}{\sqrt{n}}$ and $b_n \frac{\lambda_{n,j}}{n}$ are upper bounded by 1, while the sequences $\frac{\mathbf{X}'_n \varepsilon}{\sqrt{n}}$ and $(\hat{w}_{n,j})_{j\in\mathcal{A}}$ are tight. As for all n in $\mathbb{N}$ the matrix $\frac{\mathbf{X}'_n \mathbf{X}_n}{n}$ and its limit $\mathbf{C}$ are positive definite, $b_n(\hat{\beta}^{\text{A}}_n - \beta)' \frac{\mathbf{X}'_n \mathbf{X}_n}{n} (\hat{\beta}^{\text{A}}_n - \beta) b_n$ can be lower bounded by a scalar multiplied with $\|b_n(\hat{\beta}^{\text{A}}_n - \beta)\|^2$. Hence, $b_n(\hat{\beta}^{\text{A}}_n - \beta)$ is bounded in probability.

❑

Remark. In the case, where the adaptive LASSO estimator is uniformly tuned with a rate faster than root-n, i.e., $\lim_{n\to\infty} \frac{\lambda_n}{\sqrt{n}} \to \infty$, Lemma 7 can be directly concluded by Theorem 2 in Zou (2006). Nevertheless, the consistency of the adaptive LASSO estimator is independent of the limit of $\frac{\lambda_{n,j}}{\sqrt{n}}$ and may be attained by componentwise tuning as well.

In order to see that the actual convergence rate of $\hat{\beta}_n^{\mathrm{A}} - \beta$ towards 0 cannot be faster than b_n, we may again refer to Pötscher and Schneider (2009). Nevertheless, we would like to outline a proof for the general case as well. To achieve this, we consider the case $\beta_j < 0$ for all j. Then H_n converges in distribution to

$$u'\mathbf{C}u - 2Z_1'uc_1 - 2\sum_{j=1}^{k} \lambda_{0,j} c_2 \frac{u_j}{|\beta_j|},$$

with Z_1 being a normally distributed random variable with mean $\mathbf{0}$ and covariance matrix $\mathbf{C}$, $c_1 = \lim_{n\to\infty} \frac{b_n}{\sqrt{n}}$, $c_2 = \lim_{n\to\infty} b_n \frac{\lambda_n^*}{n}$ and $\lambda_{0,j} = \lim_{n\to\infty} \frac{\lambda_{n,j}}{\lambda_n^*}$. If one or more of these limits do not exist, we consider a subsequence n_l, such that all limits exist. Due to the compactness of $[0,1]^{k+2}$ there is always such a subsequence. Using the convexity of H_{n_l} and its limit, it is possible to show that the minimizer of H_{n_l} converges in distribution to the minimizer of its limiting function. However, the latter one is not a Dirac distributed random variable with its mass at $\mathbf{0}$.

3.2 Consistency in model selection

In Zou (2006) the consistency in model selection of the adaptive LASSO estimator was proved for uniform tuning under the conditions $\lim_{n\to\infty} \lambda_n = \infty$ and $\lim_{n\to\infty} \frac{\lambda_n}{\sqrt{n}} = 0$. With a slight modification of their proof, the same holds true in the case of component wise tuning if we rewrite the assumptions as $\lim_{n\to\infty} \lambda_{n,j} = \infty$ and $\lim_{n\to\infty} \frac{\lambda_{n,j}}{\sqrt{n}} = 0$ for all j. Interestingly, consistency in model selection may also be achieved under different circumstances. In fact, this property is independent of the limit of $\frac{\lambda_n^*}{\sqrt{n}}$. More precisely, consistency in model selection requires the conditions $\lim_{n\to\infty} \frac{\lambda_n^*}{n} = 0$ and $\lim_{n\to\infty} \lambda_{n,j} = \infty$ for all j in any case. In the uniform tuning framework these assumptions suffice. However, in the componentwise tuning framework an additional condition on the ratio of the tuning parameter's components may be necessary depending on the structure of the covariates.

Definition 8. Concerning consistency in model selection, it seems reasonable to consider the parameters to be fixed over the sample size n, i.e., $\beta_n = \beta$ for all $n \in \mathbb{N}$. Thus, we can define the active set $\mathcal{A} = \{j : \beta_j \neq 0\}$.

Proposition 9. *Suppose the adaptive LASSO estimator is tuned to perform consistent parameter estimation, i.e.,* $\lim_{n\to\infty} \frac{\lambda_n^*}{n} = 0$. *If* $\lim_{n\to\infty} \lambda_{n,j} = \infty$ *and* $\lim_{n\to\infty} \frac{\lambda_{n,j}\sqrt{n}}{\lambda_n^*} = \infty$ *for all* $j = 1, \ldots, k$, *then the adaptive LASSO estimator is consistent in model selection.*

Proof The condition $\lim_{n\to\infty} \frac{\lambda_n^*}{n} = 0$ ensures the consistency in parameter estimation, which gives us conservative model selection in the sense that the probability of underestimating a model vanishes asymptotically. Thus, we only need to prove

$$\lim_{n\to\infty} \mathbb{P}(\exists j \notin \mathcal{A} : \hat{\beta}_j^{\mathrm{A}} \neq 0) = 0$$

for all $\beta \in \mathbb{R}^k$. Consider the function $G_n(u) = \frac{1}{n}L_n(u+\beta) - \frac{1}{n}L_n(\beta)$. Rewriting $G_n(u)$ gives

$$u'\frac{\mathbf{X}_n'\mathbf{X}_n}{n}u - 2\varepsilon'\mathbf{X}_n\frac{u}{n} + 2\sum_{j=1}^{k}\frac{\lambda_{n,j}}{n}\hat{w}_{n,j}(|u_j+\beta_j| - |\beta_j|).$$

Since $\hat{\beta}^{\mathrm{A}} - \beta$ minimizes G_n, the existence of $\tilde{j} \notin \mathcal{A}$, such that $\hat{\beta}^{\mathrm{A}}_{\tilde{j}} \neq 0$, implies

$$\left|\left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}(\hat{\beta}^{\mathrm{A}} - \beta) - \frac{\mathbf{X}_n'\varepsilon}{n}\right)_{\tilde{j}}\right| = \frac{\lambda_{n,\tilde{j}}}{\sqrt{n}}\frac{\hat{w}_{n,\tilde{j}}}{\sqrt{n}}.$$

Multiplying each side with $b_n = \min(\sqrt{n}, \frac{n}{\lambda_n^*})$ yields

$$\left|\left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}b_n(\hat{\beta}^{\mathrm{A}} - \beta) - \frac{\mathbf{X}_n'\varepsilon}{\sqrt{n}}\frac{b_n}{\sqrt{n}}\right)_{\tilde{j}}\right| = b_n\frac{\lambda_{n,\tilde{j}}}{\sqrt{n}}\frac{\hat{w}_{n,\tilde{j}}}{\sqrt{n}}.$$

The left-hand side is always bounded in probability due to the b_n-consistency of the adaptive LASSO estimator. If $\lim_{n\to\infty}\frac{b_n}{\sqrt{n}} = c > 0$, then $\lim_{n\to\infty}\lambda_{n,j} = \infty$ suffices for the divergence of the right-hand side to ∞. Otherwise, the second term consists of $\frac{\lambda_{n,\tilde{j}}\sqrt{n}}{\lambda_n^*}\frac{\hat{w}_{n,\tilde{j}}}{\sqrt{n}}$, which also explodes asymptotically. In both cases the probability of $\hat{\beta}^{\mathrm{A}}_{n,\tilde{j}} \neq 0$, for a $\tilde{j} \notin \mathcal{A}$ converges to 0.

❑

The condition $\lim_{n\to\infty}\frac{\lambda_n^*}{n} = 0$ is a crucial one in the sense, that $\limsup_{n\to\infty}\frac{\lambda_n^*}{n} = c > 0$ leads to inconsistent model selection as it allows underestimation of a model with a positive asymptotic probability. To see this, define j^*, such that $\limsup_{n\to\infty}\frac{\lambda_{n,j^*}}{n} = c$. (There is always such a component, as otherwise $\lim_{n\to\infty}\frac{\lambda_n^*}{n}$ would be 0.) Now we consider the case $\beta = e_{j^*}\sqrt{\frac{c}{2\mathbf{C}_{j^*,j^*}}}$, where $\mathbf{C}_{j^*,j^*}$ denotes the j^*-th diagonal element of $\mathbf{C}$. Due to the inequality

$$\begin{aligned}\liminf_{n\to\infty}\mathbb{P}(\mathcal{A}_n = \mathcal{A}) &= \liminf_{n\to\infty}\mathbb{P}(\hat{\beta}^{\mathrm{A}}_{n,i} = 0 \text{ for all } i \neq j^* \text{ and } \hat{\beta}^{\mathrm{A}}_{n,j^*} \neq 0)\\ &\leq \liminf_{n\to\infty}\mathbb{P}(\hat{\beta}^{\mathrm{A}}_{n,j^*} \neq 0 | \hat{\beta}^{\mathrm{A}}_{n,i} = 0 \text{ for all } i \neq j^*)\end{aligned}$$

we assume $\hat{\beta}^{\mathrm{A}}_{n,i} = 0$ for all $i \neq j^*$. Considering the function G_n, $\hat{\beta}^{\mathrm{A}}_{n,j^*} = 0$ requires

$$\left|\left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}(\hat{\beta}^{\mathrm{A}} - \beta) - \frac{\mathbf{X}_n'\varepsilon}{n}\right)_{j^*}\right| \leq \frac{\lambda_{n,j^*}}{n}\hat{w}_{n,j^*}.$$

The first term simplifies to $\left|\left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}\right)_{j^*,j^*}\sqrt{\frac{c}{2\mathbf{C}_{j^*,j^*}}} + \left(\frac{\mathbf{X}_n'\varepsilon}{n}\right)_{j^*}\right|$ because of the assumption $\beta_{n,i} = \hat{\beta}^{\mathrm{A}}_{n,i}$ for all $i \neq j^*$. Hence, it converges in probability to $\sqrt{\mathbf{C}_{j^*,j^*}\frac{c}{2}}$. On the other hand, there is a subsequence n_l, such that $\frac{\lambda_{n_l,j^*}}{n_l}\hat{w}_{n_l,j^*} \xrightarrow[l\to\infty]{p} \sqrt{2c\mathbf{C}_{j^*,j^*}}$, which in turn implies that $\limsup_{n\to\infty}\mathbb{P}(\hat{\beta}^{\mathrm{A}}_{n,j^*} = 0 | \hat{\beta}^{\mathrm{A}}_{n,i} = 0 \text{ for all } i \neq j^*) = 1$.

