Asymptotic Confidence Regions Based on the Adaptive Lasso with Partial Consistent Tuning

Nicolai Amann and Ulrike Schneider
Vienna University of Technology

Abstract

We construct confidence sets based on an adaptive Lasso estimator with componentwise tuning in the framework of a low-dimensional linear regression model. We consider the case where at least one of the components is penalized at the rate of consistent model selection and where certain components may not be penalized at all. We perform a detailed study of the consistency properties and the asymptotic distribution that includes the effects of componentwise tuning within a so-called moving-parameter framework. These results enable us to explicitly provide a set $M$ such that every open superset acts as a confidence set with uniform asymptotic coverage equal to 1 whereas every proper closed subset with non-empty interior is a confidence set with uniform asymptotic coverage equal to 0. The shape of the set $M$ depends on the regressor matrix as well as the deviations within the componentwise tuning parameters. Our findings can be viewed as a generalization of (Pötscher & Schneider 2010) who considered 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 $n$ ("$p \gg n$").

Due to 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. This particular method has been used in econometrics in the context of diffusion processes (DeGregorio & Iacus 2012), for instrumental variables (Caner & Fan 2015), in the framework of stationary and non-stationary autoregressions (Kock & Callot 2015, Kock 2016) and for autoregressive distributed lag (ARDL) models (Medeiros & 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 & Schneider 2010) demonstrate that the oracle property from (Zou 2006) and (Huang et al. 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 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 detected with asymptotic probability less than
one, usually termed \textit{conservative model selection}, the one where each zero component is detected with asymptotic probability equal to one, usually referred to as \textit{consistent model selection}, and 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. Because of this, 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 this paper.

Since explicit expressions for the finite-sample distribution and 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 and leads to a compact set \( M \) that is determined by the asymptotic Gram matrix as well as the asymptotic deviations between the componentwise tuning parameters and the maximal one. This result can then be used to show how any confidence region with positive asymptotic coverage needs to include \( M \). Even more so, such confidence sets will necessarily always have asymptotic coverage equal to one, showing that it is impossible to construct classical confidence regions with arbitrary coverage in this setting.

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 estimator. The consistency properties with respect to parameter estimator, 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 the confidence regions in Section 6 where their shape is also illustrated. We summarize in Section 7 and relegate all proofs to Appendix A 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 non-stochastic 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 finite second moments, defined on some probability space \((\Omega, \mathcal{F}, \mathbb{P})\). To define the adaptive Lasso estimator first introduced by \cite{Zou2006}, let

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

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

\[
\hat{\beta}_{AL} = \arg\min_{b \in \mathbb{R}^p} L_n(b)
\]

which always exists and is uniquely defined in our setting. Note that, in contrast to \cite{Zou2006}, we allow for \textit{componentwise partial tuning} where the tuning parameter may vary over coordinates.
and may be equal to zero, so at not all components need to be penalized. This is in contrast to the
typical case of uniform tuning with a single positive tuning parameter. We also look at the
leading case of \( \omega_j = 1/|\beta_{a,j}^*| \) 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 active set \( \mathcal{A} \) to be \( \mathcal{A} = \{ j : \beta_j \neq 0 \} \). The quantity \( \lambda^* \) is given by the largest
tuning parameter, \( \lambda^* = \max_{1 \leq j \leq p} \lambda_j \) and \( \mathbb{R} \) stands for the extended real line. Finally, the symbol
\( \xrightarrow{d} \) depicts convergence in probability and convergence in distribution, respectively. For the sake
of readability, we do not show the dependence of the following quantities on \( n \) in the notation: \( y, X, \varepsilon, \beta_{aL}, \beta_{aL}, \lambda_j \) and \( \lambda^* \).

### 3 Relationship to LS estimator

The following finite-sample relationship between the adaptive Lasso estimator is essential for proving
the results in the subsequent section and will also give some insights in understanding the idea
behind the results on the shape of the confidence regions in Section 5 and 6. It 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 \( \omega \in \Omega \).

