UNIFORM ASYMPTOTICS AND CONFIDENCE REGIONS BASED ON THE ADAPTIVE LASSO WITH PARTIALLY CONSISTENT TUNING

NICOLAI AMANN
University of Vienna

ULRIKE SCHNEIDER
TU Wien

We consider the adaptive Lasso estimator with componentwise tuning in the framework of a low-dimensional linear regression model. In our setting, at least one of the components is penalized at the rate of consistent model selection and certain components may not be penalized at all. We perform a detailed study of the consistency properties and the asymptotic distribution which includes the effects of componentwise tuning within a so-called moving-parameter framework. These results enable us to explicitly provide a set $\mathcal{M}$ such that every open superset acts as a confidence set with uniform asymptotic coverage equal to 1, whereas removing an arbitrarily small open set along the boundary yields a confidence set with uniform asymptotic coverage equal to 0. The shape of the set $\mathcal{M}$ depends on the regressor matrix as well as the deviations within the componentwise tuning parameters. Our findings can be viewed as a broad generalization of Pötscher and Schneider (2009, *Journal of Statistical Planning and Inference* 139, 2775–2790; 2010, *Electronic Journal of Statistics* 4, 334–360), who considered distributional properties and confidence intervals based on components of the adaptive Lasso estimator for the case of orthogonal regressors.

1. INTRODUCTION

The least absolute shrinkage and selection operator or Lasso by Tibshirani (1996) has received tremendous attention in the statistics literature in the past two decades. The main attraction of this method lies in its ability to perform model selection and parameter estimation at very low computational cost, and the fact that the estimator can be used in high-dimensional settings where the number of variables p exceeds the number of observations n ("$p \gg n$").

For these reasons, the Lasso has also turned into a very popular and powerful tool in econometrics, and similar things can be said about the estimator's many variants, among them the adaptive Lasso estimator of Zou (2006), where the l_1-penalty term is randomly weighted according to some preliminary estimator.

Address correspondence to Ulrike Schneider, Institute of Statistics and Mathematical Methods in Economics, TU Wien, Wiedner Hauptstraße 8/E105-2, A-1040 Vienna, Austria; e-mail: ulrike.schneider@tuwien.ac.at.

1097

This particular method has been used in econometrics in the context of diffusion processes (DeGregorio and Iacus, 2012), for instrumental variables (Caner and Fan, 2015), in the framework of stationary and nonstationary autoregressions (Kock and Callot, 2015; Kock, 2016), and for autoregressive distributed lag models (Medeiros and Mendes, 2017), to name just a few.

Despite the popularity of this method, there are still many open questions on how to construct valid confidence regions in connection with the adaptive Lasso estimator. Pötscher and Schneider (2010) demonstrate that the oracle property from Zou (2006) and Huang, Ma, and Zhang (2008) cannot be used to conduct valid inference and that resampling techniques also fail. They give confidence intervals with exact coverage in finite samples as well as an extensive asymptotic study in the framework of orthogonal regressors. However, settings more general than the orthogonal case have not been considered yet.

In this paper, we consider an arbitrary low-dimensional linear regression model ("$p \leq n$") where the regressor matrix exhibits full column rank. We allow for the adaptive Lasso estimator to be tuned componentwise with some tuning parameters possibly being equal to zero, so that not all coordinates have to be penalized. Due to this componentwise structure, three possible asymptotic regimes arise: the one where each zero component is identified as such with asymptotic probability less than one, usually termed *conservative model selection*, the one where each zero component is revealed as zero with asymptotic probability equal to one, usually referred to as *consistent model selection*, as well as the mixed case where some components are tuned conservatively and some are tuned consistently. The framework we consider encompasses the latter two regimes.

The main challenge for inference in connection with the adaptive Lasso and related estimators lies in the fact that the finite-sample distribution depends on the unknown parameter in a complicated manner, and that this dependence persists in large samples. Consequently, the coverage probability of a confidence region varies over the parameter space, and in order to conduct valid inference, one needs to guard against the lowest possible coverage and consider the minimal one. This is done so in the present paper.

Since explicit expressions for the finite-sample distribution and therefore also the coverage probabilities of confidence regions are unknown when the regressors are not orthogonal, our study is set in an asymptotic framework. We determine the appropriate uniform rate of convergence and derive the asymptotic distribution of an appropriately scaled estimator that has been centered at the true parameter. While the limit distribution is still only implicitly defined through a minimization problem, the key observation and finding is that one may explicitly characterize the set of minimizers once the union over all true parameters is taken. This is done by heavily exploiting the structure of the corresponding optimization problem and leads to a compact set $\mathcal{M}$ that is determined by the asymptotic Gram matrix as well as the asymptotic deviations between the componentwise tuning parameters and the maximal one. Subsequently, this result can be used to show how the set $\mathcal{M}$ acts as a benchmark for confidence regions, since, very loosely put, any larger set will

necessarily have asymptotic coverage equal to one, and any smaller set will exhibit zero uniform coverage in the limit. We will formalize this statement, sharpen it for a wide class of tuning regimes, and demonstrate the intrinsic problem that arises for inference in connection with this estimation method.

In this article, we show that the one-dimensional case from Pötscher and Schneider (2010) can indeed be generalized to arbitrary low-dimensional models. Our investigations reveal the geometry of confidence regions based on the adaptive Lasso, which cannot be seen in the intervals of the one-dimensional setting. Our study also encompasses the effects of varying tuning schemes over different components of the parameter, which may result in a loss of dimension in the confidence set.

The paper is organized as follows. We introduce the model and the assumptions as well as the estimator in Section 2. In Section 3, we study the relationship of the adaptive Lasso to the least-squares (LS) estimator. The consistency properties with respect to parameter estimation, rates of convergence, and model selection are derived in Section 4. Section 5 looks at the asymptotic distribution of the estimator and deduces that it is always contained in a compact set, independently of the unknown parameter. These results are used to construct and discuss the confidence regions in Section 6, where their shape is also illustrated. We summarize in Section 7 and relegate all proofs to the Appendix for readability.

2. SETTING AND NOTATION

We consider the linear regression model

$$y = X\beta + \varepsilon,$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ the nonstochastic regressor matrix assumed to have full column rank, $\beta \in \mathbb{R}^p$ the unknown parameter vector, and $\varepsilon \in \mathbb{R}^n$ the unobserved stochastic error term consisting of independent and identically distributed components with mean zero and finite second moments, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. To define the adaptive Lasso estimator, first introduced by Zou (2006), let

$$L_n(b) = \|y - Xb\|^2 + 2 \sum_{j=1}^{p} \lambda_j \frac{|b_j|}{|\hat{\beta}_{\text{LS},j}|},$$

where $\|.\|$ is the euclidean norm, λ_j are nonnegative tuning parameters, and $\hat{\beta}_{\text{LS}} = (X'X)^{-1}X'y$ is the ordinary LS estimator. We assume the event $\{\hat{\beta}_{\text{LS},j} = 0\}$ to have zero probability, for all $j = 1, \ldots, p$, and do not consider this event occurring in the subsequent analysis. The adaptive Lasso estimator we employ is given by

$$\hat{\beta}_{\text{AL}} = \underset{b \in \mathbb{R}^p}{\arg\min}\, L_n(b),$$

which always exists and is uniquely defined in our setting. Note that, in contrast to Zou (2006), we allow for *componentwise partial tuning* where the tuning

parameter may vary over coordinates and may be equal to zero, so that not all components need to be penalized. This is unlike the typical case of *uniform tuning* with a single positive tuning parameter. We also look at the leading case of $\omega_j = 1/|\hat{\beta}_{\mathrm{LS},j}|^\gamma$ with $\gamma = 1$, in the notation of Zou (2006). For all asymptotic considerations, we will assume that $X'X/n$ converges to a positive definite matrix $C \in \mathbb{R}^{p \times p}$ as $n \to \infty$.

We define the true active set $\mathcal{A}$ to be $\mathcal{A} = \{j : \beta_j \neq 0\}$. The quantity λ^* is given by the largest tuning parameter, $\lambda^* = \max_{1 \leq j \leq p} \lambda_j$. We use $\overline{\overline{\mathbb{R}}}$ for the extended real line. Finally, the symbol $\xrightarrow{d}$ stands for convergence in distribution. For the sake of readability, we suppress the dependence of the following quantities on n in the notation: y, X, ε, $\hat{\beta}_{\mathrm{AL}}$, $\hat{\beta}_{\mathrm{LS}}$, λ_j, and λ^*.

3. RELATIONSHIP TO LS ESTIMATOR

The following finite-sample relationship between the adaptive Lasso and the LS estimator is essential for proving the results in the subsequent section and will also give some insights for understanding the idea behind the results on the shape of the confidence regions in Sections 5 and 6. The lemma shows that the difference between the adaptive Lasso and the LS estimator is always contained in a bounded and closed set that depends on the regressor matrix as well as on the tuning parameters. Note that the statements in Lemma 1 and Corollary 2 hold, for all $\omega \in \Omega$, i.e., "surely."

LEMMA 1 (Relationship to LS estimator).

$$\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}} \in \{z \in \mathbb{R}^p : (X'Xz)_j = 0 \text{ for } \lambda_j = 0, z_j(X'Xz)_j \leq \lambda_j \text{ for } \lambda_j > 0, j = 1, \ldots, p\},$$

for all $\omega \in \Omega$.

Lemma 1 can be used to determine under which tuning regime the adaptive Lasso is asymptotically behaving the same as the LS estimator, as is stated in the following corollary.

COROLLARY 2 (Equivalence to LS estimator). *If $\lambda^* \to 0$, $\hat{\beta}_{\mathrm{AL}}$ and $\hat{\beta}_{\mathrm{LS}}$ are asymptotically equivalent in the sense that*

$$\sqrt{n}(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}) \to 0 \quad \text{as } n \to \infty \text{ for all } \omega \in \Omega.$$

Corollary 2 shows that in case $\lambda^* \to 0$, the adaptive Lasso estimator is asymptotically equivalent to the LS estimator, so that this case becomes a trivial one. How the estimator behaves in terms of parameter estimation and model selection for different asymptotic tuning regimes is treated in the next section.

4. CONSISTENCY IN PARAMETER ESTIMATION AND MODEL SELECTION

We start our investigation by deriving the pointwise convergence rate of the estimator.

PROPOSITION 3 (Pointwise convergence rate). *Let* $a_n = \min(\sqrt{n}, n/\lambda^*)$. *Then, the adaptive Lasso estimator is pointwise* a_n-*consistent for* β *in the sense that, for every* $\delta > 0$, *there exists a real number* $M_{\beta,\delta}$ *such that*

$$\sup_{n \in \mathbb{N}} \mathbb{P}_\beta \left(a_n \| \hat{\beta}_{\mathrm{AL}} - \beta \| > M_{\beta,\delta} \right) \le \delta.$$

The fact that the pointwise convergence rate is given by $n^{1/2}$ only if $\lambda^*/n^{1/2}$ does not diverge has implicitly been noted in Zou's (2006) oracle property in Theorem 2 in that reference, reflected in the assumption of $\lambda^*/n^{1/2} \to 0$[1]. In the one-dimensional case, it can be learned from Theorem 5(2) in Pötscher and Schneider (2009) that the sequence $n^{1/2}(\hat{\beta}_{\mathrm{AL}} - \beta)$ is not stochastically bounded if $\lambda^*/n^{1/2}$ diverges.[2] However, neither of these references determine the slower rate of n/λ^* explicitly when it applies.

The uniform convergence rate is presented in the next proposition.