The condition $\lim_{n\to\infty}\lambda_{n,j} = \infty$ for all j contributes a large part in avoiding overestimation of models. To show the necessity of that condition, we define $\tilde{j}$ such that $\liminf_{n\to\infty}\lambda_{n,\tilde{j}} = c < \infty$ and consider the case $\beta_j = 0$ for all j. In order to guarantee consistency in model selection, we need $\hat{\beta}^{\mathrm{A}} = \mathbf{0}$ with asymptotic probability 1. A necessary condition for G being minimized at the point 0 is given by

$$\left|\frac{(\mathbf{X}_n'\varepsilon)_{\tilde{j}}}{\sqrt{n}}\right| \leq \frac{\hat{w}_{n,\tilde{j}}}{\sqrt{n}}\lambda_{n,\tilde{j}}.$$

However, there exists a subsequence n_l, such that this inequality is fulfilled with a probability less than 1, which is a contradiction to consistency in model selection.

The condition $\lim_{n\to\infty} \frac{\lambda_{n,j}\sqrt{n}}{\lambda_n^*} = \infty$ guarantees, that the convergence rates of the tuning parameter does not diverge too much. Without this condition, the penalty term of a component j with $\beta_j = 0$ needs not to grow faster than the other terms. As a result, $\hat{\beta}_j^{\mathrm{A}}$ is not necessarily set to 0, especially if $\mathbf{C}_{i,j} > 0$ for another component i fulfilling $\beta_i \neq 0$ and $\lim_{n\to\infty} \frac{\lambda_{n,i}}{\lambda_n^*} \neq 0$. In this case, $\hat{\beta}_j^{\mathrm{A}}$ may be penalized less than $\hat{\beta}_i^{\mathrm{A}}$, possibly causing $\hat{\beta}_j^{\mathrm{A}}$ compensating the shrinkage of $\hat{\beta}_i^{\mathrm{A}}$. However, the necessity of this condition depends on the underlying model. If all covariates are orthogonal, the condition is not needed as the minimization of the target function reduces to the one-dimensional minimization of its components. In this case, the components cannot compensate a possible over-reduction of another non-zero component.

To show the necessity of that condition in certain cases, we assume that there is a $\tilde{j}$ and j^*, such that $\mathbf{C}_{j^*,\tilde{j}} \neq 0$, $\limsup_{n\to\infty} \frac{\lambda_{n,\tilde{j}}\sqrt{n}}{\lambda_n^*} = c < \infty$ and $\lim_{n\to\infty} \frac{\lambda_{n,j^*}}{\lambda_n^*} > 0$. Considering the model $\beta = e_{j^*}$, consistency in model selection requires $\hat{\beta}_{n,i}^{\mathrm{A}} = 0$ for all $i \neq j^*$ and $\hat{\beta}_{n,j^*}^{\mathrm{A}} \neq 0$ with asymptotic probability 1. Rewriting the function $H_n(u) = \frac{b_n^2}{n}\left(L_n(\frac{u}{b_n} + \beta) - L_n(\beta)\right)$ gives

$$H_n(u) = u'\frac{\mathbf{X}_n'\mathbf{X}_n}{n}u - 2u'\mathbf{X}_n'\varepsilon\frac{b_n}{n} + 2\sum_{j=1}^{k}\lambda_{n,j}\hat{w}_{n,j}\frac{b_n^2}{n}\left(|\frac{u_j}{b_n} + \beta_j| - |\beta_j|\right).$$

Again, we stress that $m_n = \arg\min_{u\in\mathbb{R}^k} H_n(u) = b_n(\hat{\beta}_n^{\mathrm{A}} - \beta)$. On every event, on which the adaptive LASSO estimator performs correct model selection, we have $m_{n,i} = 0$ for all $i \neq j^*$. Hence, necessary optimality conditions on those events are

$$\left|\left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}\right)_{j^*,j^*} m_{n,j^*} - \left(\mathbf{X}_n'\varepsilon\right)_{j^*}\frac{b_n}{n}\right| = \lambda_{n,j^*}\hat{w}_{n,j^*}\frac{b_n}{n}$$

as well as

$$\left|\left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}\right)_{j^*,\tilde{j}} m_{n,j^*} - \left(\mathbf{X}_n'\varepsilon\right)_{\tilde{j}}\frac{b_n}{n}\right| \leq \lambda_{n,\tilde{j}}\hat{w}_{n,\tilde{j}}\frac{b_n}{n}.$$

Combining these formulas and applying the triangle inequality results in

$$\left|\left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}\right)_{j^*,\tilde{j}}\right|\left(\lambda_{n,j^*}\hat{w}_{n,j^*}\frac{b_n}{n} - \left|\left(\mathbf{X}_n'\varepsilon\right)_{j^*}\right|\frac{b_n}{n}\right) \leq \left(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}\right)_{j^*,j^*}\left(\lambda_{n,\tilde{j}}\hat{w}_{n,\tilde{j}}\frac{b_n}{n} + \left|\left(\mathbf{X}_n'\varepsilon\right)_{\tilde{j}}\right|\frac{b_n}{n}\right).$$

The left-hand side converges in probability to $|\mathbf{C}_{j^*,\tilde{j}}|\lambda_{0,j^*} > 0$, where λ_{0,j^*} denotes the limit of $\frac{\lambda_{n,j^*}}{\lambda_n^*}$. There is a subsequence n_l, such that this subsequence of the term on the right-hand side converges in distribution to $\frac{\mathbf{C}_{j^*,j^*}c}{|Z_{\tilde{j}}|}$, where $Z_{\tilde{j}}$ is a normally distributed random variable with mean 0 and variance $\sigma^2(\mathbf{C}^{-1})_{\tilde{j},\tilde{j}}$. However, that inequality cannot be fulfilled with probability 1 and therefore the event $\hat{\beta}_{n,i}^{\mathrm{A}} = 0$ for all $i \neq j^*$ and $\hat{\beta}_{n,j^*}^{\mathrm{A}} \neq 0$ does not posses asymptotic probability 1.

Remark. In the uniform tuning framework the condition $\lim_{n\to\infty} \frac{\lambda_{n,j}\sqrt{n}}{\lambda_n^*} = \infty$ is always fulfilled, regardless of the limit of $\frac{\lambda_{n,j}}{\sqrt{n}}$. If λ_n^* is not growing faster than with rate root-n, i.e., $\lim_{n\to\infty} \frac{\lambda_n^*}{\sqrt{n}} < \infty$, then that condition is fulfilled even in the componentwise tuning framework.

Remark. The proof of Proposition 9 gives another, more precise result. Suppose the adaptive LASSO estimator is tuned to perform consistent parameter estimation. Then every component j additionally fulfilling $\lim_{n\to\infty} \lambda_{n,j} = \infty$ and $\lim_{n\to\infty} \frac{\sqrt{n}\lambda_{n,j}}{\lambda_n^*} = \infty$ is consistently estimated in terms of model selection, i.e., $|\operatorname{sgn}(\hat{\beta}_j^{\mathrm{A}})| \xrightarrow[n\to\infty]{p} |\operatorname{sgn}(\beta_j)|$. This property is a matter of interest especially for the partial adaptive LASSO estimator, in which some of its components are not penalized at all. Those components are usually assumed to be relevant for the model and the model selection is reduced to the set of all models, in which these components are located in the active set. However, that result is also important for all models of the adaptive LASSO estimator containing an intercept. In practice, that component is often not penalized and therefore its influence on the consistency in model selection may be a matter of interest.

3.3 Bias of the adaptive LASSO estimator

Lemma 1 and the unbiasedness of the least squares estimator imply that the bias of the adaptive LASSO estimator is contained in the closure of the convex hull of the set in Lemma 1, i.e.,

$$\begin{aligned}\mathbb{E}(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}}) &\in cl\left(co\left(\left\{z \in \mathbb{R}^k : z_j \left(\frac{\mathbf{X'X}}{n} z\right)_j \leq \frac{\lambda_{n,j}}{n} \text{ for all } j = 1, \ldots, k\right\}\right)\right) \\ &\subseteq \left\{z \in \mathbb{R}^k : z' \left(\frac{\mathbf{X'X}}{n}\right) z \leq \sum_{j=1}^{k} \frac{\lambda_{n,j}}{n}\right\}\end{aligned}$$

Since $z'Az = \sum_{j=1}^k z_j (Az)_j$ the second set covers the set in Lemma 1. As it is closed and convex too, it is a superset of the first set.
Suppose $\lim_{n\to\infty} \frac{\lambda_{n,j}}{n} \to 0$ for all $j = 1, \ldots, k$. Then, for every fixed parameter, i.e., $\beta_n = \beta$ for all n, the bias of the adaptive LASSO estimator cannot vanish slower than with rate $\max(\frac{\lambda_n^*}{n}, n^{-\frac{1}{2}})$ according to Lemma 7. Hence, a slower rate, as suggested in Proposition 5 and the formula above, may only be obtained in the moving parameter framework. In fact, $\sup_{\beta\in\mathbb{R}^k} \mathbb{E}(\hat{\beta}_n^{\mathrm{A}} - \beta)$ is indeed of order $\sqrt{\frac{\lambda_n^*}{n}}$, as we will see in Section 4. In order to prove this, we will need the following auxiliary result.

Lemma 10. For every $j = 1, \ldots, k$ the sequence $\left(\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_{n,j}^{\mathrm{A}} - \hat{\beta}_{n,j}^{\mathrm{LS}})\right)_{n\in\mathbb{N}}$ is uniformly integrable independent of the underlying sequence $(\beta_{n,j})_{n\in\mathbb{N}}$.

Proof Let L be the infimum of all eigenvalues of the matrices $(\frac{\mathbf{X}_n'\mathbf{X}_n}{n})_{n\in\mathbb{N}}$. As all of these matrices as well as their limit $\mathbf{C}$ are positive definite we conclude $L > 0$. Together with Lemma 1 we have

$$\sum_{j=1}^{k} \frac{\lambda_{n,j}}{\lambda_n^*} \geq \left(\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})\right)' \frac{\mathbf{X}_n'\mathbf{X}_n}{n} \left(\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})\right) \geq \left\|\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}} - \hat{\beta}_n^{\mathrm{LS}})\right\|_2^2 L.$$

Hence, for every n in $\mathbb{N}$ the term $\left(\sqrt{\frac{n}{\lambda_n^*}}|\hat{\beta}_{n,j}^{\mathrm{A}} - \hat{\beta}_{n,j}^{\mathrm{LS}}|\right)$ is less or equal $\sqrt{\frac{k}{L}}$ which guarantees the uniform integrability.