**Lemma 1** (Relationship to LS estimator).

\[
\hat{\beta}_{aL} - \hat{\beta}_{aL} \in \{ z \in \mathbb{R}^p : z_j(X'Xz)_j \leq \lambda_j \text{ for all } j = 1, \ldots, p \}
\]

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

**Corollary 2** (Equivalence of LS and adaptive Lasso estimator). If \( \lambda^* \to 0 \), \( \hat{\beta}_{aL} \) and \( \hat{\beta}_{aL} \)
are asymptotically equivalent in the sense that

\[
\sqrt{n}(\hat{\beta}_{aL} - \hat{\beta}_{aL}) \to 0 \quad \text{as } n \to \infty.
\]

**Corollary 2** shows that in case \( \lambda^* \to 0 \), the adaptive Lasso estimator is asymptotically equivalent
to the LS estimator, so that this becomes a trivial case. 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 \( \beta \) in the sense that for every \( \delta > 0 \), there exists a real
number \( M \) such that

\[
\sup_{n \in \mathbb{N}} \mathbb{P}_\beta \left( a_n \| \hat{\beta}_{aL} - \beta \| > M \right) \leq \delta.
\]

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 (2006)’s oracle property in Theorem 2 in that reference, reflected in the
assumption of \( n^{1/2}/\lambda^* \to 0 \) in that reference, reflected in the
one-dimensional case, it can be learned from Theorem 5 Part 2 in Pötscher & Schneider (2009) that the sequence \( n^{1/2}(\hat{\beta}_{aL} - \beta) \) is not stochastically bounded if
\( \lambda^*/n^{1/2} \) diverges2. However, neither of these references determine the slower rate of \( n/\lambda^* \) explicitly
when it applies.

The uniform convergence rate is presented in the next proposition.

---

1 Note that \( \lambda_n \) 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 \( n\mu^2_n = \lambda^* \).
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 $\beta$ in the sense that for every $\delta > 0$, there exists a real number $M$ such that

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

Proposition 4 shows that the uniform convergence rate is slower than $n^{-1/2}$ if $\lambda^* \to \infty$, in which case it is also slower than the pointwise rate. The fact that the uniform rate may differ from the pointwise one has been noted in Pötscher & Schneider (2009).

Theorem 7 in Section 5 shows that the limit of $b_n(\hat{\beta}_{AL} - \beta_n)$ for certain sequences $\beta_n$ are non-zero, demonstrating that the uniform rate given in Proposition 4 can indeed not be improved upon.

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

(a) $\hat{\beta}_{AL}$ is pointwise consistent for $\beta$.

(b) $\hat{\beta}_{AL}$ is uniformly consistent for $\beta$.

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

(d) $\mathbb{P}_\beta(\hat{\beta}_{AL,j} = 0) \to 0$ whenever $j \in A$.

Condition (d) in Theorem 5 states that the adaptive Lasso only chooses correct and never underparametrized models with asymptotic probability equal to 1. It underlines the fact that $\lambda^*/n \to 0$ is 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

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

Remark. Inspecting the proof of Theorem 6 shows that in fact a more refined statement than Theorem 5 holds. Assume that $\lambda^*/n \to 0$. We then have that $\mathbb{P}_\beta(\hat{\beta}_{AL,j} = 0) \to 0$ whenever $j \in A$ and

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

This statement is in particular interesting for the case of partial tuning where some $\lambda_j$ 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 and subsequently construct confidence regions in Section 6. We perform our analysis for the case when $\lambda^* \to \infty$ which, by Theorem 5, 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 $\lambda^*/n \to 0$ 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 adaptive Lasso estimator is given by $(\lambda^*/n)^{-1/2}$ rather than $n^{-1/2}$, as can be seen from Proposition 4. Pötscher & Schneider (2009) and Pötscher & Schneider (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 vary over sample size. For these reasons, we study the asymptotic distribution of $(n/\lambda^*)^{1/2}(\hat{\beta}_{AL} - \beta_n)$, which is done in the following.
Throughout Section 5 and Section 6 let \( \lambda^0 \in [0,1]^p \) and \( \psi \in [0,\infty]^p \) be defined by

\[
\frac{\lambda_j}{\lambda^0_j} \to \lambda^0_j \in [0,1] \quad \text{and} \quad \frac{\sqrt{\lambda^j}}{\lambda_j} \to \psi_j \in [0,\infty],
\]