PROPOSITION 4 (Uniform convergence rate). *Let* $b_n = \min(\sqrt{n}, \sqrt{n/\lambda^*})$. *Then, the adaptive Lasso estimator is uniform* b_n-*consistent for* β *in the sense that, for every* $\delta > 0$, *there exists a real number* M_δ *such that*

$$\sup_{n \in \mathbb{N}} \sup_{\beta \in \mathbb{R}^p} \mathbb{P}_\beta \left(b_n \| \hat{\beta}_{\mathrm{AL}} - \beta \| > M_\delta \right) \le \delta.$$

Proposition 4 shows that the uniform convergence rate is slower than $n^{1/2}$ if $\lambda^* \to \infty$. The fact that the uniform rate may differ from the pointwise one has been noted in Pötscher and Schneider (2009). Unless the estimator is inconsistent in parameter estimation, the uniform convergence from Proposition 4 is slower than the pointwise one and cannot, indeed, be improved upon. The latter statement is substantiated by Theorem 7 in Section 5, which shows that the limit of $b_n(\hat{\beta}_{\mathrm{AL}} - \beta_n)$ is nonzero for certain sequences β_n.

THEOREM 5 (Consistency in parameter estimation). *The following statements are equivalent.*

(a) $\hat{\beta}_{\mathrm{AL}}$ *is pointwise consistent for* β.
(b) $\hat{\beta}_{\mathrm{AL}}$ *is uniformly consistent for* β.

---

[1] Note that λ_n in that reference corresponds to $2\lambda^*$ in our notation, assuming uniform tuning over all components.

[2] To make the connection from that reference to our notation, note that $p = 1$ there and set $\theta_n = \beta$ and $n\mu_n^2 = \lambda^*$.

(c) $\lambda^*/n \to 0$ *as $n \to \infty$.*

(d) $\lim_{n \to \infty} \mathbb{P}_\beta(\hat{\beta}_{\mathrm{AL},j} = 0) = 0$ *whenever $j \in \mathcal{A}$.*

Condition (d) in Theorem 5 states that the adaptive Lasso never chooses underparameterized models with asymptotic probability equal to 1. It underlines the fact that $\lambda^*/n \to 0$ is a basic condition that we will assume in all subsequent statements.

THEOREM 6 (Consistency in model selection). *Suppose that $\lambda^*/n \to 0$ as $n \to \infty$. If $\lambda_j \to \infty$ as well as $\sqrt{n}\lambda_j/\lambda^* \to \infty$ as $n \to \infty$, for all $j = 1, \ldots, p$, then the adaptive Lasso estimator performs consistent model selection in the sense that*

$$\lim_{n \to \infty} \mathbb{P}_\beta(\hat{\beta}_{\mathrm{AL},j} \neq 0 \iff j \in \mathcal{A}) = 1 \text{ as } n \to \infty.$$

Remark. Inspecting the proof of Theorem 6 shows that in fact a more refined statement than Theorem 6 holds. Assume that $\lambda^*/n \to 0$. We then have that $\mathbb{P}_\beta(\hat{\beta}_{\mathrm{AL},j} = 0) \to 0$ whenever $j \in \mathcal{A}$ and

$$\lambda_j \to \infty \text{ and } \frac{\sqrt{n}\lambda_j}{\lambda^*} \to \infty \implies \lim_{n \to \infty} \mathbb{P}_\beta(\hat{\beta}_{\mathrm{AL},j} = 0) = 1 \text{ for } j \notin \mathcal{A} \implies \lambda_j \to \infty.$$

This statement is, in particular, interesting for the case of partial tuning where some λ_j are set to zero and the corresponding components are not penalized, revealing that the other components can still be tuned consistently in this case.

5. ASYMPTOTIC DISTRIBUTION

In this section, we investigate the asymptotic distribution. We perform our analysis for the case when $\lambda^* \to \infty$, which, by Theorem 6, encompasses the tuning regime of consistent model selection and often is the regime of choice in applications. If the estimator is tuned uniformly over all components, the condition $\lambda^* \to \infty$ is, in fact, equivalent to consistent tuning, given the basic condition of $\lambda^*/n \to 0$.

The requirement $\lambda^* \to \infty$ also corresponds to the case where the convergence rate of the adaptive Lasso estimator is given by $(\lambda^*/n)^{1/2}$ rather than $n^{1/2}$, as can be seen from Proposition 4. Pötscher and Schneider (2009, 2010) demonstrate that in order to get a representative and full picture of the behavior of the estimator from asymptotic considerations, one needs to consider a moving-parameter framework where the unknown parameter $\beta = \beta_n$ is allowed to depend on sample size. For these reasons, we study the asymptotic distribution of $(n/\lambda^*)^{1/2}(\hat{\beta}_{\mathrm{AL}} - \beta_n)$, which is done in the following.

Throughout Sections 5 and 6, let $\lambda^0 \in [0,1]^p$ and $\psi \in [0,\infty]^p$ be defined by

$$\frac{\lambda_j}{\lambda^*} \to \lambda_j^0 \in [0,1] \text{ and}$$

$$\frac{\sqrt{\lambda^*}}{\lambda_j} \to \psi_j \in [0,\infty],$$

measuring the two different deviations between each tuning parameter to the maximal one. Note that we have $\lambda^0 = (1, \ldots, 1)'$ and $\psi = 0$ for uniform tuning, and that not penalizing the jth parameter leads to $\psi_j = \infty$ and $\lambda_j^0 = 0$. Note that assuming the existence of these limits does not pose a restriction, as we could always perform our analyses on convergent subsequences and characterize the limiting behavior for all accumulation points.

THEOREM 7 (Asymptotic distribution). *Assume that $\lambda^*/n \to 0$ and $\lambda^* \to \infty$. Moreover, define $\phi \in \overline{\mathbb{R}}^p$ by $\sqrt{n}\beta_{n,j}\sqrt{\lambda^*}/\lambda_j \to \phi_j$, for $j = 1, \ldots, p$. Then,*

$$\sqrt{\frac{n}{\lambda^*}}(\hat{\beta}_{\mathrm{AL}} - \beta_n) \xrightarrow{d} \underset{u \in \mathbb{R}^p}{\arg\min}\, V_\phi(u),$$

where

$$V_\phi(u) = u'Cu + \sum_{j=1}^{p} \begin{cases} 0 & u_j = 0 \text{ or } |\phi_j| = \infty \text{ or } \psi_j = \infty \\ \infty & u_j \neq 0 \text{ and } \phi_j = \psi_j = 0 \\ 2\frac{|u_j + \lambda_j^0 \phi_j| - |\lambda_j^0 \phi_j|}{|\phi_j + \psi_j Z_j|} & else, \end{cases}$$

with $Z \sim N(0, \sigma^2 C^{-1})$, where $X'X/n \to C$, positive definite.

There are a few things worth mentioning about Theorem 7. First of all, in contrast to the one-dimensional case, the asymptotic limit of the appropriately scaled and centered estimator may still be random. However, this can only occur if ψ_j is nonzero and finite for some component j, meaning that the maximal tuning parameter diverges faster (in some sense) than the tuning parameter for the jth component, but not too much faster. When no randomness occurs in the limit, the rate of the stochastic component of the estimator is obviously smaller by an order of magnitude compared to the bias component. In particular, this will always be the case for uniform tuning when $\psi = 0$.

As is expected, the proof of Theorem 7 will be carried out by looking at the corresponding asymptotic minimization problem of the quantity of interest, which can be shown to be the minimization of V_ϕ. However, since this limiting function is not finite on an open subset of $\mathbb{R}^p$, the reasoning of why the appropriate minimizers converge in distribution to the minimizer of V_ϕ is not as straightforward as might be anticipated.

The assumption of $n^{1/2}\beta_n\lambda^{*1/2}/\lambda_j$ converging in $\overline{\mathbb{R}}^p$ in the above theorem is not restrictive in the sense that otherwise, we simply revert to converging subsequences and characterize the limiting behavior for all accumulation points, which will prove to be all we need for Proposition 8 and the confidence regions in Section 6.

While we cannot explicitly minimize V_ϕ for a fixed $\phi \in \mathbb{R}^p$ other than in trivial cases, surprisingly, we can still explicitly deduce the set of all minimizers of V_ϕ over all $\phi \in \mathbb{R}^p$, which yields the same set regardless of the realization of Z in V_ϕ. This is done in the following proposition.

PROPOSITION 8 (Set of minimizers). *Define*

$$\mathcal{M} = \mathcal{M}(\lambda^0, \psi) = \left\{ m \in \mathbb{R}^p : (Cm)_j = 0 \text{ if } \psi_j = \infty, \, m_j(Cm)_j \leq \lambda_j^0 \text{ if } \psi_j < \infty \right\}.$$

Then, for any $\omega \in \Omega$, we have

$$\mathcal{M} = \bigcup_{\phi \in \overline{\mathbb{R}}^p} \underset{u \in \mathbb{R}^p}{\arg\min} \, V_\phi(u)(\omega).$$

So, while the limit of $(n/\lambda^*)^{1/2}(\hat{\beta}_{\text{AL}} - \beta_n)$ will, in general, be random, the set $\mathcal{M}$ is not. In fact, Proposition 8 shows that, for any ω, the union of limits over all possible sequences of unknown parameters is always given by the same compact set $\mathcal{M}$. This observation is central for the construction of confidence regions in the following section. It also shows that while in general, a stochastic component will survive in the limit, it is always restricted to have bounded support that depends on the regressor matrix and the tuning parameter through the matrix C and the quantities ψ and λ^0. Interestingly, $\mathcal{M}$ only depends on ψ for the components where $\psi_j = \infty$, in which case the set $\mathcal{M}$ loses a dimension. This can be seen as a result of the jth component being penalized much less than the maximal one, so that the scaling factor used in Theorem 7 is not large enough for this component to survive in the limit. Note that in case of uniform tuning where $\psi = 0$ and $\lambda^0 = (1, \ldots, 1)'$, $\mathcal{M}$ does not depend on the sequence of tuning parameters at all. Also, we have $\mathcal{M} = [-1, 1]$, for $p = 1$ and $C = 1$, a fact that has been shown in Pötscher and Schneider (2009) and used in Pötscher and Schneider (2010).

A simple "quick-and-dirty" way to motivate the result in Proposition 8 is to rewrite

$$\sqrt{\frac{n}{\lambda^*}}(\hat{\beta}_{\text{AL}} - \beta_n) = \sqrt{\frac{n}{\lambda^*}}(\hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}}) + \sqrt{\frac{n}{\lambda^*}}(\hat{\beta}_{\text{LS}} - \beta_n)$$

and observe that the second term on the right-hand side is $o_p(1)$, whereas the first term is always contained in the set

$$\left\{ z \in \mathbb{R}^p : z_j\left(\frac{X'X}{n}z\right)_j \leq \frac{\lambda_j}{\lambda^*} \text{ for } j = 1, \ldots, p \right\}$$

by Lemma 1, which contains the set $\mathcal{M}$ in the limit. Theorem 7 and Proposition 8 can therefore be viewed as the theory that makes this observation precise by sharpening the set and showing that it only contains the limits. This can then be used for constructing confidence regions, which is done in the following section.

6. CONFIDENCE REGIONS—COVERAGE AND SHAPE

The insights from Theorem 7 and Proposition 8 can now be used for deriving the following theorem on confidence regions.