❑

4 Asymptotic behaviour of the adaptive LASSO estimator

> Mathematics is not a deductive science – that's a cliché. When you try to prove a theorem, you don't just list the hypotheses, and then start to reason. What you do is trial and error, experimentation, guesswork.
>
> *Paul Halmos*

In the case where $\lim_{n\to\infty} \lambda_n^* = 0$ is fulfilled, Lemma 3 describes the asymptotic behaviour of the centred and scaled adaptive LASSO estimator, which can be used to set up confidence sets. Alternatively, the assumption $\lim_{n\to\infty} \lambda_n^* = \infty$ allows us to construct confidence sets with asymptotic coverage probability 1 according to Lemma 1. The reason for the counter-intuitive (asymptotic) coverage probability of 1 is the partitioning of the adaptive LASSO estimator's deviation into a bounded term of order $\lambda_n^{*\frac{1}{2}} n^{-\frac{1}{2}}$ and an unbiased deviation of order $n^{-\frac{1}{2}}$. In the case where $\lim_{n\to\infty} \lambda_n^* = \infty$ the bias term $\hat{\beta}^{\mathrm{A}} - \hat{\beta}^{\mathrm{LS}}$ vanishes asymptotically slower than the $\hat{\beta}^{\mathrm{LS}} - \beta$. Thus, confidence sets containing the bounded set of Lemma 1 may have asymptotic coverage probability 1. Nevertheless, the constructed confidence set is not necessarily the smallest set for this asymptotic coverage probability. Hence, during this and the following section we study the asymptotic behaviour of the adaptive LASSO estimator under the condition $\lim_{n\to\infty} \lambda_n^* = \infty$ in order to get the smallest confidence set. To guarantee consistency in parameter estimation, we additionally assume $\lim_{n\to\infty} \frac{\lambda_n^*}{n} = 0$ according to Theorem 6. Interestingly, the assumptions stated do not contain only the case of consistent model selection, but also some kind of "mixed" tuning. This includes the partial adaptive LASSO estimator if at least one component is tuned to perform consistent model selection.
To summarise, we extend the existing assumptions throughout this and the subsequent section by the following conditions:

$$\forall j \ \exists \lambda_{0,j} \in [0,1] : \lambda_{0,j} = \lim_{n\to\infty} \frac{\lambda_{n,j}}{\lambda_n^*}$$

$$\lim_{n\to\infty} \lambda_n^* = \infty \qquad \lim_{n\to\infty} \frac{\lambda_n^*}{n} = 0.$$

Furthermore, we assume in this section the existence of $\varphi \in \overline{\mathbb{R}}^k$ and $\psi \in [0,\infty]^k$, such that $\lim_{n\to\infty} \sqrt{n}\beta_{n,j} \frac{\sqrt{\lambda_n^*}}{\lambda_{n,j}} = \varphi_j$ as well as $\lim_{n\to\infty} \frac{\sqrt{\lambda_n^*}}{\lambda_{n,j}} = \psi_j$ for all j. In this section we postulate the following notation. Denote by $(\Omega, \mathfrak{S}, \mu)$ a probability space, while Z describes a k-dimensional, normally distributed random vector on Ω with $Z \sim N(0, \sigma^2 \mathbf{C}^{-1})$.

4.1 Asymptotic distribution

In this subsection we derive the asymptotic distribution of the properly scaled and centred adaptive LASSO estimator. For this the following definitions turned out to be fruitful concepts.

Definition 11. Let $V_n : \mathbb{R}^k \to \mathbb{R}$, given by $V_n(u) = \frac{1}{\lambda_n^*} \left(L_n \left(\sqrt{\frac{\lambda_n^*}{n}} u + \beta_n \right) - L_n(\beta_n) \right)$, i.e.,

$$u \mapsto u' \frac{\mathbf{X}_n' \mathbf{X}_n}{n} u - 2\epsilon' \mathbf{X}_n u \frac{1}{\sqrt{n\lambda_n^*}} + 2 \sum_{j=1}^{k} \frac{\lambda_{n,j} \hat{w}_{n,j}}{\sqrt{n\lambda_n^*}} \left(|u_j + \sqrt{\frac{n}{\lambda_n^*}} \beta_{n,j}| - |\sqrt{\frac{n}{\lambda_n^*}} \beta_{n,j}| \right).$$

Since $\hat{\beta}_n^{\mathrm{A}}$ uniquely minimizes L_n, V_n also possesses a unique minimizer, which is given by $\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}} - \beta_n)$. In order to construct asymptotic confidence sets we follow the approach of Ewald and Schneider (2015), which is used to obtain asymptotic confidence sets for the LASSO estimator. First we derive $\lim_{n\to\infty} V_n$ equipped with a suitable topology. Afterwards we show that for these functions the convergence carries over to the convergence of their minimizers, which gives us the asymptotic distribution of the properly scaled and centred adaptive LASSO estimator. For this, we need some more definitions.

Definition 12. Denote by $\mathfrak{I} = \{j : \max(|\varphi_j|, \psi_j) = \infty\}$ and $\mathfrak{Z} = \{j : \max(|\varphi_j|, \psi_j) = 0\}$, fulfilling $\mathfrak{I} \cup \mathfrak{Z} \subseteq \{1, \dots, k\}$.

Definition 13. Let $V^{\varphi} : \mathbb{R}^k \to \overline{\mathbb{R}}$ be the following function

$$u \mapsto u'\mathbf{C}u + \sum_{j=1}^{k} \begin{cases} \infty & j \in \mathfrak{Z} \text{ and } u_j \neq 0 \\ 0 & j \in \mathfrak{I} \text{ or } u_j = 0 \\ 2\dfrac{|u_j + \lambda_{0,j}\varphi_j| - |\lambda_{0,j}\varphi_j|}{|\varphi_j + \psi_j Z_j|} & \text{else.} \end{cases}$$

In the case where $0 < \psi_j < \infty$, $u_j \neq 0$ and $\varphi_j = -\psi_j Z_j(\omega)$, we define the summand as ∞.

Remark. In our setting $0 < \psi_j < \infty$ implies $\lambda_{0,j} = 0$. Hence, the numerator is given by $|u_j|$ and therefore non-negative, which justifies the sign in the latter definition.

Lemma 14. The "weighting" terms of V_n converge in distribution

$$\frac{\hat{w}_{n,j}\lambda_{n,j}}{\sqrt{n\lambda_n^*}} \xrightarrow[n\to\infty]{d} \begin{cases} 0 & j \in \mathfrak{I} \\ \infty & j \in \mathfrak{Z} \\ (|\varphi_j + \psi_j Z_j|)^{-1} & \text{else.} \end{cases}$$

Moreover, the stacked "weighting" vector $\left(\frac{\hat{w}_{n,j}\lambda_{n,j}}{\sqrt{n\lambda_n^*}}\right)_{j=1}^{k}$ converges in distribution to the stacked vector of the limiting distributions.

Proof The following equation holds

$$\frac{\hat{\beta}_{n,j}^{\mathrm{LS}}\sqrt{n\lambda_n^*}}{\lambda_{n,j}} = \sqrt{n}\left(\hat{\beta}_{n,j}^{\mathrm{LS}} - \beta_{n,j}\right)\frac{\sqrt{\lambda_n^*}}{\lambda_{n,j}} + \sqrt{n}\beta_{n,j}\frac{\sqrt{\lambda_n^*}}{\lambda_{n,j}}.$$

For $j \in \mathfrak{I}$ the term $\frac{\hat{w}_{n,j}\lambda_{n,j}}{\sqrt{n\lambda_n^*}}$ converges according to Lemma 27 in probability to 0. On the other hand, the vector $\sqrt{n}(\hat{\beta}_n^{\mathrm{LS}} - \beta_n)$ converges in distribution to a normally distributed random vector Z with expectation 0 and covariance $\sigma^2\mathbf{C}^{-1}$. Applying Slutsky's Theorem, $(\frac{\sqrt{n\lambda_n^*}}{\hat{w}_{n,j}\lambda_{n,j}})_{j\notin\mathfrak{I}}$ converges in distribution to $(|\varphi_j + \psi_j Z_j|)_{j\notin\mathfrak{I}}$.

❑

The limiting distribution in Lemma 14 may contain non-deterministic parts only if there is a j fulfilling $0 < \psi_j < \infty$. This property holds true also for the following results, which makes the behaviour of ψ a subject of interest.

Remark. For the partial adaptive LASSO, that is $\exists j : \lambda_{n,j} = 0$ for all n, every non-penalized component j fulfils $\psi_j = \infty$.

Remark 15. For every fixed $j \in \{1, \dots, k\}$ either $\lambda_{0,j}$ or ψ_j is zero. This immediately follows from the definitions of these quantities.

Remark 16. In the following proposition we will deal with the limit ξ_j of $\sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}$. It is linked to φ_j the following way. If $\varphi_j \neq 0$ or $\lambda_{0,j} \neq 0$, then the equation $\varphi_j = \xi_j/\lambda_{0,j}$ holds true in the sense that for $\xi_j \neq 0, \lambda_{0,j} = 0$ the limit φ_j takes the value of $\operatorname{sgn}(\xi_j)\infty$.

Remark. By Remark 16 it follows in the case of uniform tuning ($\lambda_{n,j} = \lambda_n$ for all j) that φ equals ξ and $\psi_j = 0$ for all j.

The possibility of V^φ taking the value ∞ causes some disadvantages. For the ensuing proofs the following auxiliary functions, which can be considered as the finite parts of V_n, turned out to be practicable.

Definition 17. We define the function $g_n : \mathbb{R}^k \to \mathbb{R}$

$$u \mapsto u' \frac{\mathbf{X}_n' \mathbf{X}_n}{n} u - 2\varepsilon' \mathbf{X}_n u \frac{1}{\sqrt{n\lambda_n^*}} + 2 \sum_{j \notin \mathfrak{Z}} \frac{\lambda_{n,j}\hat{w}_{n,j}}{\sqrt{n\lambda_n^*}} \left(|u_j + \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| - \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| \right)$$

and the limiting function $g : \mathbb{R}^k \to \mathbb{R}$

$$u \mapsto u'\mathbf{C}u + 2 \sum_{j \notin (\mathfrak{I} \cup \mathfrak{Z})} \frac{|u_j + \lambda_{0,j}\varphi_j| - |\lambda_{0,j}\varphi_j|}{|\varphi_j + \psi_j Z_j|}.$$

The set $\mathcal{N} = \{\omega \in \Omega \mid \exists j \notin (\mathfrak{I} \cup \mathfrak{Z}) : \varphi_j + \psi_j Z_j(\omega) = 0\}$ possesses probability 0 and hence is of no interest if considering convergence in law / in probability. For convenience, we define the term in the third sum as 0, whenever $\varphi_j + \psi_j Z_j(\omega) = 0$. Thus, g and g_n are continuous, strict convex and finite.