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 \( j \)-th parameter leads to \( \psi_i = \infty \). 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 \( \phi \in \mathbb{R}^p \) by \( \sqrt{n}\lambda_{n,j}\sqrt{n}/\lambda_j \to \phi_j \) for \( j = 1,\ldots,p \). Then

\[
\sqrt{\frac{n}{\lambda^*}}(\beta_{AL} - \beta_{n}) \xrightarrow{d} \arg \min_{u \in \mathbb{R}^p} V_{\phi}(u),
\]

where

\[
V_{\phi}(u) = u'C u + \sum_{j=1}^{p} \begin{cases} 0 & u_j = 0 \text{ or } |\phi_j| = \infty \text{ or } \psi_j = \infty \hfill \\
\infty & u_j \neq 0 \text{ and } \phi_j = \psi_j = 0 \\
\frac{|u_j + \lambda_{n,j}^0 \phi_j| - |\lambda_{n,j}^0 \phi_j|}{|\phi_j + \psi_j|} & \text{else},
\end{cases}
\]

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

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 \( \psi_j \) is non-zero and finite for some component \( j \), meaning that the maximal tuning parameter diverges faster (in some sense) than the tuning parameter for the \( j \)-th 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 shown to be the minimization of \( V_{\phi} \). 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 \( V_{\phi} \) is not as straightforward as might be anticipated.

The assumption of \( n^{1/2}\beta_{n}\lambda^{1/2}/\lambda_j \) converging in \( \mathbb{R}^p \) 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 \( V_{\phi} \) for a fixed \( \phi \in \mathbb{R}^p \) other than in trivial cases, surprisingly, we can still explicitly adduce the set of all minimizers of \( V_{\phi} \) over all \( \phi \in \mathbb{R}^p \), which yields the same set regardless of the realization of \( Z \) in \( V_{\phi} \). This is done in the following proposition.

**Proposition 8** (Set of minimizers). Define

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

Then for any \( \omega \in \Omega \) we have

\[
\mathcal{M} = \bigcup_{\phi \in \mathbb{R}^p} \arg \min_{u \in \mathbb{R}^p} V_{\phi}(u)(\omega).
\]
So, while the limit of \((n/\lambda)^{1/2}(\hat{\beta}_{AL} - \beta_n)\) will, in general, be random, the set \(M\) is not. In fact, Proposition 8 shows that for any \(\omega\), the union of limits over all possible sequences of unknown parameters is always given by the same compact set \(M\). 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 \(\psi\) and \(\lambda^0\). Interestingly, \(M\) only depends on \(\psi\) for the components when \(\psi_j = \infty\), in which case the set \(M\) loses a dimension. This can be seen as a result of the \(j\)-th component being penalized much less than the maximal so that the scaling factor used in Theorem 7 is not enough for this component to survive in the limit. Note that in case of uniform tuning where \(\psi = 0\) and \(\lambda = (1, \ldots, 1)'\), \(M\) does not depend on the sequence of tuning parameters at all. Also, we have \(M = [-1, 1]^{p\times1}\) for \(p = 1\) and \(C = 1\), a fact that has been shown in Pötscher & Schneider (2009) and used in Pötscher & Schneider (2010).

A very “quick-and-dirty” way to motivate the result in Proposition 8 is to rewrite

\[
\sqrt{n} \lambda^* (\hat{\beta}_{AL} - \beta_n) = \sqrt{n} \lambda^* (\hat{\beta}_{AL} - \hat{\beta}_{LS}) + \sqrt{n} \lambda^* (\hat{\beta}_{LS} - \beta_n)
\]

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 \(M\) in the limit. Theorem 7 and Proposition 8 can therefore be viewed as the theory that make 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 \(O\) of \(M\) satisfies

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

For \(d > 0\), define \(M_d = (d\lambda^0, \psi)\). We then have that

\[
\lim_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_\beta (\beta \in \hat{\beta}_{AL} - \sqrt{\frac{\lambda^*}{n}} M_d) = 0
\]

for any \(0 < d < 1\).