THEOREM 9 (Confidence regions). *Let $\lambda^*/n \to 0$ and $\lambda^* \to \infty$. Then, every open superset $\mathcal{O}$ of $\mathcal{M}$ satisfies*

$$\lim_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_\beta \left(\beta \in \hat{\beta}_{\mathrm{AL}} - \sqrt{\frac{\lambda^*}{n}} \mathcal{O} \right) = 1.$$

For $d > 0$, define $\mathcal{M}_d = \mathcal{M}(d\lambda^0, \psi)$. We then have that

$$\lim_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_\beta \left(\beta \in \hat{\beta}_{\mathrm{AL}} - \sqrt{\frac{\lambda^*}{n}} \mathcal{M}_d \right) = 0,$$

for any $0 < d < 1$.

Remark. The statements in Theorem 9 can be strengthened in the following way. Let $\lambda^*/n \to 0$ and $\lambda^* \to \infty$.

(a) If $\lambda^0 \in (0,1]^p$, then, for any $d > 1$, we have

$$\lim_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_\beta \left(\beta \in \hat{\beta}_{\mathrm{AL}} - \sqrt{\frac{\lambda^*}{n}} \mathcal{M}_d \right) = 1.$$

(b) If $\psi \in \{0, \infty\}^p$, then any closed and proper subset $\mathcal{C}$ of $\mathcal{M}$ fulfills

$$\lim_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_\beta \left(\beta \in \hat{\beta}_{\mathrm{AL}} - \sqrt{\frac{\lambda^*}{n}} \mathcal{C} \right) = 0.$$

Note that for uniform tuning, both refinements hold, since $\psi = 0$ and $\lambda^0 = (1, \ldots, 1)'$.

Part (a) holds since under the given assumptions, $\mathcal{M}_d$ has nonempty interior and therefore contains an open superset of $\mathcal{M}$. Part (b) hinges on the fact that the limits in Theorem 7 are always nonrandom under the given assumptions.

Casually put, Theorem 9 and the subsequent remark show the following. The set $\mathcal{M} = \mathcal{M}_1$ acts as a benchmark for confidence sets in the sense that if we take a "slightly larger" set, multiplied with the appropriate factor and centered at the adaptive Lasso estimator, we get a confidence region with minimal asymptotic coverage probability equal to 1. If, however, we base the region on a "slightly smaller" set than $\mathcal{M}$, we end up with a confidence set of asymptotic minimal coverage 0. Nothing can be revealed from the above when using $\mathcal{M}$ itself. We get into a deeper discussion in the following.

We focus on the case where $\lambda^0 \in (0,1]^p$, i.e., the case where all components of λ^0 are nonzero (implying $\psi = 0$). This means that all components are penalized at the same rate, which is obviously fulfilled for uniform tuning. In this case, the asymptotic distribution is mere point-mass with no stochastic part surviving in the limit, as can be seen from Theorem 7. The reason for this is the fact that when controlling for the bias of the estimator (by scaling with the reciprocal of the uniform convergence rate), the stochastic part vanishes asymptotically. In other words, the appropriate scaling factor is simply not large enough to keep the random

component alive in the limit, illustrating that the bias is of larger order than the stochastic component when viewed under a uniform lens[3] —a fact that is generally inherent to penalized estimators.

Given the above considerations, one might ask what happens when the confidence region is based on $\mathcal{M}_{d_n}$ where d_n may vary? The following theorem addresses this question by giving upper and lower bounds for the corresponding coverage probabilities. To state the theorem, we define the finite sample version of $\mathcal{M}_d$ as

$$\widehat{\mathcal{M}}_d = \left\{ m \in \mathbb{R}^p : (X'Xm)_j = 0 \text{ if } \lambda_j = 0, m_j \left(\frac{X'X}{n} m \right)_j \leq \frac{\lambda_j}{\lambda^*} d \text{ if } \lambda_j > 0 \right\},$$

which differs from $\mathcal{M}_d$ only in that C and λ^0 are replaced by their finite-sample equivalents $X'X/n$ and λ/λ^*, respectively, so that $\widehat{\mathcal{M}}_d$ converges to $\mathcal{M}_d$ (in the Hausdorff metric). We now provide lower and upper bounds depending on if and how d_n converges to 1 in relation to λ^*.

THEOREM 10. *Assume that $\lambda^0 \in (0,1]^p$ and let $\nu = \lim_{n\to\infty} \sqrt{\lambda^*}(d_n - 1) \in \overline{\mathbb{R}}$. We then have*

$$\limsup_{n\to\infty} \inf_{\beta\in\mathbb{R}^p} \mathbb{P}_\beta \left(\beta \in \hat{\beta}_{\mathrm{AL}} - \sqrt{\frac{\lambda^*}{n}} \widehat{\mathcal{M}}_{d_n} \right) \leq \min_{1\leq j\leq p} \Phi \left(\frac{\nu\sqrt{\lambda_j^0}}{\sigma\sqrt{3 + (C^{-1})_{jj}C_{jj}}} \right),$$

and, for $\nu > 0$,

$$\liminf_{n\to\infty} \inf_{\beta\in\mathbb{R}^p} \mathbb{P}_\beta \left(\beta \in \hat{\beta}_{\mathrm{AL}} - \sqrt{\frac{\lambda^*}{n}} \widehat{\mathcal{M}}_{d_n} \right) \geq \min_{1\leq j\leq p} F_{\chi_p^2} \left(\frac{(\lambda_j^0 \nu)^2}{4\kappa_C l_0 \sigma^2} \right),$$

where $l_0 = \sum_{j=1}^p \lambda_j^0$, and Φ and $F_{\chi_p^2}$ denote the cdf of a standard normal and a chi-square distribution with p degrees of freedom, respectively. The symbol κ_C stands for the condition number of C with respect to the spectral norm, i.e., the ratio of the largest and the smallest eigenvalue.

Remark.

(a) Theorem 10 can be shown to still hold true when $\widehat{\mathcal{M}}_{d_n}$ is replaced by its counterpart $\mathcal{M}_{d_n}$, with a slight adaptation of the constant ν involving the convergence rate of $X'X/n$ to C and λ/λ^* to λ^0.
(b) If $d_n = 1$, for all n, implying that the confidence region is based on $\widehat{\mathcal{M}}_1$, the above theorem provides 0 as lower and 1/2 as upper bound. The lower bound can, in fact, be shown to be *strict*, implying that using $\widehat{\mathcal{M}}_1$ will always yield a *positive* asymptotic coverage (bounded by 1/2) when all components of λ^0 are nonzero.

---

[3]Note that Proposition 8 shows that in *all* settings where at least one component is tuned consistently, even if a stochastic component survives in the limit, it always has bounded support contained in $\mathcal{M}$, leaving very limited possibilities for the construction of confidence regions based on the asymptotic distribution.

(c) Pötscher and Schneider (2010) prove that in the one-dimensional and Gaussian case, the upper bound of Theorem 10 is sharp: The interval $[\hat{\beta}_{\text{AL}} - (\lambda^*/n)^{1/2}d_n, \hat{\beta}_{\text{AL}} + (\lambda^*/n)^{1/2}d_n]$ possesses asymptotic infimal coverage probability of $\Phi(\nu/(2\sigma))$, which is precisely the upper bound in the above theorem.

(d) Lemma 13, on which the proof of the second statement in the above theorem is based, reveals that, for any $d > 1$, the convergence rate of the coverage probability of $\widehat{\mathcal{M}}_d$ (converging to 1) is *at least* $1/\lambda^*$.

Theorem 10 furthermore allows to illustrate the following. Assume that the confidence region $\widehat{\mathcal{M}}_{d_n}$ has asymptotic coverage strictly between 0 and 1 (implying that $d_n \to 1$). Then, this region will asymptotically not differ in volume from sets that exhibit asymptotic coverage of probability 1. In fact, it can be shown that there exists a sequence $\tilde{d}_n$ such that $\widehat{\mathcal{M}}_{\tilde{d}_n}$ has asymptotic coverage 1, satisfying

$$(\lambda^*)^q \left(\frac{\mu_p(\widehat{\mathcal{M}}_{\tilde{d}_n})}{\mu_p(\widehat{\mathcal{M}}_{d_n})} - 1 \right) \longrightarrow 0 \quad \text{for all } q < \frac{1}{2},$$

where μ_p denotes p-dimensional Lebesgue measure. This states that the ratio of volumes will tend to 1, even faster than rate $(\lambda^*)^q$, for any $q < 1/2$. It demonstrates a peculiar nature inherent to the estimation method, differing strongly from the standard approach through the LS estimator.

One might wonder now how this type of confidence region does indeed compare to the confidence ellipse based on the LS estimator. Note that the regions will be multiplied by a different factor and centered at a different estimator. In general, the following observation can be made. For $0 < \alpha < 1$, let $E_\alpha = \{z \in \mathbb{R}^p : z'Cz \leq k_\alpha\}$ with $k_\alpha > 0$ be such that $\hat{\beta}_{\text{LS}} - n^{-1/2}E_\alpha$ is an asymptotic $(1-\alpha)$-confidence region for β. If we contrast this with $\hat{\beta}_{\text{AL}} - (\frac{\lambda^*}{n})^{1/2}\mathcal{M}$, we see that since both E_α and $\mathcal{M}$ have positive, finite volume and since $\lambda^* \to \infty$, the regions based on the adaptive Lasso are always larger by an order of magnitude. This phenomenon is a special case of what has been found for any consistently tuned model selection estimator in Pötscher (2009).

Finally, we illustrate the shape of $\mathcal{M}$. We start with $p = 2$ and the matrix

$$C = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}.$$

We consider the case of uniform tuning, so that $\lambda^0 = (1,1)'$ and $\psi = (0,0)'$ and show the resulting set $\mathcal{M}$ in Figure 1. The color indicates the value of $\max_{j=1,2} m_j(Cm)_j$ at the specific point m inside the set. The higher the absolute value of the correlation of the covariates, the flatter and more stretched the confidence set becomes. As one may expect intuitively, in case of negative correlation, the confidence set covers more of the area where the signs of the covariates are equal, as can be seen in Figure 1. A positive correlation causes the opposite behavior. Note that the corners of the set $\mathcal{M}$ touch the boundary of the ellipse E_α for a certain value of k_α.
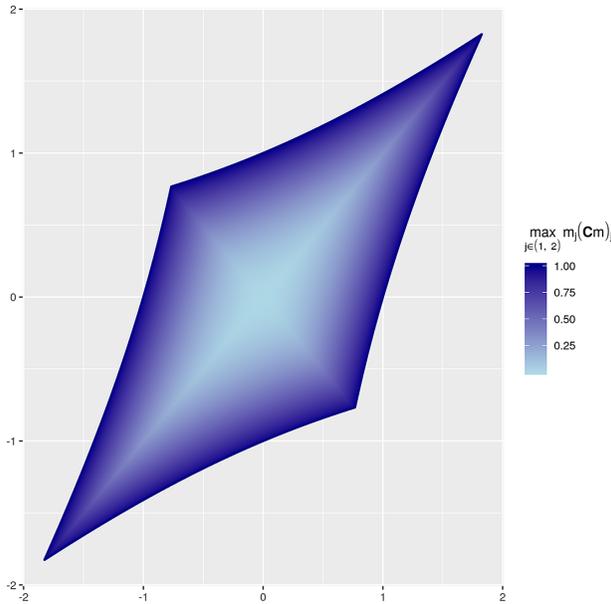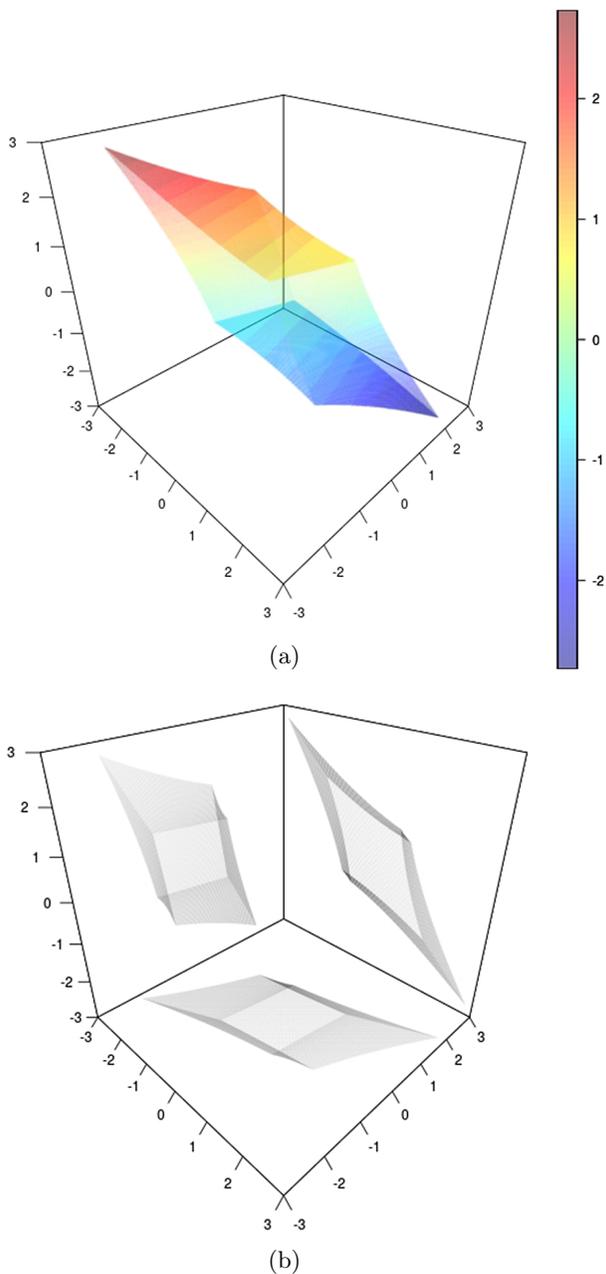
FIGURE 1. An example for the set $\mathcal{M}$ with uniform tuning in $p = 2$ dimensions.