In order to apply Lemma 25 we need another result concerning the pointwise convergence in law of the functions V_n to V^φ and g_n to g.

Proposition 18. *For every finite subset $\{s_1, \dots, s_l\} \subset \mathbb{R}^k$ we have*

$$(V_n(s_1), \dots, V_n(s_l)) \xrightarrow[n\to\infty]{d} (V^\varphi(s_1), \dots, V^\varphi(s_l)).$$

The same holds true for the sequence of the functions g_n with its limiting function g.

Proof First, we rewrite V_n as a sum of two functions

$$V_n^1(u) = u' \frac{\mathbf{X}_n' \mathbf{X}_n}{n} u - 2\varepsilon' \mathbf{X}_n u \frac{1}{\sqrt{n\lambda_n^*}} + 2 \sum_{j \in \mathfrak{Z} \cup \mathfrak{I}} \frac{\lambda_{n,j}\hat{w}_{n,j}}{\sqrt{n\lambda_n^*}} \left(|u_j + \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| - |\sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| \right)$$

$$V_n^2(u) = 2 \sum_{j \notin \mathfrak{Z} \cup \mathfrak{I}} \frac{\lambda_{n,j}\hat{w}_{n,j}}{\sqrt{n\lambda_n^*}} \left(|u_j + \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| - |\sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| \right)$$

The function V_n^1 represents those parts of V_n converging in probability to some limit, which allows us to conclude joint convergence from marginal convergence. For every fixed $u \in \mathbb{R}^k$, $u' \frac{\mathbf{X}_n' \mathbf{X}_n}{n} u$ converges to $u'\mathbf{C}u$, while $\varepsilon' \mathbf{X}_n u \frac{1}{\sqrt{n\lambda_n^*}}$ vanishes with rate $\frac{1}{\sqrt{\lambda_n^*}}$. For the remaining part of V_n^1 we remark that

$$\left| |u_j + \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| - |\sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| \right| \leq |u_j| \quad \text{as well as} \tag{4}$$

$$\sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j} = \frac{\lambda_{n,j}}{\lambda_n^*} \left(\sqrt{n}\beta_{n,j} \frac{\sqrt{\lambda_n^*}}{\lambda_{n,j}} \right) \quad \text{holds.} \tag{5}$$

Let j be arbitrary. In the case of $u_j = 0$ the corresponding term in the sum is 0. If $j \in \mathfrak{I}$, the convergence to 0 follows by Equation (4) and Lemma 14. Let $j \in \mathfrak{Z}$ and $u_j \neq 0$ hold true, then by equation (5) the term $|u_j + \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| - |\sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}|$ converges to $|u_j|$. Together with Lemma 14 the term converges to ∞ in probability. Altogether we have $(V_n^1(s_1), \dots, V_n^1(s_l)) \xrightarrow[n\to\infty]{p}$ $(V^1(s_1), \dots, V^1(s_l))$ with $V^1 : \mathbb{R}^k \to \mathbb{R}$ fulfilling

$$u \mapsto \begin{cases} u'\mathbf{C}u & u_j = 0 \text{ for all } j \in \mathfrak{Z} \\ \infty & \text{else.} \end{cases}$$

For every $j \notin \mathfrak{Z} \cup \mathfrak{I}$ by Equation (5) the expression $|u_j + \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}| - |\sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}|$ converges to $|u_j + \varphi_j\lambda_{0,j}| - |\varphi_j\lambda_{0,j}|$. From Lemma 14 we additionally have

$$\left(\frac{\hat{w}_{n,j}\lambda_{n,j}}{\sqrt{n\lambda_n^*}}\right)_{j\notin\mathfrak{I}\cup\mathfrak{Z}} \xrightarrow[n\to\infty]{d} \left(\frac{1}{|\varphi_j + \psi_j Z_j|}\right)_{j\notin\mathfrak{I}\cup\mathfrak{Z}}.$$

Denote by $V^2 : \mathbb{R}^k \to \mathbb{R}$ the function

$$u \mapsto 2\sum_{j\notin\mathfrak{I}\cup\mathfrak{Z}} \frac{|u_j + \lambda_{0,j}\varphi_j| - |\lambda_{0,j}\varphi_j|}{|\varphi_j + \psi_j Z_j|}$$

then, by the Continuous Mapping Theorem, it follows that $(V_n^2(s_1), \dots, V_n^2(s_l))$ converges in distribution to $(V^2(s_1), \dots, V^2(s_l))$. The convergence of g_n to g can be concluded analogously by simply reducing the summands in V_n^1 to $j \in \mathfrak{I}$ and redefining V^1 as $u \mapsto u'\mathbf{C}u$.

❑

The functions g_n, $n \in \mathbb{N}$, are strictly convex and fulfil

$$g_n(u) \xrightarrow[\|u\|\to\infty]{p} \infty. \tag{6}$$

The same property holds true for V_n, g and V^φ. Thus, all of them possess unique minimizers. Denote m the minimizer of V^φ and m_n the minimizer of V_n, which is given by $\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}} - \beta_n)$. We then immediately conclude $m_j = 0$ for $j \in \mathfrak{Z}$, as otherwise V^φ would become ∞.

Theorem 19. *The scaled and centred adaptive LASSO estimator converges in distribution to the minimizer of V^φ, i.e,*

$$\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^A - \beta_n) \xrightarrow[n\to\infty]{d} \operatorname*{arg\,min}_{u\in\mathbb{R}^k} V^\varphi(u). \tag{7}$$

Proof To prove the statement above directly, V^φ needs to be finite on $\mathbb{R}^k$. As this is not necessarily fulfilled, we split the proof into two parts. In the first one, we prove the convergence in probability of $m_{n,j}$ to m_j for all $j \in \mathfrak{Z}$. In the second part we restrict the definition range of our functions to those dimensions, on which V^φ is finite in order to apply Lemma 25. However, g_n fulfils the conditions of Lemma 25, which gives us

$$\min_{u\in\mathbb{R}^k} g_n(u) \xrightarrow[n\to\infty]{d} \min_{u\in\mathbb{R}^k} g(u).$$

By the Prohorov Theorem the sequence $\min_{u\in\mathbb{R}^k} g_n(u)$ is tight. From $V_n(m_n) \leq 0$ we have for every $j \in \mathfrak{Z}$

$$0 \leq \frac{\lambda_{n,j}}{\sqrt{n\lambda_n^*}}\hat{w}_{n,j}\left|m_{n,j} + \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}\right| \leq -g_n(m_n) + \sum_{i\in\mathfrak{Z}} \hat{w}_{n,i}\frac{\lambda_{n,i}}{\lambda_n^*}|\beta_{n,i}|.$$

To see the boundedness of $\sum_{j\in\mathfrak{Z}} \frac{\lambda_{n,j}}{\lambda_n^*}\hat{w}_{n,j}|\beta_{n,j}|$ we apply, again, the Prohorov Theorem. Due to the compactness of $\overline{\mathbb{R}}$, every subsequence of $\hat{w}_{n,j}|\beta_{n,j}|$ contains a subsubsequence fulfilling $\lim_{m\to\infty} \sqrt{(}n_{l_m})\beta_{n_{l_m},j} = c$ for some $c \in \overline{\mathbb{R}}$. This subsubsequence converges in distribution to $\frac{|c|}{|c+Z_j|}$ (becoming one in the case $|c| = \infty$), where Z_j stands for the j-th element of the k-dimensional, normally distributed random vector of Lemma 14. As every subsequence of $\hat{w}_{n,j}|\beta_{n,j}|$ contains a subsubsequence converging in distribution, the sum on the right-hand side is tight. Together with the fact, that $\limsup_{n\to\infty} \mathbb{P}(-g_n(m_n) \leq K) \geq \mathbb{P}(-\min_{u\in\mathbb{R}^k} g(u) \leq K)$ for every $K \in \mathbb{R}$, we conclude the tightness of the right-hand side. Since $\frac{\lambda_{n,j}}{\sqrt{n\lambda_n^*}}\hat{w}_{n,j}$ converges to ∞ for every $j \in \mathfrak{Z}$, the term $|m_{n,j} + \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j}|$ converges in probability to 0. Finally, $\lim_{n\to\infty} \sqrt{\frac{n}{\lambda_n^*}}\beta_{n,j} = \lim_{n\to\infty} \frac{\sqrt{n\lambda_n^*}}{\lambda_{n,j}}\frac{\lambda_{n,j}}{\lambda_n^*}\beta_{n,j} = 0$ implies $m_{n,j} \xrightarrow[n\to\infty]{p} 0$. As mentioned above, m_j equals 0 for $j \in \mathfrak{Z}$, which gives us $m_{n,\mathfrak{Z}} \xrightarrow[n\to\infty]{p} m_{\mathfrak{Z}}$. For the second part of the proof, denote s the cardinality of the set $\mathfrak{Z}$ and, without loss of generality, let $\mathfrak{Z} = \{1, \ldots, s\}$. If $s = k$, then the proof is complete. Otherwise we define embedding functions $\iota_n : \mathbb{R}^{k-s} \to \mathbb{R}^k$ at the point m_n

$$[\iota_n(u)]_j = \begin{cases} u_{j-s} & j > s \\ m_{n,j} & j \leq s \end{cases}.$$

The embedding function $\iota : \mathbb{R}^{k-s} \to \mathbb{R}^k$ at m is defined analogously as

$$[\iota(u)]_j = \begin{cases} u_{j-s} & j > s \\ m_j & j \leq s \end{cases}.$$

Now we can define $\tilde{g}_n : \mathbb{R}^{k-s} \to \mathbb{R}$ as the function composition $g_n \circ \iota_n$

$$v \mapsto g_n(\iota_n(v))$$

and $\tilde{g}$ as $g \circ \iota$ analogously. The functions $\tilde{g}_n$ and $\tilde{g}$ are strictly convex and, due to equation (6), possess a unique minimizer. On the set $\mathfrak{E} = \{x \in \mathbb{R}^k : x_j = 0 \text{ for all } j \in \mathfrak{Z}\}$ the functions g and V^φ coincide almost surely. Therefore the minimizer of $\tilde{g}$ equals the minimizer of $V^\varphi \circ \iota$ almost surely, which in turn is – according to its definition – $m_{\mathfrak{Z}^\mathsf{C}}$. Furthermore, the functions $\tilde{g}_n$ and $V_n \circ \iota_n$ differ only by a constant, which implies the equality of its minimizers. For every fixed $v \in \mathbb{R}^{k-s}$ the sequence $(\tilde{g}_n(v))_{n\in\mathbb{N}}$ converges in distribution to $\tilde{g}(v)$. Hence, we conclude

$$\operatorname*{arg\,min}_{v\in\mathbb{R}^{k-s}} g_n(\iota_n(v)) \xrightarrow[n\to\infty]{d} \operatorname*{arg\,min}_{v\in\mathbb{R}^{k-s}} g(\iota(v)).$$

Altogether we have

$$m_{n,\mathfrak{Z}^\mathsf{C}} = \operatorname*{arg\,min}_{v\in\mathbb{R}^{k-s}} V_n(\iota_n(v)) = \operatorname*{arg\,min}_{v\in\mathbb{R}^{k-s}} g_n(\iota_n(v)) \xrightarrow[n\to\infty]{d} \operatorname*{arg\,min}_{v\in\mathbb{R}^{k-s}} g(\iota(v)) = \operatorname*{arg\,min}_{v\in\mathbb{R}^{k-s}} V^\varphi(\iota(v)) = m_{\mathfrak{Z}^\mathsf{C}},$$

where the equality after the limit is fulfilled almost surely. Together with the fact $m_{n,\mathfrak{Z}} \xrightarrow[n\to\infty]{p} m_{\mathfrak{Z}}$ we conclude the desired result.