Theorem 9 essentially shows the following. The set \(M = M_1\) is the “boundary case” for confidence sets in the sense that if we take a “slightly larger”, 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 take a “slightly smaller” set, we end up with a confidence region of asymptotic minimal coverage 0. Nothing can be said in general about the case when we use \(M\) itself. All this entails that constructing valid confidence regions based on the Lasso is not possible in classical sense where one can go for an arbitrary prescribed coverage level, if at least one component of the estimator is tuned to perform consistent model selection. The reason for this is the fact that when controlling the bias of the estimator, the stochastic component either has the same bounded support as the bias component or completely vanishes at all, as has been pointed out in Section 5.
Remark. The statements in Theorem 7 can be strengthened in the following way. Let $\lambda^*/n \to 0$ and $\lambda^* \to \infty$.

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

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

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

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

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

Part [a] holds since then $\mathcal{M}_d$ has non-empty interior and therefore contains an open superset of $\mathcal{M}$. Part [b] hinges on the fact that the limits in Theorem 7 are always non-random under the given assumptions.

One might wonder how this type of confidence region compares 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 $\hat{\beta}_{LS} - n^{-1/2}E_\alpha$ is an asymptotic $1(\alpha)$-confidence region for $\beta$. If we contrast this with $\hat{\beta}_{AL} - (\frac{\lambda^*}{n})^{1/2}\mathcal{M}$, we see that since both $E_\alpha$ 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.

We now 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 $M$ 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 is, the flatter and more stretched the confidence set becomes. As one may expect intuitively, in the case of positive correlation, the confidence set covers more of the area where the signs of the covariates are equal. A negative correlation causes the opposite behavior seen in Figure 1. Note that the corners of the set $\mathcal{M}$ touch the boundary of the ellipse $E_\alpha$ for a certain value of $k_\alpha$.

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 covariate stretches the set in the direction where the signs of the covariates differ. Figure 2 shows the projections of the three-dimensional set of Figure 2 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 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...
Figure 1: An example for the set $M$ with uniform tuning in $p = 2$ dimensions.

is an intersection of a plane with the set in Figure 2A. 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 so that the factor $(\lambda^*/n)^{1/2}$ with which $M$ is multiplied is not enough for this component to survive in the limit.

7 Summary and conclusions

We give a detailed study the asymptotic behavior of the adaptive Lasso estimator with partial consistent tuning in a low-dimensional linear regression model. We do so within a framework that takes into account the non-uniform behavior of the estimator, non-trivially generalizing results from Pötscher & Schneider (2009) that were derived for the case of orthogonal regressors. We use these distributional results to show that valid confidence regions based on the estimator can essentially only have asymptotic coverage equal to 0 or to 1, a fact that has been observed before for the one-dimensional case in Pötscher & Schneider (2010). We illustrate the shape of these regions and demonstrate the effect of componentwise tuning at different rates as well as the implications of partial tuning on the confidence set.
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.

A Appendix – Proofs

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

$$D_r g(u) = \lim_{h \to 0} \frac{g(u + hr) - g(u)}{h}.$$ 

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

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

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

$$u'X'Xu + 2 \sum_{j=1}^{p} \lambda_j \frac{|u_j + \hat{\beta}_{\text{LS},j}| - |\hat{\beta}_{\text{LS},j}|}{|\beta_{\text{LS},j}|}.$$
Note that $G_n$ is minimized at $\hat{\beta}_{AL} - \hat{\beta}_{LS}$ and that, since all directional derivatives have to be non-negative at the minimizer of a convex function. After some basic calculations we get

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

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

(1)

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

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

and therefore

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

(2)

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

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

(3)

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

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

which completes the proof.