For the case of $p = 3$, we again start with an example with uniform tuning, so that $\lambda^0 = (1, 1, 1)'$ and $\psi = (0, 0, 0)'$ and consider the matrix

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

The resulting set $\mathcal{M}$ is depicted in Figure 2. To give a better impression of the shape, the set is colored depending on the value of the third coordinate. Here, the high correlation between the first and third covariates stretches the set in the direction where the signs of the covariates differ. Figure 2(b) shows the projections of the three-dimensional set of Figure 2(a) onto three planes where one component is held fixed at a time. The projection onto the plane where the second component is held constant clearly shows the behavior 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.

Finally, Figure 3 illustrates the partially tuned case with the same matrix C. The first component is not penalized, whereas the remaining ones are tuned uniformly. This implies that $\lambda^0 = (0, 1, 1)'$ and $\psi = (\infty, 0, 0)'$. Due to the condition $(Cm)_1 = 0$, for all $m \in \mathcal{M}$, the resulting set is an intersection of a plane with the set in Figure 2(a). The fact that the confidence set is only two-dimensional might appear odd and is due to the fact that the unpenalized component exhibits a faster convergence rate,

(a)

(b)

FIGURE 2. An example for the set $\mathcal{M}$ with uniform tuning and $p = 3$ dimensions. The three-dimensional set is depicted in (a), whereas its two-dimensional projections are shown in (b).
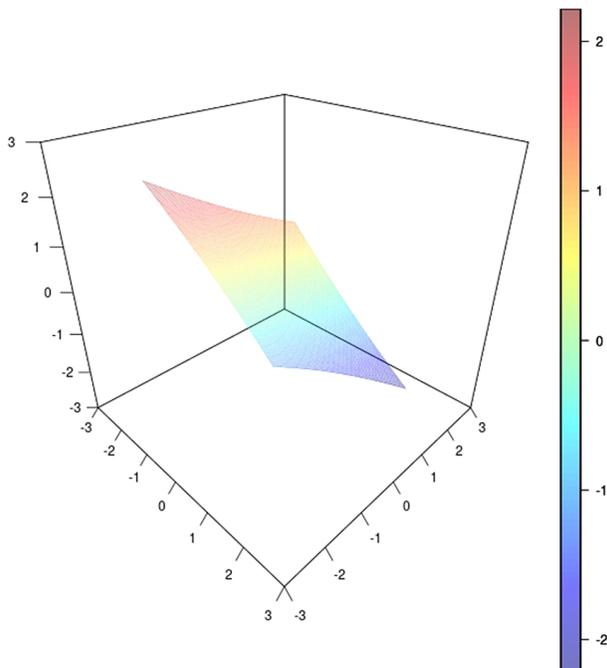
FIGURE 3. An example of the set $\mathcal{M}$ with partial tuning and $p = 3$ dimensions. The first component is not penalized resulting in the set being part of a two-dimensional subspace.

so that the factor $(\lambda^*/n)^{1/2}$ with which $\mathcal{M}$ is multiplied is not large enough for this component to survive in the limit.

7. SUMMARY AND CONCLUSIONS

We give a detailed study of the asymptotic behavior of the adaptive Lasso estimator with partially consistent and partial tuning in a low-dimensional linear regression model in terms of consistency and distributional properties. We do so within a framework that takes into account the nonuniform behavior of the estimator, nontrivially generalizing results from Pötscher and Schneider (2009) that were derived for the case of orthogonal regressors. We also demonstrate and formalize what these distributional results imply for valid confidence regions, namely that there exists a "benchmark" set $\mathcal{M}$, such that open supersets have asymptotic coverage equal to 1, whereas "slightly smaller" sets exhibit 0 uniform coverage in the limit. The reason for this phenomenon lies in the different rates of the bias component and the stochastic component of the estimator. A similar effect has been observed before for the one-dimensional case in Pötscher and Schneider (2010). We illustrate the shape of $\mathcal{M}$ and demonstrate the effect of componentwise tuning at different rates, as well as the implications of partial tuning on the confidence set.

APPENDIX

A. Proofs

We introduce the following additional notation for the proofs. The symbol e_j denotes the jth unit vector in $\mathbb{R}^p$, and the sign function is given by $\mathrm{sgn}(x) = \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x<0\}}$, for $x \in \mathbb{R}$. For a function $g : \mathbb{R}^p \to \mathbb{R}$, the one-sided directional derivative of g at u in the direction of $r \in \mathbb{R}^p$ is denoted by $\mathcal{D}_r g(u)$, given by

$$\mathcal{D}_r g(u) = \lim_{h \searrow 0} \frac{g(u+hr) - g(u)}{h}.$$

For a vector $u \in \mathbb{R}^p$ and an index set $I \subseteq \{1, \ldots, p\}$, $u_I \in \mathbb{R}^{|I|}$ contains only the components of u corresponding to indices in I. Finally, $\xrightarrow{p}$ denotes convergence in probability.

A.1. Proofs for Section 3

Proof of Lemma 1. Consider the function $G_n : \mathbb{R}^p \to \mathbb{R}$

$$u \mapsto L_n(u + \hat{\beta}_{\mathrm{LS}}) - L_n(\hat{\beta}_{\mathrm{LS}}),$$

which can, using the normal equations of the LS estimator, be rewritten to

$$u'X'Xu + 2\sum_{j=1}^p \lambda_j \frac{|u_j + \hat{\beta}_{\mathrm{LS},j}| - |\hat{\beta}_{\mathrm{LS},j}|}{|\hat{\beta}_{\mathrm{LS},j}|}.$$

Note that G_n is minimized at $\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}$ and that, since all directional derivatives have to be nonnegative at the minimizer of a convex function, after some basic calculations, we get

$$\mathcal{D}_{e_j} G_n(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}) = 2(X'X(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}))_j + 2\frac{\lambda_j}{|\hat{\beta}_{\mathrm{LS},j}|} \left(\mathbb{1}_{\{\hat{\beta}_{\mathrm{AL},j} \geq 0\}} - \mathbb{1}_{\{\hat{\beta}_{\mathrm{AL},j} < 0\}} \right) \geq 0,$$

$$\mathcal{D}_{-e_j} G_n(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}) = -2(X'X(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}))_j + 2\frac{\lambda_j}{|\hat{\beta}_{\mathrm{LS},j}|} \left(\mathbb{1}_{\{\hat{\beta}_{\mathrm{AL},j} \leq 0\}} - \mathbb{1}_{\{\hat{\beta}_{\mathrm{AL},j} > 0\}} \right) \geq 0,$$

$$\tag{1}$$

for all $j = 1, \ldots, p$. When $\hat{\beta}_{\mathrm{AL},j} = 0$, this implies that

$$|(X'X(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}))_j| \leq \frac{\lambda_j}{|\hat{\beta}_{\mathrm{LS},j}|},$$

and therefore,

$$|(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}})_j (X'X(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}))_j| \leq \lambda_j \tag{2}$$

holds. When $\hat{\beta}_{\mathrm{AL},j} \neq 0$, the equations in (1) imply

$$(X'X(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}))_j = -\lambda_j \frac{\mathrm{sgn}(\hat{\beta}_{\mathrm{AL},j})}{|\hat{\beta}_{\mathrm{LS},j}|}. \tag{3}$$

If $|\hat{\beta}_{\mathrm{AL},j} - \hat{\beta}_{\mathrm{LS},j}| \leq |\hat{\beta}_{\mathrm{LS},j}|$, clearly, (2) also holds. If $|\hat{\beta}_{\mathrm{AL},j} - \hat{\beta}_{\mathrm{LS},j}| > |\hat{\beta}_{\mathrm{LS},j}|$, we have $\mathrm{sgn}(\hat{\beta}_{\mathrm{AL},j} - \hat{\beta}_{\mathrm{LS},j}) = \mathrm{sgn}(\hat{\beta}_{\mathrm{AL},j}) \neq 0$ yielding

$$(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}})_j (X'X(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}))_j = -\lambda_j \frac{|\hat{\beta}_{\mathrm{AL},j} - \hat{\beta}_{\mathrm{LS},j}|}{|\hat{\beta}_{\mathrm{LS},j}|} \leq 0.$$

In any case, $\lambda_j = 0$ implies $(X'X(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}))_j = 0$, which completes the proof. $\qquad \square$

Proof of Corollary 2. By Lemma 1, we have

$$0 \leq \sqrt{n}(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}})' \frac{X'X}{n} \sqrt{n}(\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}) \leq \sum_{j=1}^{p} \lambda_j \leq p\lambda^* \to 0.$$

Since $X'X/n \to C$ with C being positive definite, the claim follows. $\qquad \square$

A.2. Proofs for Section 4

Proof of Proposition 3. Consider the function $H_{n,\beta} : \mathbb{R}^p \to \mathbb{R}$ defined by $H_{n,\beta}(u) = a_n^2(L_n(u/a_n + \beta) - L_n(\beta))/n$, which can be written as

$$H_{n,\beta}(u) = u' \frac{X'X}{n} u - \frac{2a_n}{n} u'X'\varepsilon + 2\sum_{j=1}^{p} \lambda_j \frac{a_n^2}{n|\hat{\beta}_{\mathrm{LS},j}|} \left(|\frac{u_j}{a_n} + \beta_j| - |\beta_j| \right).$$