❑

From the foregoing result we now can conclude that the convergence rate of $\sup_{\beta\in\mathbb{R}^k}(\hat{\beta}_n^{\mathrm{A}}-\beta)$ is indeed of order $\sqrt{\frac{n}{\lambda_n^*}}$. Since there must be at least one j^* fulfilling $\psi_{j^*}=0$, we can always find a φ, such that $\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}}-\beta)$ converges to a random vector not being 0. Together with Proposition 5, which gives an upper bound for the convergence rate, we conclude the desired result. Moreover, with a slightly improved argumentation we can also show that the convergence rate of $\sup_{\beta\in\mathbb{R}^k}\mathbb{E}(\hat{\beta}^{\mathrm{A}}-\beta)$ equals the same rate. Defining $(\beta_n)_{n\in\mathbb{N}}$ such that $\varphi_i=\infty$ for all $i\neq j^*$ and $|\varphi_{j^*}|$ being neither 0 nor ∞, the limiting random vector is deterministic. Furthermore, the expectation of its j^*-th component does not equal 0 either. On the other hand, Lemma 10 ensures the uniform integrability of $\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}}-\hat{\beta}_n^{\mathrm{LS}})_{j^*}$. As $\lim_{n\to\infty}\lambda_n^*=\infty$, the latter expression converges in probability to the same limit as $\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}}-\beta_n)_{j^*}$. Thus, we have

$$\lim_{n\to\infty}\mathbb{E}\left(\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}}-\beta_n)_{j^*}\right)=\lim_{n\to\infty}\mathbb{E}\left(\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}}-\hat{\beta}_n^{\mathrm{LS}})_{j^*}\right)=$$
$$\mathbb{E}\left(\lim_{n\to\infty}\sqrt{\frac{n}{\lambda_n^*}}(\hat{\beta}_n^{\mathrm{A}}-\hat{\beta}_n^{\mathrm{LS}})_{j^*}\right)=\mathbb{E}\left(\operatorname*{arg\,min}_{u\in\mathbb{R}^k}V^{\varphi}(u)\right)_{j^*}\neq 0.$$

Now the upper bound of $\sup_{\beta\in\mathbb{R}^k}\mathbb{E}(\hat{\beta}^{\mathrm{A}}-\beta)$ in the subsection 3.3 implies that its convergence rate coincides with $\sqrt{\frac{n}{\lambda_n^*}}$.

4.2 Alternative representation

In order to get an asymptotically valid confidence set one may consider the union of the minimizers of the functions V^{φ}. However, they are defined implicitly and may be hard to calculate. Therefore we are looking for a more convenient representation of that set.

Definition 20. Let $\lambda_0\in[0,1]^k$ and $\psi\in[0,\infty]^k$. Then $\mathcal{M}$ denotes the following set

$$\mathcal{M}=\mathcal{M}^{\lambda_0,\psi}=\{m\in\mathbb{R}^k : m_j(\mathbf{C}m)_j\leq\lambda_{0,j}\text{ for all }j\}\cap\{m\in\mathbb{R}^k : (\psi_j=\infty)\Rightarrow((\mathbf{C}m)_j=0)\}.$$

Proposition 21. *Let $\psi\in[0,\infty]^k$ and denote $(\Omega,\mathfrak{S},\mu)$ the probability space of Lemma 14. Then for all $\omega\in\Omega$ the set $\mathcal{M}$ coincides with the set of all minimizers of the function V^{φ}.*

$$\mathcal{M}=\bigcup_{\varphi\in\overline{\mathbb{R}}^k}\operatorname*{arg\,min}_{u\in\mathbb{R}^k}V^{\varphi}(u)(\omega)\quad\text{ for all }\omega\in\Omega.$$

In particular, the sets coincide μ-almost surely.

Proof Take an arbitrary $\omega\in\Omega$ and denote $\widetilde{M}=\bigcup_{\varphi\in\overline{\mathbb{R}}^k}\operatorname*{arg\,min}_{u\in\mathbb{R}^k}V^{\varphi}(u)(\omega)$.

First, we want to prove $\widetilde{M}\subseteq\mathcal{M}$. Let $m\in\widetilde{M}$ and $j\in\{1,\dots,k\}$. By definition there is a $\varphi\in\overline{\mathbb{R}}^k$, such that $m=\operatorname*{arg\,min}_{u\in\mathbb{R}^k}V^{\varphi}(u)(\omega)$. If $j\in\mathfrak{I}$, that is to say $\max(|\varphi_j|,|\psi_j|)=\infty$, then the directional derivative in the direction of the j-th unit vector e_j exists. Since m is a minimizer, it follows that

$$0=\frac{1}{2}\frac{\partial V^{\varphi}}{\partial u_j}(m)(\omega)=(\mathbf{C}m)_j\,.$$

We would like to emphasize that $j \in \mathfrak{I}$ already contains the special case $\psi_j = \infty$, for which the stricter condition $(\mathbf{C}m)_j = 0$ has to be fulfilled. Now we consider the case $\varphi_j = -\psi_j Z_j(\omega)$, which includes $j \in \mathfrak{Z}$ as well as $\varphi_j = 0 = \psi_j Z_j(\omega)$. Here, the objective function $V^{\varphi}(\cdot)(\omega)$ is finite if and only if $m_j = 0$. This, again, gives us $m_j(\mathbf{C}m)_j \leq 0 \leq \lambda_{0,j}$. For the following case analysis we always additionally suppose $\varphi_j \neq -\psi_j Z_j(\omega)$. In the case of $0 < \max(|\varphi_j|, |\psi_j Z_j(\omega)|) < \infty$ and $\lambda_{0,j} = 0$, m_j either equals 0 or the directional derivative of V^{φ} at the point m in the direction e^j exists. In the latter case we have

$$0 = \frac{1}{2}\frac{\partial V^{\varphi}}{\partial u_j}(m)(\omega) = (\mathbf{C}m)_j + \frac{\operatorname{sgn}(m_j)}{|\varphi_j + \psi_j Z_j(\omega)|}$$

Thus, m_j fulfills

$$m_j\,(\mathbf{C}m)_j = -\frac{|m_j|}{|\varphi_j + \psi_j Z_j(\omega)|} < 0 = \lambda_{0,j}.$$

Considering the case $0 < \max(|\varphi_j|, |\psi_j Z_j(\omega)|) < \infty, \lambda_{0,j} > 0$ and $m_j = -\lambda_{0,j}\varphi_j$ we conclude $\psi_j = 0$ due to remark 15. The one-sided partial derivatives of V^{φ} at m have to be nonnegative, which gives us $|\,(\mathbf{C}m)_j\,| \leq \dfrac{1}{|\varphi_j|}$ and furthermore

$$\lambda_{0,j} \geq \lambda_{0,j}|\,(\mathbf{C}m)_j\,\varphi_j| = |\,(\mathbf{C}m)_j\,m_j|.$$

If $0 < \max(|\varphi_j|, |\psi_j Z_j(\omega)|) < \infty, \lambda_{0,j} > 0$ and $m_j \neq -\lambda_{0,j}\varphi_j$, then the j-partial derivative of V^{φ} at m exists and equals 0. Hence, we conclude

$$0 = \frac{1}{2}\frac{\partial V^{\varphi}}{\partial u_j}(m)(\omega) = (\mathbf{C}m)_j + \frac{\operatorname{sgn}(m_j + \lambda_{0,j}\varphi_j)}{|\varphi_j|}. \tag{8}$$

In the subcase $|m_j| \leq \lambda_{0,j}|\varphi_j|$ it follows by equation (8)

$$\lambda_{0,j} = \lambda_{0,j}|\,(\mathbf{C}m)_j\,\varphi_j| \geq |\,(\mathbf{C}m)_j\,m_j|,$$

while the other subcase $|m_j| > \lambda_{0,j}|\varphi_j|$ implies

$$m_j\,(\mathbf{C}m)_j = -\frac{m_j \operatorname{sgn}(m_j + \lambda_{0,j}\varphi_j)}{|\varphi_j|} < 0.$$

To summarise, every $m \in \widetilde{M}$ fulfils $m_j(\mathbf{C}m)_j \leq \lambda_{0,j}$ and $(\psi_j = \infty) \Rightarrow ((\mathbf{C}m)_j = 0)$, which gives us $\widetilde{M} \subseteq \mathcal{M}$. For the opposite direction we assume $m \in \mathcal{M}$. We want to find φ, such that m is the minimizer of $V^{\varphi}(\cdot)(\omega)$. In order to do so, we define φ as follows

$$\varphi_j = \begin{cases} \infty & (\mathbf{C}m)_j = 0 \\ -\frac{m_j}{\lambda_{0,j}} & (\mathbf{C}m)_j \neq 0, \lambda_{0,j} > 0 \text{ and } |m_j(\mathbf{C}m)_j| \leq \lambda_{0,j} \\ -\psi_j Z_j(\omega) + \frac{1}{|(\mathbf{C}m)_j|} & \text{else} \end{cases}.$$

Denote $\mathcal{Q} := \{j : \varphi_j = 0 \text{ and } \lambda_{0,j} > 0\}$ and $\mathcal{P} := \{j : (\mathbf{C}m)_j = 0\}$. For a given $j \in \mathcal{Q}$ the function $V^{\varphi}(u)(\omega)$ is finite if and only if $u_j = 0$. As $\mathcal{Q}$ is a subset of $\{j : m_j = 0\}$, m fulfils this property. Moreover, $\underset{u\in\mathbb{R}^k}{\arg\min} V^{\varphi}(u)$ coincides with $\underset{u\in\mathbb{R}^k : u_j = 0\ \forall j\in\mathcal{Q}}{\arg\min} V^{\varphi}(u)$. Thus, it

suffices to concern only about one-sided directional derivatives of $V^{\varphi}(\cdot)(\omega)$ with direction in $\{r \in \mathbb{R}^k : r_j = 0 \text{ for all } j \in \mathcal{Q}\}$.