Proof of Corollary[26] By Lemma[14] we have

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

Since $X'X/n \rightarrow C$ with $C$ positive definite, the claim follows.

\[\square\]

A.2 Proofs for Section 4

Proof of Proposition[20] Consider the function $H_{n,\beta} : \mathbb{R}^p \rightarrow \mathbb{R}$ defined by $H_{n,\beta}(u) = a_n^2 (L_n(u/a_n + \beta) - L_n(\beta))/n$ which can be written as

$$
H_{n,\beta}(u) = u X'Xu - 2a_n X' + 2 \sum_{j=1}^{p} \lambda_j u_j \left( u_j / a_n + \beta_j \right).
$$

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

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

where in the latter sum we have dropped the non-positive terms for $j \notin A$ and have used the fact that $|\beta_j| - |u_j/a_n - \beta_j| \leq |u_j/a_n|$ on the terms for $j \in A$. Now note that both $a_n/\sqrt{n}$ and $a_n \lambda_j/n$ are bounded by 1 and that the sequences $X'X/\sqrt{n}$ and $1/|\hat{\beta}_{LS,j}|$ for $j \in A$ are tight, so that we can bound the right-hand side of the above inequality by a term that is stochastically bounded times $|a_n(\hat{\beta}_{AL} - \beta)|$. 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 $|a_n(\hat{\beta}_{AL} - \beta)|$, so that we can arrive at

$$
\sqrt{n} a_n(\hat{\beta}_{AL} - \beta) \leq O_p(1) \sqrt{n} a_n(\hat{\beta}_{AL} - \beta)
$$

which proves the claim.

\[\square\]

11
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 $b_n^2X^*/n \leq 1$. By Lemma 7 we have

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

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

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

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

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

Proof of Theorem 5. We have (c) $\Rightarrow$ (b) by Proposition 4 and clearly, (b) $\Rightarrow$ (a) holds. To show (a) $\Rightarrow$ (c), assume that $\hat{\beta}_{AL}$ is consistent for $\beta$ 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 $\hat{\beta}_{AL,j} \neq 0$, which by consistency has asymptotic probability equal to 1, we have

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

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) $\Rightarrow$ (d) since for $j \in A$

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

by consistency in parameter estimation.

The final implication we show is (d) $\Rightarrow$ (e). For this, assume that $\lambda^*/n_k \not\to 0$ 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 $\hat{\beta}_{AL}$ is stochastically bounded, since $L_n(\hat{\beta}_{AL}) \leq L_n(0) = \|y\|^2$ implies

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

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

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

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

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

whenever $\hat{\beta}_{AL,j} \neq 0$. The left-hand side is bounded by $L$ whereas the right-hand side converges to $c/|\beta_j|$ in probability. We therefore get $P_{\beta}(\hat{\beta}_{AL,j} = 0) \to 1$ for all $\beta_j \in \mathbb{R}$ satisfying $|\beta_j| < c/L$, also yielding a contraction to (d).
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 \( \hat{\beta}_{AL,j} \neq 0 \). The partial derivative of \( L_n \) with respect to \( b_j \neq 0 \) is given by
\[
\frac{\partial L_n}{\partial b_j} = 2(X'Xb_j) - 2(X'y_j) + 2\frac{\lambda_j}{|\beta_{LS,j}|} \text{sgn}(b_j) = 2(X'X(b - \beta))_j - 2(X'\varepsilon)_j + 2\frac{\lambda_j}{|\beta_{LS,j}|} \text{sgn}(b_j),
\]
which yields
\[
\left( \frac{X'X}{n}(a_n(\hat{\beta}_{AL} - \beta)) \right)_j - a_n \frac{1}{\sqrt{n}} \sqrt{n}(X'\varepsilon)_j = \frac{\lambda_j}{\sqrt{n}|\beta_{LS,j}|} a_n.
\]
Since \( \hat{\beta}_{AL} \) is \( a_n \)-consistent for \( \beta \), \( X'X/n \) 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 \( \sqrt{n}\beta_{LS,j} \) 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 \( \lambda_j \) 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.