$H_{n,\beta}$ is minimized at $a_n(\hat{\beta}_{\mathrm{AL}} - \beta)$ and, since $H_{n,\beta}(0) = 0$, we have $H_{n,\beta}(a_n(\hat{\beta}_{\mathrm{AL}} - \beta)) \leq 0$, which implies that

$$a_n(\hat{\beta}_{\mathrm{AL}} - \beta)' \frac{X'X}{n} a_n(\hat{\beta}_{\mathrm{AL}} - \beta) \leq \frac{a_n}{\sqrt{n}} a_n(\hat{\beta}_{\mathrm{AL}} - \beta)' \frac{2}{\sqrt{n}} X'\varepsilon + 2\sum_{j \in \mathcal{A}} \frac{1}{|\hat{\beta}_{\mathrm{LS},j}|} \frac{a_n \lambda_j}{n} |a_n(\hat{\beta}_{\mathrm{AL}} - \beta)_j|,$$

where in the latter sum we have dropped the nonpositive terms for $j \notin \mathcal{A}$ and have used the fact that $|\beta_j| - |u_j/a_n + \beta_j| \leq |u_j/a_n|$ on the terms for $j \in \mathcal{A}$. Now, note that both $a_n/\sqrt{n}$ and $a_n \lambda_j/n$ are bounded by 1 and that the sequences $X'\varepsilon/\sqrt{n}$ and $1/\hat{\beta}_{\mathrm{LS},j}$, for $j \in \mathcal{A}$, are tight, so that we can bound the right-hand side of the above inequality by a term that is stochastically bounded times $\|a_n(\hat{\beta}_{\mathrm{AL}} - \beta)\|$. Moreover, since $X'X/n$ converges to C and all matrices are positive definite, we can bound the left-hand side of the above inequality from below by a positive constant times $\|a_n(\hat{\beta}_{\mathrm{AL}} - \beta)\|^2$, so that we can arrive at

$$\|a_n(\hat{\beta}_{\mathrm{AL}} - \beta)\|^2 \leq O_p(1)\|a_n(\hat{\beta}_{\mathrm{AL}} - \beta)\|,$$

which proves the claim. $\qquad \square$

Proof of Proposition 4. Let $L > 0$ denote the infimum of all eigenvalues of $X'X/n$ and C taken over n and note that $b_n^2 \lambda^*/n \leq 1$. By Lemma 1, we have

$$b_n^2 \|\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}\|^2 \leq \frac{b_n^2}{L} (\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}})' \frac{X'X}{n} (\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}) \leq p \frac{b_n^2}{L} \frac{\lambda^*}{n} \leq \frac{p}{L}.$$

For any $M \geq 2\sqrt{\frac{p}{L}}$, we therefore have

$$\mathbb{P}_\beta(b_n\|\hat{\beta}_{\mathrm{AL}} - \beta\| > M) \leq \mathbb{P}_\beta(b_n\|\hat{\beta}_{\mathrm{AL}} - \hat{\beta}_{\mathrm{LS}}\| > M/2) + \mathbb{P}(b_n\|\hat{\beta}_{\mathrm{LS}} - \beta\| > M/2)$$

$$= \mathbb{P}(b_n \|\hat{\beta}_{\text{LS}} - \beta\| > M/2).$$

The claim now follows from the uniform $\sqrt{n}$-consistency of the LS estimator. $\qquad\square$

Proof of Theorem 5. We have (c) $\implies$ (b) by Proposition 4 and, clearly, (b) $\implies$ (a) holds. To show (a) $\implies$ (c), assume that $\hat{\beta}_{\text{AL}}$ is consistent for β and that $\lambda_j/n_k \to c \in (0, \infty]$, for some j, along a subsequence n_k. Let $\beta_j \neq 0$. On the event $\hat{\beta}_{\text{AL},j} \neq 0$, which by consistency has asymptotic probability equal to 1, we have

$$\left| \left(\frac{X'X}{n_k} (\hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}}) \right)_j \right| = \frac{\lambda_j}{n_k |\hat{\beta}_{\text{LS},j}|}$$

by equation (3). By consistency and the convergence of $X'X/n$, the left-hand side converges to zero in probability, whereas the right-hand side converges to $c/|\beta_j| > 0$ in probability along the subsequence n_k, yielding a contradiction. This shows the equivalence of the first three statements.

Moreover, (a) $\implies$ (d) since, for $j \in \mathcal{A}$,

$$\mathbb{P}_\beta(\hat{\beta}_{\text{AL},j} = 0) \leq \mathbb{P}_\beta(|\hat{\beta}_{\text{AL},j} - \beta_j| > |\beta_j|/2) \to 0$$

by consistency in parameter estimation.

The final implication we show is (d) $\implies$ (c). For this, assume that $\lambda^*/n \not\to 0$, so that there exists a subsequence n_k such that $\lambda_j/n_k \to c > 0$ as $n_k \to \infty$ for some j. We first look at the case of $c = \infty$. Note that $\hat{\beta}_{\text{AL}}$ is stochastically bounded, since $L_n(\hat{\beta}_{\text{AL}}) \leq L_n(0) = \|y\|^2$ implies

$$\hat{\beta}'_{\text{AL}} \frac{X'X}{n} \hat{\beta}_{\text{AL}} \leq \hat{\beta}'_{\text{AL}} \frac{X'X}{n} \hat{\beta}_{\text{AL}} + 2 \sum_{j=1}^{p} \lambda_j \frac{|\hat{\beta}_{\text{AL},j}|}{|\hat{\beta}_{\text{LS},j}|} \leq \hat{\beta}'_{\text{AL}} \frac{2}{n} X'y.$$

As $X'X/n \to C$ and $X'y/n \to X'X\beta$, the quadratic term on the left-hand side dominates the linear term on the right-hand side, which is only possible if $\hat{\beta}_{\text{AL}}$ is $O_p(1)$. Now, note that by equation (3), $\hat{\beta}_{\text{AL},j} \neq 0$ implies

$$\left| \left(\frac{X'X}{n_k} (\hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}}) \right)_j \right| = \frac{\lambda_j}{n_k} \frac{1}{|\hat{\beta}_{\text{LS},j}|}.$$

The fact that $X'X/n_k \to C$ and that $\hat{\beta}_{\text{AL}}$ and $\hat{\beta}_{\text{LS}}$ are stochastically bounded for fixed β shows that the left-hand side of the above display is bounded in probability also. The right-hand side, however, diverges to ∞ regardless of the value of β_j. We therefore have $\mathbb{P}_\beta(\hat{\beta}_{\text{AL},j} = 0) \to 1$, for all $\beta_j \in \mathbb{R}$, which is a contradiction to (d). If $c < \infty$, we first observe that $X'X/n(\hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}})$ is always contained in a compact set by Lemma 1 and the convergence of $X'X/n$ to C. This implies that $\|X'X/n(\hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}})\|_\infty \leq L < \infty$, for some $L > 0$ and for all β. Again, by equation (3),

$$\left| \left(\frac{X'X}{n_k} (\hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}}) \right)_j \right| = \frac{\lambda_j}{n_k} \frac{1}{|\hat{\beta}_{\text{LS},j}|},$$

whenever $\hat{\beta}_{\text{AL},j} \neq 0$. The left-hand side is bounded by L, whereas the right-hand side converges to $c/|\beta_j|$ in probability. We therefore get $\mathbb{P}_\beta(\hat{\beta}_{\text{AL},j} = 0) \to 1$, for all $\beta_j \in \mathbb{R}$, satisfying $|\beta_j| < c/L$, also yielding a contraction to (d). $\qquad\square$

Proof of Theorem 6. Since the condition $\lambda^*/n \to 0$ guards against false negatives asymptotically by Theorem 5, we only need to show that the estimator detects all zero coefficients with asymptotic probability equal to one. Assume that $\beta_j = 0$ and that $\hat{\beta}_{\text{AL},j} \neq 0$. The partial derivative of L_n with respect to $b_j \neq 0$ is given by

$$\frac{\partial L_n}{\partial b_j} = 2(X'Xb)_j - 2(X'y)_j + 2\frac{\lambda_j}{|\hat{\beta}_{\text{LS},j}|}\,\text{sgn}(b_j) = 2(X'X(b-\beta))_j - 2(X'\varepsilon)_j + 2\frac{\lambda_j}{|\hat{\beta}_{\text{LS},j}|}\,\text{sgn}(b_j),$$

which yields

$$\left| \left(\frac{X'X}{n}(a_n(\hat{\beta}_{\text{AL}} - \beta)) \right)_j - \frac{a_n}{\sqrt{n}}\frac{1}{\sqrt{n}}(X'\varepsilon)_j \right| = \frac{\lambda_j}{\sqrt{n}|\hat{\beta}_{\text{LS},j}|}\frac{a_n}{\sqrt{n}}.$$

Since $\hat{\beta}_{\text{AL}}$ is a_n-consistent for β, $X'X/n$ converges, $a_n/n^{1/2} \leq 1$, and $X'\varepsilon/\sqrt{n}$ is tight, the left-hand side of the above display is stochastically bounded. The behavior of the right-hand side is governed by $\lambda_j a_n/\sqrt{n}$, as $\sqrt{n}\hat{\beta}_{\text{LS},j}$ is also stochastically bounded for $\beta_j = 0$. If $a_n/\sqrt{n}$ does not converge to zero, then the right-hand side diverges because λ_j does. If $a_n/\sqrt{n} \to 0$, we have $a_n = n/\lambda^*$ eventually, so that $\lambda_j a_n/\sqrt{n} = \sqrt{n}\lambda_j/\lambda^*$, which also diverges by assumption. $\qquad\square$

A.3. Proofs for Section 5

LEMMA 11. *Assume that $\lambda^*/n \to 0$ and $\lambda^* \to \infty$. Moreover, suppose that $\psi_{n,j} = \sqrt{\lambda^*}/\lambda_j \to \psi_j \in [0,\infty]$ and $\phi_{n,j} = \sqrt{n}\beta_{n,j}\sqrt{\lambda^*}/\lambda_j \to \phi_j \in \overline{\mathbb{R}}$. Then, for any $u_j \in \mathbb{R}$, the term*

$$A_{n,\beta_n,j}(u_j) = \frac{\lambda_j}{\sqrt{n\lambda^*}}\frac{1}{|\hat{\beta}_{\text{LS},j}|}\left(|u_j + \sqrt{\frac{n}{\lambda^*}}\beta_{n,j}| - |\sqrt{\frac{n}{\lambda^*}}\beta_{n,j}| \right)$$

satisfies $A_{n,\beta_n,j}(u_j) \xrightarrow{d} A_{\phi,j}(u_j)$, where

$$A_{\phi,j}(u_j) = \begin{cases} 0 & u_j = 0 \text{ or } |\phi_j| = \infty \text{ or } \psi_j = \infty \\ \infty & u_j \neq 0 \text{ and } \phi_j = \psi_j = 0 \\ 2\frac{|u_j + \lambda_j^0\phi_j| - |\lambda_j^0\phi_j|}{|\psi_j Z_j + \phi_j|} & \text{else} \end{cases}$$

with $Z \sim N(0, \sigma^2 C^{-1})$. Moreover,

$$\sum_{j=1}^p A_{n,\beta_n,j}(u_j) \xrightarrow{d} \sum_{j=1}^p A_{\phi,j}(u_j),$$

for all $u \in \mathbb{R}^p$.