$$\frac{1}{2}\frac{\partial V^{\varphi}}{\partial r}(m)(\omega) - r'\mathbf{C}m = \sum_{j\in\mathcal{Q}} 0 + \sum_{j\in\mathcal{P}} 0 + \sum_{\substack{\lambda_{0,j}=0, j\notin\mathcal{Q}\cup\mathcal{P} \\ m_j=0}} \frac{|r_j|}{|\varphi_j + \psi_j Z_j(\omega)|} +$$

$$\sum_{\substack{\lambda_{0,j}=0, j\notin\mathcal{Q}\cup\mathcal{P} \\ m_j\neq 0}} \frac{r_j \operatorname{sgn}(m_j)}{|\varphi_j + \psi_j Z_j(\omega)|} + \sum_{\substack{\lambda_{0,j}>0, j\notin\mathcal{Q}\cup\mathcal{P} \\ |m_j(\mathbf{C}m)_j|\leq\lambda_{0,j}}} \frac{|r_j|}{|\varphi_j|} + \sum_{\substack{\lambda_{0,j}>0, j\notin\mathcal{Q}\cup\mathcal{P} \\ |m_j(\mathbf{C}m)_j|>\lambda_{0,j}}} \frac{r_j \operatorname{sgn}(m_j + \lambda_{0,j}\varphi_j)}{|\varphi_j|}$$

In order to recognize that this summation contains all possible cases, we mention that $\lambda_{0,j} > 0$ already implies $\psi_j = 0$. The third sum's terms correspond with $|r_j(\mathbf{C}m)_j|$ and are greater or equal than $-r_j(\mathbf{C}m)_j$. The fourth sum implies $m_j(\mathbf{C}m)_j < 0$ and its summands equal $-r_j(\mathbf{C}m)_j$. In the penultimate sum the terms equal $|r_j \frac{\lambda_{0,j}}{m_j}|$ and hence are greater or equal than $|r_j(\mathbf{C}m)_j|$. For the last sum's term it holds $\psi_j = 0$ due to Remark 15 and $\varphi_j = \frac{1}{|(\mathbf{C}m)_j|}$. From $|m_j| > \frac{\lambda_{0,j}}{|(\mathbf{C}m)_j|}$ and $m_j(\mathbf{C}m)_j \leq \lambda_{0,j}$ we conclude $\operatorname{sgn}(m_j + \lambda_{0,j}\varphi_j) = \operatorname{sgn}(m_j) = -\operatorname{sgn}((\mathbf{C}m)_j)$. Hence, all summands of the last sum equal $-r_j(\mathbf{C}m)_j$. Altogether, the one-sided directional derivative is nonnegative and – due to the strict convexity of the objective function – m is the minimizer of $V^{\varphi}(\cdot)(\omega)$.

❑

5 Confidence sets

5.1 Construction of the Confidence sets

> Mathematics consists in proving the most obvious thing in the least obvious way.
>
> *George Polya*

Theorem 22. *Let* $\frac{\sqrt{\lambda_n^*}}{\lambda_{n,j}} \to \psi_j \in [0,\infty]$ *for all* j*. Then, for every open superset* $\mathcal{O}$ *of* $\mathcal{M}$*, it holds*

$$\lim_{n\to\infty} \inf_{\beta\in\mathbb{R}^k} \mathbb{P}\left(\hat{\beta}_n^A - \beta \in \sqrt{\frac{\lambda_n^*}{n}}\mathcal{O}\right) = 1.$$

On the other hand, for all $d < 1$ *we have*

$$\lim_{n\to\infty} \inf_{\beta\in\mathbb{R}^k} \mathbb{P}\left(\hat{\beta}_n^A - \beta \in \sqrt{\frac{\lambda_n^*}{n}}\mathcal{M}^{d\lambda_0,\psi}\right) = 0.$$

Proof We start proving the first claim. Let $c = \liminf_{n\to\infty} \inf_{\beta\in\mathbb{R}^k} \mathbb{P}(\beta^A - \beta \in \sqrt{\frac{\lambda_n^*}{n}}\mathcal{O})$. Hence, there is a subsequence n_l, such that

$$c = \lim_{l\to\infty} \inf_{\beta\in\mathbb{R}^k} \mathbb{P}(\beta^A - \beta \in \sqrt{\frac{\lambda_{n_l}^*}{n_l}}\mathcal{O})$$

For an arbitrary $\epsilon > 0$ there is a sequence β_{n_l}, such that

$$\left(\mathbb{P}_{n_l}(\hat{\beta}^{\mathrm{A}} - \beta_{n_l} \in \sqrt{\frac{\lambda^*_{n_l}}{n_l}}\mathcal{O}) - \inf_{\beta \in \mathbb{R}^k} \mathbb{P}_{\beta}(\hat{\beta}^{\mathrm{A}} - \beta \in \sqrt{\frac{\lambda^*_{n_l}}{n_l}}\mathcal{O}) \right) \leq \varepsilon \text{ for all } l$$

holds true. If we could show

$$\lim_{l \to \infty} \mathbb{P}_{\beta_{n_l}}(\hat{\beta}^{\mathrm{A}}_{n_l} - \beta_{n_l} \in \sqrt{\frac{\lambda^*_{n_l}}{n_l}}\mathcal{O}) = 1,$$

then the proof would be completed due to the arbitrariness of ϵ. However, we can neither assume the existence of this limit, nor the convergence of the vector $(\sqrt{n_l}\beta_{n_l,j}\frac{\sqrt{\lambda^*_{n_l,j}}}{\lambda_{n_l}})_{j=1}^k$, and therefore have to deal again with subsequences. Denote

$$d = \liminf_{l \to \infty} \mathbb{P}_{\beta_{n_l}} \left(\hat{\beta}^{\mathrm{A}}_{n_l} - \beta_{n_l} \in \sqrt{\frac{\lambda^*_{n_l}}{n_l}}\mathcal{O} \right).$$

The sequence n_l contains a subsequence n_{l_p}, such that

$$d = \lim_{p \to \infty} \mathbb{P}_{\beta_{n_{l_p}}} \left(\hat{\beta}^{\mathrm{A}}_{n_{l_p}} - \beta_{n_{l_p}} \in \sqrt{\frac{\lambda^*_{n_{l_p}}}{n_{l_p}}}\mathcal{O} \right)$$

holds true. There is another subsequence $n_{l_{pq}}$, such that the vector

$$\left(\sqrt{n_{l_{pq}}}\beta_{n_{l_{pq}},j}\frac{\sqrt{\lambda^*_{n_{l_{pq}}}}}{\lambda_{n_{l_{pq}},j}} \right)_{j=1}^k$$

converges to $\varphi \in \overline{\mathbb{R}}^k$. According to Proposition 19 it follows

$$\sqrt{\frac{n_{l_{pq}}}{\lambda^*_{n_{l_{pq}}}}} \left(\hat{\beta}^{\mathrm{A}}_{n_{l_{pq}}} - \beta_{n_{l_{pq}}} \right) \xrightarrow[q \to \infty]{d} \operatorname*{arg\,min}_{u \in \mathbb{R}^k} V^{\varphi}(u).$$

Finally, the Portemanteau-Theorem implies

$$\liminf_{q \to \infty} \mathbb{P}_{\beta_{n_{l_{pq}}}} \left(\sqrt{\frac{n_{l_{pq}}}{\lambda^*_{n_{l_{pq}}}}} \left(\hat{\beta}^{\mathrm{A}}_{n_{l_{pq}}} - \beta_{n_{l_{pq}}} \right) \in \mathcal{O} \right) \geq \mathbb{P}_{\varphi} \left(\operatorname*{arg\,min}_{u \in \mathbb{R}^k} V^{\varphi}(u) \in \mathcal{M} \right) = 1$$

For the second part of the proof, denote $\mathcal{S} = \{j : \lambda_{0,j} = 0\}$ and $r = \mathbf{C}^{-1}\lambda_0 \in \mathbb{R}^k$. As there is at least one j with $\lambda_{0,j} = 1$, the set $\{1, \ldots, k\} \backslash \mathcal{S}$ is not empty and r is not the zero vector. Due to the positive definiteness of $\mathbf{C}$,

$$0 < r' \mathbf{C} r = \sum_{j \notin \mathcal{S}} r_j \lambda_{0,j},$$

implies that there exists at least one strictly positive r_j. Defining m as $r \left(\max_{l \notin \mathcal{S}} r_l \right)^{-\frac{1}{2}}$, it satisfies the equations $m_j(\mathbf{C}m)_j = \lambda_0 \frac{r_j}{\max_{l \notin \mathcal{S}} r_l}$ and is contained in $\mathcal{M}^{\lambda_0, \psi} \backslash \mathcal{M}^{d\lambda_0, \psi}$. In order to see this, we

stress the fact that $|\psi_j| = \infty$ implies $j \in \mathcal{S}$. Let

$$\varphi_j = \begin{cases} \infty & j \in \mathcal{S} \\ -\frac{m_j}{\lambda_{0,j}} & j \notin \mathcal{S} \text{ and } |m_j(\mathbf{C}m)_j| \leq \lambda_{0,j} \\ \frac{1}{|(\mathbf{C}m)_j|} & \text{else} \end{cases},$$

then, analogously to the second part of the proof of Lemma 21, m is the unique minimizer of V^{φ}. As for all j in $\{j : 0 < \psi_j < \infty\} \subseteq \mathcal{S}$ the value of φ_j is infinite, the objective function V^{φ} is non-stochastic. This in turn implies for all sequences $(\tilde{\beta}_n)_{n\in\mathbb{N}}$ with limit φ

$$m_n = \underset{u\in\mathbb{R}^k}{\arg\min} V_n^{\tilde{\beta}_n,\lambda_n}(u) \xrightarrow[n\to\infty]{p} \underset{u\in\mathbb{R}^k}{\arg\min} V^{\varphi}(u) = m.$$

Because of the continuity of the function $m \mapsto m_j(\mathbf{C}m)_j$, $m_n \in \mathcal{M}^{d\lambda_0,\psi}$ implies $\|m_n - m\| \geq \epsilon$ and therefore

$$\begin{aligned} \limsup_{n\to\infty} \inf_{\beta\in\mathbb{R}^k} \mathbb{P}\left(\hat{\beta}_n^{\mathrm{A}} - \beta \in \sqrt{\frac{\lambda_n^*}{n}}\mathcal{M}^{d\lambda_0,\psi}\right) &\leq \limsup_{n\to\infty} \mathbb{P}_{\tilde{\beta}_n}\left(\hat{\beta}_n^{\mathrm{A}} - \tilde{\beta}_n \in \sqrt{\frac{\lambda_n^*}{n}}\mathcal{M}^{d\lambda_0,\psi}\right) \\ &= \limsup_{n\to\infty} \mathbb{P}_{\tilde{\beta}_n}\left(m_n \in \mathcal{M}^{d\lambda_0,\psi}\right) \\ &\leq \limsup_{n\to\infty} \mathbb{P}_{\tilde{\beta}_n}\left(\|m_n - m\| \geq \epsilon\right) = 0 \end{aligned}$$

❑

Remark. Pötscher and Schneider (2010) derived one-dimensional confidence intervals based on the adaptive LASSO estimator for orthonormal regresssors and uniform tuning. In this special case the confidence sets of Theorem 22 are given by open supersets of $[\hat{\beta}_j^{\mathrm{A}} - \sqrt{\frac{n}{\lambda_n}}, \hat{\beta}_j^{\mathrm{A}} + \sqrt{\frac{n}{\lambda_n}}]$. Hence, our results essentially coincide with the proposed intervals $(\hat{\beta}_j^{\mathrm{A}} - d\sqrt{\frac{n}{\lambda_n}}, \hat{\beta}_j^{\mathrm{A}} + d\sqrt{\frac{n}{\lambda_n}})$ (with $d > 1$).