A.3 Proofs for Section 5

Lemma 10. Assume that \( \lambda^*/n \to 0 \) and \( \lambda^* \to \infty \). Moreover, suppose that \( \psi_{n,j} = \sqrt{X}/\lambda_j \to \psi_j \in [0, \infty] \) and \( \phi_{n,j} = \sqrt{n}\beta_{n,j}/\lambda_j \to \phi_j \in \mathbb{R} \). Then for any \( u_j \in \mathbb{R} \), the term
\[
A_{n,\beta_{n,j}}(u_j) = \frac{\lambda_j}{\sqrt{n}\lambda^*} a_n \frac{1}{|\beta_{LS,j}|} \left( |u_j| + \sqrt{\frac{n}{X}}|\beta_{n,j}| - |\sqrt{\frac{n}{X}}|\beta_{n,j}| \right).
\]
satisfies \( A_{n,\beta_{n,j}}(u_j) \overset{d}{\to} A_{\phi,j}(u_j) \) where
\[
A_{\phi,j}(u_j) = \begin{cases} 
0 & u_j = 0 \text{ or } |\phi_j| = \infty \text{ or } \psi_j = \infty \\
\infty & u_j \neq 0 \text{ and } |\phi_j| = \psi_j = 0 \\
2|u_j + \lambda_0^*|\phi_j|/|\psi_j,\phi_j + \psi_j| & \text{else}
\end{cases}
\]
with \( Z \sim N(0, \sigma^2 C^{-1}) \). Moreover,
\[
\sum_{j=1}^{p} A_{n,\beta_{n,j}}(u_j) \overset{d}{\to} \sum_{j=1}^{p} A_{\phi,j}(u_j)
\]
for all \( u \in \mathbb{R}^p \).

Proof of Lemma 10. 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 \( \zeta_{n,j} = \sqrt{n}/\lambda \beta_{n,j} \to \zeta_j \in \mathbb{R} \) and notice that \( |\zeta_j| \leq |\phi_j| \), as well as \( \zeta_j = \lambda_0^* \phi_j \) when \( \lambda_0^* > 0 \) or \( |\phi_j| < \infty \). Moreover, let \( Z_n = \sqrt{n}(\hat{\beta}_{n,j} - \beta_n) \) which satisfies \( Z_n \overset{d}{\to} Z \) 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 \( \lambda_0^* |u_j|/|\zeta_j| \) in probability. But in this case, we necessarily have \( \lambda_0^* = 0 \), so that the limit also equals zero. If \( \zeta_j = 0 \), rewrite the above bound to
\[
|u_j|/|\psi_{n,j} Z_{n,j} + \phi_{n,j}|
\]
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

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

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 $\zeta_{n,j} \to \zeta = \lambda^0 \phi_j$.

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

\emph{Proof of Theorem 2} Define $V_{n,\beta_n}(u) = \frac{1}{n} \left( L_n(\sqrt{n}X u + \beta_n) - L_n(\beta_n) \right)$ and notice that $V_{n,\beta_n}$ is minimized at $\sqrt{n}/\lambda^* (\beta_{n,\lambda^*} - \beta_n)$. The function $V_{n,\beta_n}$ 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 10. Since $X'X/n \to C$, $X' \varepsilon/\sqrt{n}$ is stochastically bounded and $\lambda^* \to \infty$, invoking Lemma 10 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 2 the limiting function $V_{\phi}$ 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 $V_{n,\beta_n} : \mathbb{R}^p \to \mathbb{R}$ defined by

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

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

$$\tilde{V}_{n,\beta_n}(\tilde{u}) = \tilde{V}_{n,\beta_n} \left( m_n^{\tilde{u}}, u_{\tilde{r}} \right) \quad \text{and} \quad \tilde{V}_{\phi}(\tilde{u}) = \tilde{V}_{\phi} \left( \tilde{u}, m_{\tilde{r}} \right).$$

We first show that $m_{n,\tilde{r}} \overset{p}{\to} 0$. Note that $V_{n,\beta_n}(m_n) \leq V_{n,\beta_n}(0) = 0$ implies that