Proof of Lemma 11. Note that if $u_j = 0$, the term $A_{n,\beta_n,j}$ is clearly equal to 0, so that we assume $u_j \neq 0$ in the following. Define $\zeta_{n,j} = \sqrt{n/\lambda^*}\beta_{n,j} \to \zeta_j \in \overline{\mathbb{R}}$ and notice that $|\zeta_j| \leq |\phi_j|$, as well as $\zeta_j = \lambda_j^0\phi_j$ when $\lambda_j^0 > 0$ or $|\phi_j| < \infty$. Moreover, let $Z_n = \sqrt{n}(\hat{\beta}_{\text{LS}} - \beta_n)$, which satisfies $Z_n \xrightarrow{d} Z$ with $Z \sim N(0, \sigma^2 C^{-1})$.

We now look at the case where $|\phi_j| = \infty$. The term $|A_{n,\beta_n,j}(u_j)|$ is bounded by

$$\frac{\lambda_j}{\lambda^*} \frac{|u_j|}{|Z_{n,j}/\sqrt{\lambda^*} + \zeta_{n,j}|},$$

where $Z_{n,j}/\sqrt{\lambda^*}$ is $o_p(1)$. If $|\zeta_j| = \infty$ also, the above expression tends to zero in probability. If $0 < |\zeta_j| < \infty$, the same expression converges to $\lambda_j^0 |u_j|/|\zeta_j|$ in probability. But in this case, we necessarily have $\lambda_j^0 = 0$, so that the limit also equals zero. If $\zeta_j = 0$, rewrite the above bound to

$$\frac{|u_j|}{|\psi_{n,j}Z_{n,j} + \phi_{n,j}|},$$

which clearly converges to zero in probability when $\psi_j < \infty$. If $\psi_j = \infty$, note that the above display converges to zero in probability if and only if, for any $\delta > 0$, the expression

$$\mathbb{P}\left(\frac{1}{|\psi_{n,j}Z_{n,j} + \phi_{n,j}|} \geq \delta\right) = \mathbb{P}\left(|\psi_{n,j}Z_{n,j} + \phi_{n,j}| \leq \frac{1}{\delta}\right)$$
$$= \mathbb{P}\left(\frac{-1/\delta - \phi_{n,j}}{\psi_{n,}} \leq Z_n \leq \frac{1/\delta - \phi_{n,j}}{\psi_{n,}}\right)$$

converges to zero, which it does by Polya's Theorem.

We next turn to the case where $\psi_j = \infty$. If $|\phi_j| = \infty$ also, the limit equals zero by the above. If $|\phi_j| < \infty$, since $|A_{n,\beta_n,j}(u_j)|$ is bounded by

$$\frac{|u_j|}{|\psi_{n,j}Z_{n,j} + \phi_{n,j}|},$$

it will converge to zero in probability.

Let us now consider the case where $\phi_j = \psi_j = 0$. We write $A_{n,\beta_n,j}(u_j)$ as

$$\frac{|u_j + \zeta_{n,j}| - |\zeta_{n,j}|}{|\psi_{n,j}Z_{n,j} + \phi_{n,j}|},$$

which clearly diverges as $u_j \neq 0$, $|\zeta_{n,j}| \leq |\phi_{n,j}| \to 0$, and the denominator tends to 0 in probability.

For the remaining cases where $u_j \neq 0$, $|\phi_j|, \psi_j < \infty$, and $\max(|\phi_j|, \psi_j) > 0$, note that $A_{n,\beta_n,j}(u_j)$ can also be written as

$$\frac{|u_j + \zeta_{n,j}| - |\zeta_{n,j}|}{|\psi_{n,j}Z_{n,j} + \phi_{n,j}|}$$

and $\zeta_{n,j} \to \zeta_j = \lambda_j^0 \phi_j$.

The joint distributional convergence of $\sum_j A_{n,\beta_n,j}(u_j)$ to $\sum_j A_{\phi,j}(u_j)$ follows trivially. $\qquad \square$

Proof of Theorem 7. Define $V_{n,\beta_n}(u) = \frac{1}{\lambda^*}\left(L_n(\sqrt{\lambda^*/n}u + \beta_n) - L_n(\beta_n)\right)$ and notice that V_{n,β_n} is minimized at $\sqrt{n/\lambda^*}(\hat{\beta}_{AL} - \beta_n)$. The function V_{n,β_n} can be shown to equal

$$V_{n,\beta_n}(u) = u'\frac{X'X}{n}u - \frac{2}{\sqrt{n\lambda^*}}u'X'\varepsilon + 2\sum_{j=1}^{p} A_{n,\beta_n,j}(u_j),$$

where $A_{n,\beta_n,j}(u_j)$ is defined in Lemma 11. Since $X'X/n \to C$, $X'\varepsilon/\sqrt{n}$ is stochastically bounded and $\lambda^* \to \infty$, invoking Lemma 11 shows that $V_{n,\beta_n}(u)$ converges in distribution to $V_\phi(u)$. We now wish to deduce the same for the corresponding minimizers m_n and m. As explained in Section 5, the limiting function V_ϕ is not finite on an open subset of $\mathbb{R}^p$ and we cannot invoke the usual theorems employed in such a context. Instead, we define a new sequence of functions whose minimizers behave similarly but whose limiting function remains finite. To this end, we let $I = \{j : \max(|\phi_j|, \psi_j) > 0\}$ and assume, without loss of generality, that $I = \{1, \ldots, \tilde{p}\}$ with $\tilde{p} \leq p$ to ease notation with indices. Now, consider $\bar{V}_{n,\beta_n} : \mathbb{R}^p \to \mathbb{R}$ defined by

$$\bar{V}_{n,\beta_n}(u) = u'\frac{X'X}{n}u - \frac{2}{\sqrt{n\lambda^*}}u'X'\varepsilon + 2\sum_{j\in I}A_{n,\beta_n,j}(u_j),$$

and let $\tilde{V}_{n,\beta_n}, \tilde{V}_\phi : \mathbb{R}^{\tilde{p}} \to \mathbb{R}$ with

$$\tilde{V}_{n,\beta_n}(\tilde{u}) = \bar{V}_{n,\beta_n}\begin{pmatrix} \tilde{u} \\ m_{n,I^c} \end{pmatrix} \quad \text{and} \quad \tilde{V}_\phi(\tilde{u}) = V_\phi\begin{pmatrix} \tilde{u} \\ 0 \end{pmatrix}.$$

We first show that $m_{n,I^c} \xrightarrow{p} 0$. Note that $V_{n,\beta_n}(m_n) \leq V_{n,\beta_n}(0) = 0$ implies that

$$m_n'\frac{X'X}{n}m_n - \frac{2}{\sqrt{n\lambda^*}}m_n'X'\varepsilon + 2\sum_{j\in I}A_{n,\beta_n,j}(m_{n,j}) \leq -2\sum_{j\notin I}A_{n,\beta_n,j}(m_{n,j}).$$

The sequence m_n is stochastically bounded by Proposition 4. But then so is the left-hand side of the above inequality by Lemma 11. The right-hand side, however, tends to $-\infty$ whenever m_{n,I^c} does not tend to zero in probability, yielding a contradiction.

Since $m_{n,I^c} \xrightarrow{p} 0$, it is straightforward to see that $\tilde{V}_{n,\beta_n}(\tilde{u}) \xrightarrow{d} \tilde{V}_\phi(\tilde{u})$ for each $\tilde{u} \in \mathbb{R}^{\tilde{p}}$ by Lemma 11. Inspired by the Convexity Lemma of Pollard (1991), it can be shown that the functions also converge uniformly on compact sets of $\mathbb{R}^{\tilde{p}}$. Since $\tilde{V}_{n,\beta_n}$ and $\tilde{V}_\phi$ are convex and finite, this means that $\tilde{V}_{n,\beta_n}$ epiconverges to $\tilde{V}_\phi$ (cf. Geyer, 1996, p. 2). Through Theorem 3.2 in that same reference, we may deduce that

$$\arg\min_{\tilde{u}\in\mathbb{R}^{\tilde{p}}} \tilde{V}_{n,\beta_n}(\tilde{u}) \xrightarrow{d} \arg\min_{\tilde{u}\in\mathbb{R}^{\tilde{p}}} \tilde{V}_\phi(\tilde{u}).$$

To piece together the missing parts for the minimizers m_n and m of $V_{n,\beta_n}(u)$ and $V_\phi(u)$, respectively, we do the following. First note that $m_{I^c} = 0$ since otherwise V_ϕ is infinite, so that we have

$$m_{n,I^c} \xrightarrow{p} m_{I^c}.$$

To finish, observe that

$$m_{n,I} = \arg\min_{\tilde{u}\in\mathbb{R}^{\tilde{p}}} \tilde{V}_{n,\beta_n}(\tilde{u}) \xrightarrow{d} \arg\min_{\tilde{u}\in\mathbb{R}^{\tilde{p}}} \tilde{V}_\phi(\tilde{u}) = m_I.$$

$\square$

PROPOSITION 12. *The point $m \in \mathbb{R}^p$ is a minimizer of V_ϕ if and only if*

$$
\begin{cases}
m_j = 0 & \phi_j = \psi_j = 0 \\
(Cm)_j = 0 & |\phi_j| = \infty \text{ or } \psi_j = \infty \\
(Cm)_j = -\frac{\text{sgn}(m_j + \lambda_j^0 \phi_j)}{|\psi_j Z_j + \phi_j|} & 0 < \max(|\phi_j|, \psi_j) < \infty \text{ and } m_j \neq -\lambda_j^0 \phi_j \\
|(Cm)_j| \leq \frac{1}{|\psi_j Z_j + \phi_j|} & 0 < \max(|\phi_j|, \psi_j) < \infty \text{ and } m_j = -\lambda_j^0 \phi_j.
\end{cases}
$$

Proof of Proposition 12. Clearly, $m_j = 0$ if $\phi_j = \psi_j = 0$ as otherwise V_ϕ is infinite. The other conditions immediately follow by noting that m is a minimizer of the convex function V_ϕ if and only if 0 is a subgradient of V_ϕ at m. $\qquad \square$

Proof of Proposition 8. "$\subseteq$": We first show that the union of minimizers is contained in the set $\mathcal{M}$. For this, let $m = \arg\min_u V_\phi(u)$, for some $\phi \in \mathbb{R}^p$. We distinguish three cases.

First, if $\phi_j = \psi_j = 0$, we have $m_j = 0$, which immediately implies $m_j(Cm)_j = 0 \leq \lambda_j^0$.

Second, if $|\phi_j| = \infty$ or $\psi_j = \infty$, Proposition 12 implies that $(Cm)_j = 0$, which also yields $m_j(Cm)_j = 0 \leq \lambda_j^0$.