Remark. Since $\lambda_{0,j}$ may take the value 0, a classification of the confidence sets as in Ewald and Schneider (2015) is not possible. Assume that $\mathbf{C}$ is a diagonal matrix and there is at least one $\tilde{j}$ fulfilling $\lambda_{0,\tilde{j}} = 0$. Then, for every $d \in \mathbb{R}$, $d\mathcal{M}$ does not contain any elements with $m_{\tilde{j}} \neq 0$. Hence, it cannot be an open strict superset.

Remark. The set $\mathcal{M}$ is symmetric. In Pötscher and Schneider (2010) it was shown for the case of orthogonal regressors, that symmetric intervals are the shortest confidence sets for a given infimal coverage probability. However, this result was proven in the finite sample framework.

5.2 Illustration

Figure 1 illustrates the set $\mathcal{M}$ in the case of $\lambda_{0,j} = 1$ and $\psi_j = 0$ for all j with $\mathbf{C}$ being $\left[\begin{smallmatrix} 1 & -0.7 \\ -0.7 & 1 \end{smallmatrix}\right]$, which is the case for uniform tuning and negative correlated covariates. The colour indicates the value of $\max_{1\leq j\leq 2} m_j(\mathbf{C}m)_j$ at the specific point m inside the set. The higher the absolute value of the correlation of the covariates is, the flatter and more stretched the confidence set becomes. As one may expect intuitively, in the case of negative correlation, the confidence set covers more of the area, where the covariates' signs equal. A positive correlation causes the opposite behaviour.

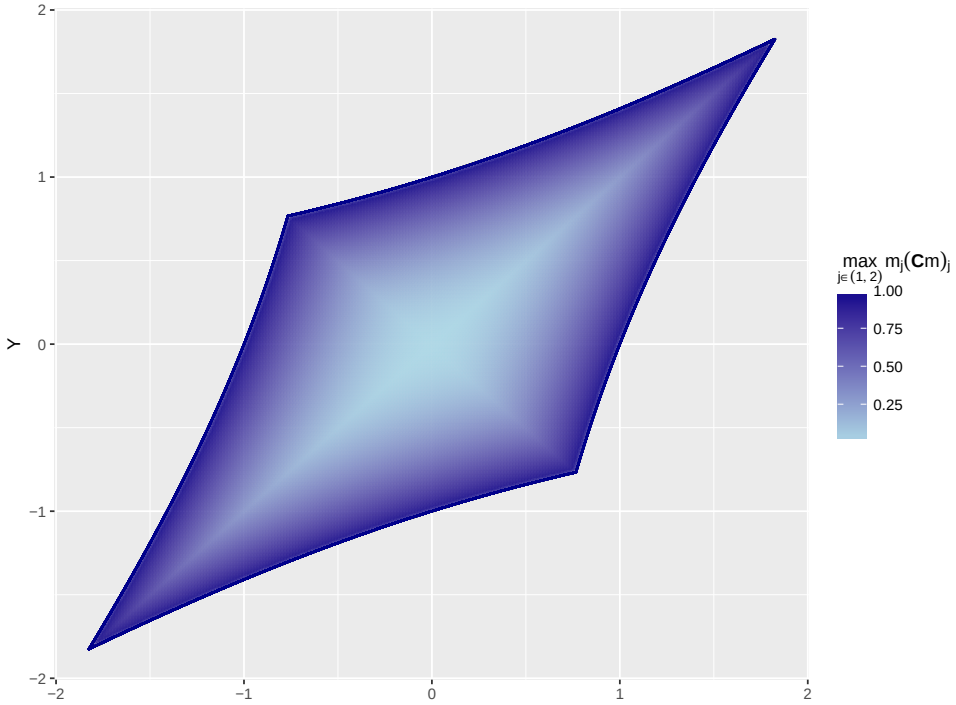

Figure 1: Two-dimensional example of $\mathcal{M}$ in the case of uniform tuning

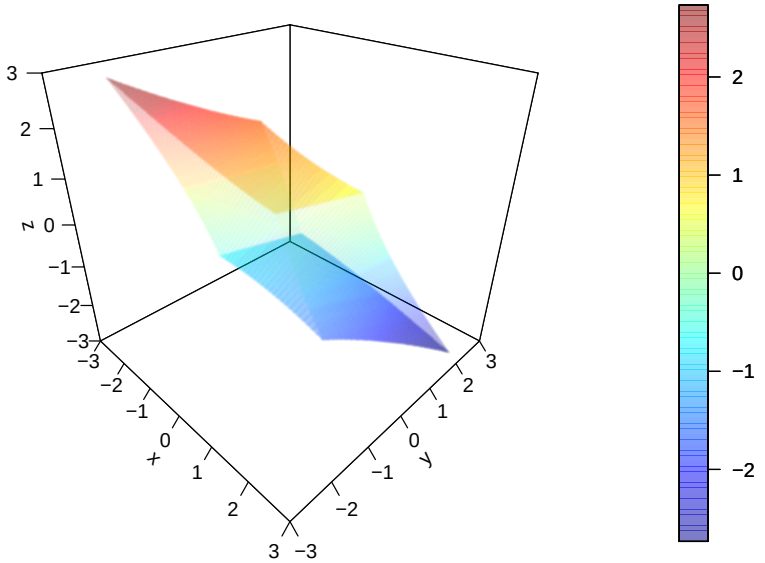

Figure 2: Three-dimensional example of $\mathcal{M}$ in the case of uniform tuning

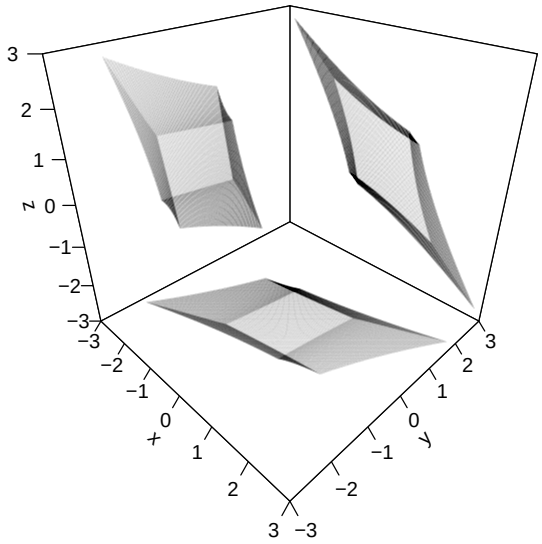

Figure 3: Projections of the three-dimensional set of Figure 2

As a three-dimensional example of $\mathcal{M}$ we consider again uniform tuning with

$$\mathbf{C} = \begin{bmatrix} 1 & -0.3 & 0.7 \\ -0.3 & 1 & 0.2 \\ 0.7 & 0.2 & 1 \end{bmatrix},$$

which is illustrated in Figure 2. To give a better impression of the structure, the set is coloured depending on the value of the third coordinate. Here, the high correlation of the first and the third covariate „stretches“ the set in the direction, where the covariates' signs differ. From an interpretative point of view this can be read as follows: Due to the hight correlation of the regressor variables, the underestimation of one component compensating the overestimation of the other one would lead to similar results in the observated values and hence is more likely to occur. Figure 3 shows the projections of the three-dimensional set of Figure 2 onto three planes where one component is held fixed. The projection on that plane, where the second component is fixed, clearly shows the behaviour explained above. On the other hand, the other two projections emphasize that for covariates with a lower correlation (in absolute value) the confidence set is less distorted. Figure 4 shows the case of a partial adaptive LASSO estimator with the same matrix $\mathbf{C}$, where the first component is not penalized while the other ones are uniformly tuned. Hence, we have $\lambda_{0,1} = 0$, $\psi_1 = \infty$, $\lambda_{0,j} = 1$ as well as $\psi_j = 0$ for $j > 1$. Due to the condition $(\mathbf{C}m)_1 = 0$ for all $m \in \mathcal{M}$ that set is the intersection of a plane with the set of Figure 2.

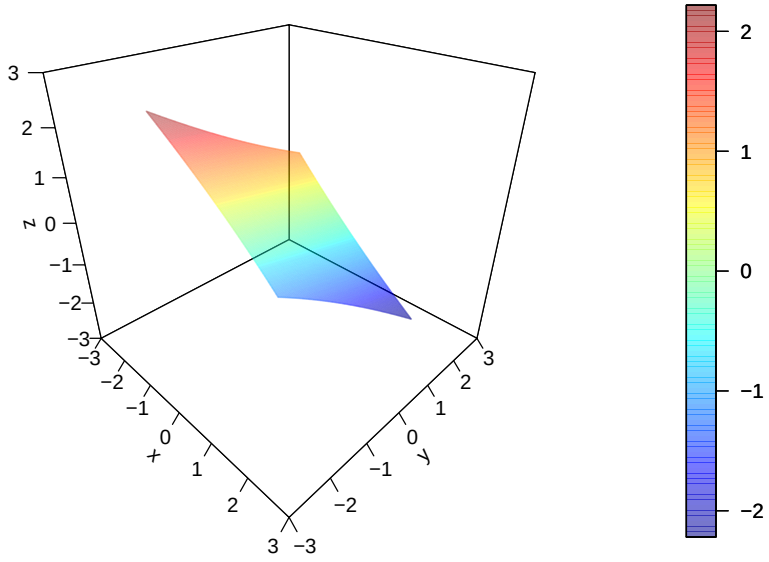

Figure 4: Three-dimensional example of $\mathcal{M}$ in the case of partial tuning

6 Conclusion

> Education is what you get when you read the fine print. Experience is what you get when you don't.
>
> *Arthur Levitt*

We have studied the asymptotic behaviour of the adaptive LASSO estimator in the framework of componentwise tuning and covariates being not necessarily orthogonal. Hence, our results generalize the findings in Pötscher and Schneider (2009), which are also derived in the low-dimensional setting. Regarding consistency in parameter estimation in the moving-parameter framework, the necessary and sufficient condition on the tuning parameter barely differs from those of the one-dimensional case and intuitively carries over to componentwise tuning. Consistency in parameter estimation is necessary to guarantee consistency in model selection as well. Besides that condition, a multidimensional equivalent of the uniform tuning's condition for consistency in model selection has to be fulfilled, too. Depending on the underlying model structure, however, another condition limiting the deviation of the tuning parameter's convergence rates may be required to ensure consistent model selection.