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

The sequence $m_n$ is stochastically bounded by Proposition 4. But then so is the left-hand side of the above inequality by Lemma 10. The right-hand side, however, tends to $-\infty$ whenever $m_{n,\tilde{r}}$ does not tend to zero in probability, yielding a contradiction.
Since \( m_{n,t} \xrightarrow{p} 0 \), it is straightforward to see that \( \tilde{V}_{n,\beta_n}(\tilde{u}) \xrightarrow{d} \tilde{V}_{\phi}(\tilde{u}) \) for each \( \tilde{u} \in \mathbb{R}^p \) by Lemma \[10\]. By the Convexity Lemma of Pollard (1991), this means that the functions also converge uniformly on compact sets of \( \mathbb{R}^p \). Since \( \tilde{V}_{n,\beta_n} \) and \( \tilde{V}_{\phi} \) are convex and finite, this means that \( \tilde{V}_{n,\beta_n} \) epiconverges to \( \tilde{V}_{\phi} \) (c.f. Geyer 1996, p. 2). Through Theorem 3.2 in that same reference, we may deduce that

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

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 \( m_{1^*} = 0 \) since otherwise \( V_\phi \) is infinite, so that we have

\[
m_{n,t^*} \xrightarrow{p} m_{1^*}.
\]

To finish, observe that

\[
m_{n,t} = \arg \min_{\tilde{u} \in \mathbb{R}^p} \tilde{V}_{n,\beta_n}(\tilde{u}) \xrightarrow{d} \arg \min_{\tilde{u} \in \mathbb{R}^p} \tilde{V}_{\phi}(\tilde{u}) = m_t.
\]

\[
\text{Proposition 11. The point } m \in \mathbb{R}^p \text{ is a minimizer of } V_\phi \text{ if and only if}
\]

\[
\begin{aligned}
m_j &= 0 & \phi_j &= \psi_j &= 0 \\
(Cm)_j &= 0 & |\phi_j| &= \infty \text{ or } \psi_j &= \infty \\
(Cm)_j &= -\frac{\text{sgn}(m_j + \lambda^0_j \phi_j)}{|\psi_j Z_j + \phi_j|} & 0 < \max(|\phi_j|, \psi_j) &= \infty \text{ and } m_j &\neq -\lambda^0_j \phi_j \\
|\langle Cm \rangle_j| &\leq \frac{1}{{|\psi_j Z_j + \phi_j|}} & 0 < \max(|\phi_j|, \psi_j) &= \infty \text{ and } m_j &\neq -\lambda^0_j \phi_j.
\end{aligned}
\]

\[\text{Proof of Proposition 11.}\]

Clearly, \( m_j = 0 \) if \( \phi_j = \psi_j = 0 \) as otherwise \( V_\phi \) is infinite. The other conditions immediately follow by noting that \( m \) is a minimizer of the convex function \( V_\phi \) if and only if \( 0 \) is a subgradient of \( V_\phi \) at \( m \). □

\[\text{Proof of Proposition 3.}\]

"\( \subseteq \)" : 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.

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

If secondly \( |\phi_j| = \infty \) or \( \psi_j = \infty \), Proposition 11 implies that \( (Cm)_j = 0 \) which also yields \( m_j(Cm)_j = 0 \leq \lambda^0_j \).

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

\[
m_j(Cm)_j = -\frac{|m_j|}{|\psi_j Z_j + \phi_j|} < 0 = \lambda^0_j
\]

by Proposition 11. The other subcase of \( \psi_j = 0 \) can be treated as follows. If \( m_j = -\lambda^0_j \phi_j \), Proposition 11 yields

\[
|\langle Cm \rangle_j| \leq \frac{1}{|\phi_j|}
\]

so that

\[
m_j(Cm)_j \leq |m_j(Cm)_j| \leq \frac{|\lambda^0_j \phi_j|}{|\phi_j|} = \lambda^0_j.
\]

If \( m_j \neq -\lambda^0_j \phi_j \), the same proposition gives

\[
(Cm)_j = -\frac{\text{sgn}(m_j + \lambda^0_j \phi_j)}{|\phi_j|}.
\]

15
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.