Third, if $0 < \max(|\phi_j|, \psi_j) < \infty$, we consider two subcases. When $\psi_j > 0$, $\lambda_j^0 = 0$ necessarily holds. Here, if $m_j = 0$, we immediately have $m_j(Cm)_j = 0 = \lambda_j^0$. Otherwise, $m_j \neq 0$ implies

$$
m_j(Cm)_j = -\frac{|m_j|}{|\psi_j Z_j + \phi_j|} < 0 = \lambda_j^0
$$

by Proposition 12. The other subcase of $\psi_j = 0$ can be treated as follows. If $m_j = -\lambda_j^0 \phi_j$, Proposition 12 yields

$$
|(Cm)_j| \leq \frac{1}{|\phi_j|},
$$

so that

$$
m_j(Cm)_j \leq |m_j(Cm)_j| \leq \frac{|\lambda_j^0 \phi_j|}{|\phi_j|} = \lambda_j^0.
$$

If $m_j \neq -\lambda_j^0 \phi_j$, the same proposition gives

$$
(Cm)_j = -\frac{\text{sgn}(m_j + \lambda_j^0 \phi_j)}{|\phi_j|}.
$$

If $|m_j| > |\lambda_j^0 \phi_j|$, we have $\text{sgn}(m_j) = \text{sgn}(m_j + \lambda_j^0 \phi_j)$ and

$$
m_j(Cm)_j = -\frac{|m_j|}{|\phi_j|} < 0 \leq \lambda_j^0.
$$

Finally, if $|m_j| \leq |\lambda_j^0 \phi_j|$, similarly to above, we get

$$
m_j(Cm)_j \leq |m_j(Cm)_j| = \frac{|m_j|}{|\phi_j|} \leq \frac{|\lambda_j^0 \phi_j|}{|\phi_j|} = \lambda_j^0.
$$

"⊇": We now need to show that, for any $m \in \mathcal{M}$, we can construct a $\phi \in \overline{\mathbb{R}}^p$, such that $m = \arg\min_u V_\phi(u)$. To this end, we define

$$\phi_j = \begin{cases} \infty & (Cm)_j = 0 \\ -\dfrac{m_j}{\lambda_j^0} & (Cm)_j \neq 0 \text{ and } \lambda_j^0 > 0 \text{ and } |m_j(Cm)_j| \leq \lambda_j^0 \\ \dfrac{1}{(Cm)_j} - \psi_j Z_j & \text{else} \end{cases} \tag{4}$$

and show that m is a minimizer of the resulting function V_ϕ. First, note that since $m \in \mathcal{M}$, $\psi_j = \infty$ immediately implies $(Cm)_j = 0$, satisfying the second condition of Proposition 12. We therefore assume that $\psi_j < \infty$ in the following and go through the three definitions in (4).

If $(Cm)_j = 0$, then the second condition in Proposition 12 is satisfied.

When $\phi_j = -m_j/\lambda_j^0$, the condition $\lambda_j^0 > 0$ implies that $\psi_j = 0$. So when $m_j = 0$, we are in the case where $\phi_j = \psi_j = 0$ and the first condition in Proposition 12 is fulfilled. If $m_j \neq 0$, we have

$$|(Cm)_j| \leq \frac{\lambda_j^0}{|m_j|} = \frac{1}{|\phi_j|},$$

and the fourth condition in Proposition 12 is satisfied.

Finally, when $\phi_j = 1/(Cm)_j - \psi_j Z_j$ and $\lambda_j^0 > 0$, we again have $\psi_j = 0$ and therefore $\phi_j = 1/(Cm)_j$. In that case, we also have $|m_j(Cm)_j| > \lambda_j^0$, which, since $m \in \mathcal{M}$, implies that $m_j(Cm)_j < 0$, so that we have $\mathrm{sgn}((Cm)_j) = -\mathrm{sgn}(m_j)$. But this also entails $|m_j| > \lambda_j^0/|(Cm)_j| = |\lambda_j^0 \phi_j|$, so that $m_j \neq -\lambda_j^0 \phi_j$ as well as $\mathrm{sgn}(m_j) = \mathrm{sgn}(m_j + \lambda_j^0 \phi_j)$. Thus,

$$(Cm)_j = \mathrm{sgn}((Cm)_j)|(Cm)_j| = -\frac{\mathrm{sgn}(m_j)}{|\phi_j|} = -\frac{\mathrm{sgn}(m_j + \lambda_j^0 \phi_j)}{|\phi_j|},$$

and the third condition in Proposition 12 holds. Finally, if $\lambda_j^0 = 0$ here and $m_j = 0$, it is easily seen that the fourth condition of Proposition 12 is satisfied. If $m_j \neq 0$, we are again in the case where $m_j \neq -\lambda_j^0 \phi_j$. Since $m \in \mathcal{M}$, we get $m_j(Cm)_j \leq \lambda_j^0 = 0$ and $m_j \neq 0$ and $(Cm)_j \neq 0$ implies $\mathrm{sgn}((Cm)_j) = -\mathrm{sgn}(m_j)$. Therefore, similarly as above,

$$(Cm)_j = \mathrm{sgn}((Cm)_j)|(Cm)_j| = -\frac{\mathrm{sgn}(m_j)}{|\psi_j Z_j + \phi_j|}$$

holds, satisfying the third condition in Proposition 12. □

A.4. Proofs for Section 6

Proof of Theorem 9. We start by proving the first statement. Let $g_n(\beta) = P_\beta(\beta \in \hat{\beta}_{\mathrm{AL}} - \sqrt{\frac{\lambda*}{n}} \mathcal{O})$ and $c_n = \inf_{\beta \in \mathbb{R}^p} g_n(\beta)$. We have to show that $c_n \to 1$ as $n \to \infty$. Since c_n are the infima of g_n, we can choose sequences $(\tilde{\beta}_{n,k})_{k \in \mathbb{N}} \subseteq \mathbb{R}^p$ such that

$$|c_n - g_n(\tilde{\beta}_{n,k})| \leq \frac{1}{k},$$

for all $n, k \in \mathbb{N}$. Let $\beta_n = \tilde{\beta}_{n,n}$ and note that $|c_n - g_n(\beta_n)| = o(1)$ as $n \to \infty$, so that we can look at the limiting behavior of $g_n(\beta_n)$ instead. For $\sqrt{n}\beta_n \frac{\sqrt{\lambda*}}{\lambda_j} \to \phi_j \in \mathbb{R}$, by Theorem 7, the Portmanteau Theorem, and Proposition 8, we immediately get

$$1 \geq \limsup_n g_n(\beta_n) \geq \liminf_n g_n(\beta_n) = \liminf_n P_{\beta_n} \left(\sqrt{\frac{n}{\lambda*}}(\hat{\beta}_{\mathrm{AL}} - \beta_n) \in \mathcal{O} \right)$$
$$\geq P_\phi(\arg\min_u V_\phi(u) \in \mathcal{O}) \geq P_\phi(\arg\min_u V_\phi(u) \in \mathcal{M}) = 1,$$

proving that $\lim_n c_n = \lim_n g_n(\beta_n) = 1$.

To show the second statement, we define a specific point m on the boundary of $\mathcal{M}$, as well as $\phi \in \overline{\mathbb{R}}^p$ such that $m = \arg\min_u V_\phi(u)$ and $\sqrt{n/\lambda*}(\hat{\beta}_{\mathrm{AL}} - \beta) \xrightarrow{P} m$, implying that the limiting distribution is nonrandom. Hence, excluding an open set around that m of $\mathcal{M}$ will give an infimal coverage probability tending to 0. Toward this end, let $\mathcal{S} = \{j : \lambda_j^0 > 0\}$ and note that $\mathcal{S} \neq \varnothing$, so that we have $r = C^{-1}\lambda^0 \neq 0$. Moreover,

$$0 < r'Cr = \sum_{j \in \mathcal{S}} \lambda_j^0 r_j$$

implies that there is at least one positive component r_j with $j \in \mathcal{S}$. Now, define $r_0 = \max_{j \in \mathcal{S}} r_j > 0$, let $m = r_0^{-1/2} r$, and note that this m satisfies $m \in \mathcal{M} \setminus \mathcal{M}_d$, since $Cm = r_0^{-1/2}\lambda^0$ and

$$m_j(Cm)_j = \lambda_j^0 \frac{r_j}{r_0},$$

implying that $(Cm)_j = 0$, for $j \notin \mathcal{S}$, $m_j(Cm)_j \leq \lambda_j^0$, for $j \in \mathcal{S}$, and $m_j(Cm)_j = \lambda_j^0 > d\lambda_j^0$, for some $j \in \mathcal{S}$. Also, note that $\psi_j = \infty$ implies $j \notin S$. Now, let $\phi \in \overline{\mathbb{R}}^p$ with

$$\phi_j = \begin{cases} \infty & (Cm)_j = 0 \\ -\frac{m_j}{\lambda_j^0} & (Cm)_j \neq 0 \text{ and } |m_j(Cm)_j| \leq \lambda_j^0 \\ \frac{1}{(Cm)_j} & \text{else.} \end{cases}$$

According to (4) in the proof of Proposition 8, m then is the unique minimizer of the corresponding function V_ϕ. This can be seen by noting that $(Cm)_j = 0$ if and only if $\lambda_j^0 = 0$, as well as $\psi_j > 0$ implying that $\lambda_j^0 = 0$. It is crucial to observe that the function V_ϕ is nonrandom in this case and that $\mathcal{M}_d$ is closed. Now, take any sequence $(\beta_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}^p$ converging to ϕ and let $f_n(\beta) = P_\beta(\beta \in \hat{\beta}_{\mathrm{AL}} - \sqrt{\frac{\lambda*}{n}}\mathcal{M}_d)$. By Theorem 7 and the Portmanteau Theorem, we have

$$0 \leq \liminf_n \inf_{\beta \in \mathbb{R}^p} f_n(\beta) \leq \limsup_n \inf_{\beta \in \mathbb{R}^p} f_n(\beta) \leq \limsup_n P_{\beta_n} \left(\sqrt{\frac{n}{\lambda*}}(\hat{\beta}_{\mathrm{AL}} - \beta_n) \in \mathcal{M}_d \right)$$
$$\leq P_\phi(\arg\min_u V_\phi(u) \in \mathcal{M}_d) = \mathbb{1}_{\{m \in \mathcal{M}_d\}} = 0.$$

$\square$

The following lemma is the basis to prove Theorem 10. For a symmetric matrix A, we denote by κ_A the condition number of A with respect to the spectral norm, i.e., the ratio of the largest by the smallest eigenvalue of A (in absolute value).

LEMMA 13. *Let*

$$c_n = \min_{1 \leq j \leq p} \lambda_j \left(\frac{d_n - 1}{2\sqrt{\kappa_{X'X} l_n \lambda^*}} \right)$$

with $l_n = \sum_{j=1}^{p} \lambda_j / \lambda^*$. *If* $d_n \geq 1$, *we have*

$$\inf_{\beta \in \mathbb{R}^p} \mathbb{P}_\beta \left(\beta \in \hat{\beta}_{\text{AL}} - \sqrt{\frac{\lambda^*}{n}} \widehat{\mathcal{M}}_{d_n} \right) \geq \mathbb{P} \left(\sqrt{\varepsilon' X (X'X)^{-1} X' \varepsilon} \leq c_n \left(1 - \frac{c_n}{2\sqrt{l_n \lambda^*}} \right) \right).$$

Proof. Let $a = c_n (1 - c_n/(2\sqrt{l_n \lambda^*}))$. The above statement is trivial when $a < 0$. Note that by Lemma 1, $\hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}}$ is an element of $\sqrt{\lambda^*/n} \widehat{\mathcal{M}}_1$. If $a = 0$, the event on the right-hand side implies $X'\varepsilon = 0$ and therefore $\hat{\beta}_{\text{LS}} - \beta = (X'X)^{-1} X'\varepsilon = 0$. But then, we get $\hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}} = \hat{\beta}_{\text{AL}} - \beta \in \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_1$, which implies the claim since $\widehat{\mathcal{M}}_1 \subseteq \widehat{\mathcal{M}}_{d_n}$.