The main result of this work consists in the construction of confidence sets, which are derived in the framework of partial consistent tuning: At least one component is tuned to perform consistent model selection, while the other components may be tuned arbitrarily. First, the asymptotic distribution of the appropriately scaled and centred adaptive LASSO estimator is derived implicitly and expressed as a function's minimizer. Afterwards, we present a more practicable expression of the set of all possible minimizers, which makes it far more easier to

compute. These results are used to create confidence sets with asymptotic infimal coverage probability 1. As a side result, we derived the convergence rate (towards 0) of the supremum of the adaptive LASSO estimator's bias in the case of partial consistent tuning.

A Appendix

> If ℓ_2 was the norm of the 20th century, then ℓ_1 is the norm of the 21st century ... OK, maybe that statement is a bit dramatic, but at least so far, there's been a frenzy of research involving the ℓ_1 norm and its sparsity-inducing properties.
>
> *Ryan J. Tibshirani* (in R. J. Tibshirani and Wasserman, 2015)

A.1 Random functions

Definition 23. Let $(\Omega, \mathcal{A})$ be a measurable space and denote $\mathcal{B}_b$ the Borel-σ-algebra of $\mathbb{R}^b$ equipped with the Euclidean metric. A mapping $f : \Omega \times \mathbb{R}^a \to \mathbb{R}^b$ is called a stochastic function if for every $x \in \mathbb{R}^a$ the mapping $f(\cdot, x) : \Omega \to \mathbb{R}^b$ is n Borel random vector on $(\Omega, \mathcal{A})$.

A.2 Asymptotics of minimizers

In this section we recapitulate some results for the asymptotic properties of minimizers of convex functions derived in Geyer (1996). First we start with a definition using the notation in that paper.

Definition 24. A sequence t_n is called an approximate minimizing sequence for a sequence of functions g_n if for some sequences $\nu_n \searrow 0$ and $r_n \searrow -\infty$

$$g_n(t_n) \leq \begin{cases} \inf g_n + \nu_n & \text{if } \inf g_n > -\infty \\ r_n & \text{if } \inf g_n = -\infty. \end{cases}$$

Lemma 25. Suppose g_n is a sequence of random convex functions on $\mathbb{R}^d$ and g is another such function. Let D be a countable dense set in $\mathbb{R}^d$ and denote t_n an approximate minimizing sequence of g_n. If for each finite subset $\{s_1, \ldots, s_k\}$ of D, the random vector $(g_n(s_1), \ldots, g_n(s_k))$ converges in law to the random vector $(g(s_1), \ldots, g(s_k))$, and if with probability one g has a unique minimizer t, then t_n converges in law to t and $g_n(t_n)$ converges in law to $g(t)$.

The lemma above mainly bases on Theorem 3.2 in Geyer (1996), which deals with extended-real-valued functions that are finite on some nonempty open set. However, we will only use a version for finite functions as the extension does not yield any advantages for our application. Due to Lemma 3.1 in Geyer (1996) it suffices to consider the convergence of g_n on a countable dense subset D of $\mathbb{R}^d$. As mentioned in Chapter 3 of that paper, this condition reduces to the convergence of finite-dimensional distributions, which results in the condition of Lemma 25. Altogether Lemma 25 is a slight modification for finite functions of a statement in Chapter 1 of Geyer (1996).

A.3 Auxiliary results

First we need the following lemma, also known as Polya's Theorem.

Lemma 26. Let $(F_n)_{n\in\mathbb{N}}$ as well as F be cumulative distribution functions on $\mathbb{R}$. Furthermore, let $(F_n)_{n\in\mathbb{N}}$ converge pointwise to F. If F is continuous, then the sequence $(F_n)_{n\in\mathbb{N}}$ converges

uniformly to F. For every convergent sequence $(x_n)_{n\in\mathbb{N}}$ with limit $x \in \overline{\mathbb{R}}$ it therefore holds

$$\lim_{n\to\infty} F_n(x_n) = F(x).$$

In all cases, where either $|a|$ or $|b|$ is finite, the following lemma directly follows from Slutsky's Lemma. However, the consideration of the case $\min(|a|, |b|) = \infty$ makes it practicable for our use.

Lemma 27. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables in $\mathbb{R}$, converging in distribution to a continuous random variable X in $\mathbb{R}$. Furthermore, let $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$ be sequences of real numbers with limits $a \in \overline{\mathbb{R}}$ and $b \in \overline{\mathbb{R}}$, respectively. If $a \neq 0$, suppose the limit $c \in \overline{\mathbb{R}}$ of $\frac{b_n}{a_n}$ exists. If $\max(|a|, |b|) = \infty$ holds true, then it follows that

$$\frac{1}{|a_n X_n + b_n|} \xrightarrow[n\to\infty]{p} 0. \tag{9}$$

Proof The case $a = 0$ implies $|b| = \infty$ and therfore (9) is fulfilled. Otherwise, take an arbitrary $\epsilon > 0$. Then the following equalities hold

$$\mathbb{P}\left(\frac{1}{|a_n X_n + b_n|} \geq \epsilon\right) = \mathbb{P}(|a_n X_n + b_n| \leq \frac{1}{\epsilon}) =$$
$$\mathbb{P}(a_n X_n \leq \frac{1}{\epsilon} - b_n) - \mathbb{P}(a_n X_n < -\frac{1}{\epsilon} - b_n).$$

Since $a \neq 0$, the sign of a_n equals $\operatorname{sgn}(a)$ for sufficiently large n. Thus, the term above is eventually upper bounded by

$$\operatorname{sign}(a)\left(\mathbb{P}(X_n \leq \frac{\frac{2}{\epsilon} - b_n}{a_n}) - \mathbb{P}(X_n \leq \frac{-\frac{2}{\epsilon} - b_n}{a_n})\right) = \operatorname{sign}(a)\left(F_n\left(\frac{\frac{2}{\epsilon} - b_n}{a_n}\right) - F_n\left(\frac{-\frac{2}{\epsilon} - b_n}{a_n}\right)\right). \tag{10}$$

Denote $y_n = \frac{\frac{2}{\epsilon} - b_n}{a_n}, z_n = \frac{-\frac{2}{\epsilon} - b_n}{a_n}$. In the case where $0 < |a| < \infty$, $|b| = \infty$ and therefore both sequences converge either to ∞ or $-\infty$. If $|a| = \infty$, then the distance between the sequences vanishes asymptotically and therefore they converge to the same limit $-c$ as well. Thus, according to Lemma 26, the expression in equation (10) converges to 0.

❑

Remark. The condition on the continuity of X is – in the case where $\min(|a|, |b|) = \infty$ – a crucial one. To see this, define $X_n = -\frac{b_n}{a_n} + \frac{Z}{a_n}$, with Z being bounded in probability. Then $\min(|a|, |b|) = \infty$ implies $\frac{1}{|X_n a_n + b_n|} \xrightarrow[n\to\infty]{p} \frac{1}{|Z|}$. However, $X_n \xrightarrow[n\to\infty]{p} -c$, which is not a continuous random variable.

References

Breiman, Leo (1995). "Better Subset Regression Using the Nonnegative Garrote". In: *Technometrics* 37.4, pp. 373–384.

Donoho, David L and Iain M Johnstone (1994). "Minimax Risk over ℓ_p-Balls for ℓ_q-Error". In: *Probability Theory and Related Fields* 99.2, pp. 277–303.

Efron, Bradley et al. (2004). "Least Angle Regression". In: *The Annals of statistics* 32.2, pp. 407–499.

Ewald, Karl and Ulrike Schneider (2015). "Confidence Sets Based on the Lasso Estimator". In: *arXiv preprint arXiv:1507.05315*.

Fan, Jianqing and Runze Li (2001). "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties". In: *Journal of the American statistical Association* 96.456, pp. 1348–1360.

Frank, Ildiko E and Jerome H Friedman (1993). "A Statistical View of Some Chemometrics Regression Tools". In: *Technometrics* 35.2, pp. 109–135.

Geyer, Charles J (1996). "On the Asymptotics of Convex Stochastic Optimization". In: *Unpublished manuscript* 37.

Knight, Keith and Wenjiang Fu (2000). "Asymptotics for Lasso-Type Estimators". In: *Annals of statistics* 28, pp. 1356–1378.

Leeb, Hannes and Benedikt M Pötscher (2005). "Model Selection and Inference: Facts and Fiction". In: *Econometric Theory* 21.1, pp. 21–59.

— (2008). "Sparse Estimators and the Oracle Property, or the Return of Hodges′ Estimator". In: *Journal of Econometrics* 142.1, pp. 201–211.

Pötscher, Benedikt M (2009). "Confidence Sets Based on Sparse Estimators Are Necessarily Large". In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 1–18.

Pötscher, Benedikt M and Hannes Leeb (2009). "On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and thresholding". In: *Journal of Multivariate Analysis* 100.9, pp. 2065–2082.

Pötscher, Benedikt M and Ulrike Schneider (2009). "On the Distribution of the Adaptive LASSO Estimator". In: *Journal of Statistical Planning and Inference* 139.8, pp. 2775–2790.

— (2010). "Confidence Sets Based on Penalized Maximum Likelihood Estimators in Gaussian Regression". In: *Electronic Journal of Statistics* 4, pp. 334–360.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B* 58, pp. 267–288.

Tibshirani, Ryan J et al. (2013). "The Lasso Problem and Uniqueness". In: *Electronic Journal of Statistics* 7, pp. 1456–1490.

Tibshirani, Ryan J and Larry Wasserman (2015). *Sparsity and the Lasso*. URL: `www.stat.cmu.edu/~larry/=sml/sparsity.pdf`. Last visited on 2018/05/21.

Zou, Hui (2006). "The Adaptive Lasso and Its Oracle Properties". In: *Journal of the American statistical association* 101.476, pp. 1418–1429.

List of Figures