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

\[\phi_j = \begin{cases} \infty & (Cm)_j = 0 \\ \frac{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 \\ \frac{1}{(Cm)_j} - \psi_j Z_j & \text{else} \end{cases} \quad (4)

and show that } m \text{ is a minimizer of the resulting function } V_\phi. \text{First note that since } m \in \mathcal{M}, \psi_j = \infty \text{ immediately implies } (Cm)_j = 0, \text{satisfying the second condition of Proposition 11}. \text{We therefore assume that } \psi_j < \infty \text{ in the following and go through the three definitions in (4).}

\[\text{If } (Cm)_j = 0 \text{ then the second condition in Proposition 11 is satisfied.} \]

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 11 is fulfilled. If \(m_j \neq 0\), we have

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

and the fourth condition in Proposition 11 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 \(\text{sgn}((Cm)_j) = -\text{sgn}(m_j)\). 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 \(\text{sgn}(m_j) = \text{sgn}(m_j + \lambda_j^0 \phi_j)\). Thus,

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

and the third condition in Proposition 11 holds. Lastly, if \(\lambda_j^0 = 0\) here and \(m_j = 0\), it is easily seen that the fourth condition of Proposition 11 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 \(\text{sgn}((Cm)_j) = -\text{sgn}(m_j)\). Therefore, similarly as above,

\[(Cm)_j = \text{sgn}((Cm)_j)|(Cm)_j| = -\frac{\text{sgn}(m_j)}{\psi_j Z_j + \phi_j}\]

holds, satisfying the third condition in Proposition 11. 

\[\square\]

A.4 Proofs for Section 6

\textit{Proof of Theorem 1} We start by proofing the first statement. Let \(g_n(\beta) = P_g(\beta \in \hat{\beta}_{\lambda L} - \sqrt{n} \mathcal{O})\) 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 \((\hat{\beta}_{n,k})_{k \in \mathbb{N}} \subseteq \mathbb{R}^p\) such that

\[|c_n - g_n(\hat{\beta}_{n,k})| \leq \frac{1}{k}\]
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 $\sqrt{n} \beta_n \sqrt{\lambda_n} \to \phi_j \in \mathbb{R}$, by Theorem 7, the Portmanteau Theorem and Proposition 8 we immediately get

$$1 \geq \limsup_n g_n(\beta_n) \geq \liminf_n g_n(\beta_n) = \liminf_n P_{\beta_n}(\sqrt{n} \lambda^{0}_{\hat{\beta}}(\beta_{\Delta L} - \beta_n) \in \mathcal{O}) \geq P_\phi(\arg \min_u V_\phi(u) \in \mathcal{O}) \geq P_\phi(\arg \min_u V_\phi(u) \in \mathcal{M}) = 1,$$

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

To show the second statement, let $S = \{ j : \lambda_{0j}^0 > 0 \}$ and note that $S \neq \emptyset$ so that we have $r = C^{-1} \lambda^0 \neq 0$. Moreover,

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

implies that there is at least one positive component $r_j$ with $j \in S$. Now define $r_0 = \max_{j \in S} r_j > 0$, 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

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

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

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

According to (4) in the proof of Proposition 8, $m$ then is the unique minimizer of the corresponding function $V_\phi$. This can be seen by noting that $(Cm)_j = 0$ if and only if $\lambda_{0j}^0 = 0$ as well as $\psi_j > 0$ implying that $\lambda_{0j}^0 = 0$. It is crucial to observe that the function $V_\phi$ is non-random 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 $\phi$ and let $f_n(\beta) = P_\phi(\beta \in \hat{\beta}_{\Delta L} - \sqrt{\lambda_n} \mathcal{M}_d)$. By Theorem 7 and the Portmanteau Theorem we have

$$0 \leq \liminf_{\beta \in \mathbb{R}^p} f_n(\beta) \leq \limsup_{\beta \in \mathbb{R}^p} f_n(\beta) \leq \limsup_n P_{\beta_n}(\sqrt{n} \lambda^{0}_{\hat{\beta}}(\beta_{\Delta L} - \beta_n) \in \mathcal{M}_d) \leq P_\phi(\arg \min_u V_\phi(u) \in \mathcal{M}_d) = 1_{\{m \in \mathcal{M}_d\}} = 0.$$
References

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

DeGregorio, A. & Iacus, S. (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., Ma, S. & Zhang, C.-H. (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. & Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics* **186**, 325–344.

Medeiros, M. C. & Mendes, E. F. (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. & Schneider, U. (2009). On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference* **139**, 2775–2790.

Pötscher, B. M. & Schneider, U. (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.