We now prove the statement for $a > 0$. If we can show that whenever $z'X'Xz \leq a^2$ and $m \in \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_1$, we get $z + m \in \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_{d_n}$, then the following holds:

$$\mathbb{P} \left(\varepsilon' X (X'X)^{-1} X' \varepsilon \leq a^2 \right) = \mathbb{P}_\beta \left((\hat{\beta}_{\text{LS}} - \beta)' X'X (\hat{\beta}_{\text{LS}} - \beta) \leq a^2, \hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}} \in \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_1 \right)$$

$$\leq \mathbb{P}_\beta \left(\hat{\beta}_{\text{LS}} - \beta + \hat{\beta}_{\text{AL}} - \hat{\beta}_{\text{LS}} \in \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_{d_n} \right) = \mathbb{P}_\beta \left(\beta \in \hat{\beta}_{\text{AL}} - \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_{d_n} \right),$$

for all $\beta \in \mathbb{R}^p$, which is what we have to prove. It only remains to show that $z + m \in \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_{d_n}$ whenever $z'X'Xz \leq a^2$ and $m \in \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_1$. To do so, we show that $(z + m)_j (X'X(z+m))_j \leq \lambda_j d_n$, for all j. As $a > 0$ implies $\lambda_j > 0$, for all j, this suffices to conclude $z + m \in \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_{d_n}$. Clearly, $m \in \sqrt{\lambda^*/n} \widehat{\mathcal{M}}_1$ implies $m'X'Xm \leq \sum_{j=1}^{p} \lambda_j$. We also have

$$L_n \|m\|_\infty^2 \leq L_n \|m\|_2^2 \leq m'X'Xm$$

and

$$\|X'Xm\|_\infty^2 \leq \|X'Xm\|_2^2 \leq U_n m'X'Xm,$$

where L_n and U_n are the smallest and largest eigenvalues of $X'X$, respectively. With the same argument, we get $\|z\|_\infty^2 \leq a^2/L_n$ and $\|X'Xz\|_\infty^2 \leq a^2 U_n$. Equipped with these inequalities, we conclude, for every j, that

$$(z+m)_j (X'X(z+m))_j = z_j (X'Xz)_j + m_j (X'Xm)_j + z_j (X'Xm)_j + m_j (X'Xz)_j$$

$$\leq a^2 \sqrt{U_n/L_n} + \lambda_j + 2a\sqrt{(U_n/L_n) l_n \lambda^*}$$

$$\leq \lambda_j + 2c_n (1 - c_n/(2\sqrt{l_n \lambda^*})) \sqrt{\kappa_{X'X} l_n \lambda^*} + a^2 \sqrt{\kappa_{X'X}}$$

$$\leq \lambda_j + 2c_n \sqrt{\kappa_{X'X} l_n \lambda^*} - c_n^2 \sqrt{\kappa_{X'X}} + a^2 \sqrt{\kappa_{X'X}}$$

$$\leq \lambda_j d_n + (a^2 - c_n^2) \sqrt{\kappa_{X'X}} \leq \lambda_j d_n,$$

which completes the proof. $\square$

Remark. Lemma 13 bases on a purely algebraic argument and is still valid if X and λ are stochastic (possibly depending on ε and each other) and ε follows an arbitrary distribution. The only condition needed is the regularity of $X'X$ with probability 1.

Proof of Theorem 10. We start by proving the second statement. Note that if $\nu > 0$, we have $\delta_n > 1$ eventually, allowing to apply Lemma 13. We have that

$$c_n = \min_{1 \leq j \leq p} \lambda_j \left(\frac{d_n - 1}{2\sqrt{l_n \kappa_{X'X} \lambda^*}} \right) = \min_{1 \leq j \leq p} \left(\frac{\sqrt{\lambda^*}(d_n - 1)}{2\sqrt{l_n \kappa_{X'X}}} \frac{\lambda_j}{\lambda^*} \right) \xrightarrow[n \to \infty]{} \min_{1 \leq j \leq p} \frac{\nu \lambda_j^0}{2\sqrt{l_0 \kappa_C}}.$$

Moreover, $\varepsilon' X (X'X)^{-1} X' \varepsilon / \sigma^2$ converges to a chi-square random variable with p degrees of freedom. The second claim then follows by Lemma 13 and Polya's Theorem.

The main idea to show the first claim is the following. We pick a sequence β_n close to the boundary of, but outside the set $\sqrt{\lambda^*/n} \, \widehat{\mathcal{M}}_{d_n}$. As $\sqrt{\lambda^*/n} \, \widehat{\mathcal{M}}_{d_n}$ converges to $\sqrt{\lambda^*/n} \, \widehat{\mathcal{M}}_1$, we expect the LS estimator to lie in the set $\sqrt{\lambda^*/n} \, \widehat{\mathcal{M}}_1$ with a positive probability. (This is actually the fact if and only if $\nu \in \mathbb{R}$, because then the gap between $\sqrt{\lambda^*/n} \, \widehat{\mathcal{M}}_1$ and $\sqrt{\lambda^*/n} \, \widehat{\mathcal{M}}_{d_n}$ is of order $n^{-1/2}$.) However, $\hat{\beta}_{\mathrm{LS}} \in \sqrt{\lambda^*/n} \, \widehat{\mathcal{M}}_1$ guarantees $\hat{\beta}_{\mathrm{AL}} = 0$. In that case, $\hat{\beta}_{\mathrm{AL}} - \beta_n = -\beta_n$ is located outside of $\widehat{\mathcal{M}}_{d_n}$. Hence, $\mathbb{P}_{\beta_n}(\hat{\beta}_{\mathrm{LS}} \notin \sqrt{\lambda^*/n} \, \widehat{\mathcal{M}}_1)$ gives an upper bound for the infimal coverage probability.

For an arbitrary but fixed component $1 \leq s \leq p$, we define

$$\beta_n = \sqrt{\frac{\lambda_s d_n + \delta_n}{(X'X)_{ss}}} e_s,$$

where $\delta_n > 0$ and $\lim_{n \to \infty} \delta_n = 0$. From $\beta_{n,s}(X'X\beta_n)_s = \lambda_s d_n + \delta_n$, it follows that $\beta_n \notin \sqrt{\frac{\lambda^*}{n}} \widehat{\mathcal{M}}_{d_n}$. Hence,

$$\sup_{\beta \in \mathbb{R}^p} \mathbb{P}_\beta \left(\beta \notin \hat{\beta}_{\mathrm{AL}} + \sqrt{\frac{\lambda^*}{n}} \widehat{\mathcal{M}}_{d_n} \right) \geq \mathbb{P}_{\beta_n} \left(\beta_n \notin \hat{\beta}_{\mathrm{AL}} + \sqrt{\frac{\lambda^*}{n}} \widehat{\mathcal{M}}_{d_n} \right) \geq \mathbb{P}_{\beta_n} \left(\hat{\beta}_{\mathrm{AL}} = 0 \right).$$

If $|\hat{\beta}_{\mathrm{LS},j}(X'X\hat{\beta}_{\mathrm{LS}})_j| < \lambda_j$, for all j, then $G_n(u) = L_n(u + \hat{\beta}_{\mathrm{LS}}) - L_n(\hat{\beta}_{\mathrm{LS}})$ is minimized at $u = -\hat{\beta}_{\mathrm{LS}}$, which gives $\hat{\beta}_{\mathrm{AL}} = 0$. So in order to finish the proof, we only have to show that

$$\lim_{n \to \infty} \mathbb{P}_{\beta_n} \left(|\hat{\beta}_{\mathrm{LS},j}(X'X\hat{\beta}_{\mathrm{LS}})_j| < \lambda_j \text{ for all } j \right) = \Phi \left(\frac{-\nu\sqrt{\lambda_s^0}}{\sigma\sqrt{3 + (C^{-1})_{ss}C_{ss}}} \right).$$

Since $\beta_{n,j}(X'X\beta_n)_j = 0$, for $j \neq s$, we have $\hat{\beta}_{\mathrm{LS},j}(X'X\hat{\beta}_{\mathrm{LS}})_j \sim \mathcal{O}_p(\sqrt{\lambda^*})$, implying $|\hat{\beta}_{\mathrm{LS},j}(X'X\hat{\beta}_{\mathrm{LS}})_j| < \lambda_j$ with asymptotic probability 1 as $\lambda_j^0 > 0$. So it only remains to show that

$$\lim_{n \to \infty} \mathbb{P}_{\beta_n} \left(|\hat{\beta}_{\mathrm{LS},s}(X'X\hat{\beta}_{\mathrm{LS}})_s| < \lambda_s \right) = \Phi \left(\frac{-\nu\sqrt{\lambda_s^0}}{\sigma\sqrt{3 + (C^{-1})_{ss}C_{ss}}} \right)$$

holds true. For this, we use the equality

$$\frac{1}{\sqrt{\lambda_s d_n + \delta_n}} \left(\hat{\beta}_{\mathrm{LS},s}(X'X\hat{\beta}_{\mathrm{LS}})_s - \beta_{n,s}(X'X\beta_n)_s \right)$$

$$= \frac{1}{\sqrt{\lambda_s d_n + \delta_n}} \Big(((X'X)^{-1} X' \varepsilon)_s (X' \varepsilon)_s + \beta_{n,s} (X' \varepsilon)_s$$
$$+ ((X'X)^{-1} X' \varepsilon)_s (X' X \beta_n)_s \Big) \xrightarrow{d} Z,$$

where $Z \sim \mathcal{N}(0, \sigma^2 (C_{ss} (C^{-1})_{ss} + 3))$. This implies

$$\mathbb{P}_{\beta_n} \left(\hat{\beta}_{\text{LS},s} (X' X \hat{\beta}_{\text{LS}})_s < \lambda_s \right)$$
$$= \mathbb{P}_{\beta_n} \left(\frac{1}{\sqrt{\lambda_s d_n + \delta_n}} \left(\hat{\beta}_{\text{LS},s} (X' X \hat{\beta}_{\text{LS}})_s - \beta_{n,s} (X' X \beta_n)_s \right) < \frac{\lambda_s (1 - d_n) + \delta_n}{\sqrt{\lambda_s d_n + \delta_n}} \right),$$

where the right-hand side inside the probability converges to $-\nu \sqrt{\lambda_s^0}$, even in the case where $\lim_{n \to \infty} d_n \neq 1$. Since $\lim_{n \to \infty} \mathbb{P}_{\beta_n} \left(\hat{\beta}_{\text{LS},s} (X' X \hat{\beta}_{\text{LS}})_s \leq -\lambda_s \right) = 0$ follows by $\hat{\beta}_{\text{LS},s} (X' X \hat{\beta}_{\text{LS}})_s / \lambda_s \xrightarrow{p} \lim_{n \to \infty} d_n \geq 0$, the proof is complete. $\qquad \square$

REFERENCES

Caner, M. & M. Fan (2015) Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive Lasso. *Journal of Econometrics* 187, 256–274.

DeGregorio, A. & S. Iacus (2012) Adaptive Lasso-type estimation for multivariate diffusion processes. *Econometric Theory* 28, 838–860.

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

Huang, J., S. Ma, & C.-H. Zhang (2008) Adaptive Lasso for sparse high-dimensional regression models. *Statistical Science* 18, 1603–1618.

Kock, A.B. (2016) Consistent and conservative model selection with the adaptive LASSO in stationary and nonstationary autoregressions. *Econometric Theory* 32, 243–259.

Kock, A.B. & L. Callot (2015) Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics* 186, 325–344.

Medeiros, M.C. & E.F. Mendes (2017) Adaptive LASSO estimation for ARDL models with GARCH innovations. *Econometric Reviews* 36, 622–637.

Pollard, D. (1991) Asymptotics for least absolute deviation regression estimators. *Econometric Theory* 7, 186–199.

Pötscher, B.M. (2009) Confidence sets based on sparse estimators are necessarily large. *Sankya* 71-A, 1–18.

Pötscher, B.M. & U. Schneider (2009) On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference* 139, 2775–2790.

Pötscher, B.M. & U. Schneider (2010) Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electronic Journal of Statistics* 4, 334–360.

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

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