Confidence Sets Based on the Lasso Estimator

Karl Ewald and Ulrike Schneider
Vienna University of Technology

Abstract

In a linear regression model with fixed dimension, we construct confidence sets for the unknown parameter vector based on the Lasso estimator in finite samples as well as in an asymptotic setup, thereby quantifying estimation uncertainty of this estimator. In finite samples with Gaussian errors and asymptotically in the case where the Lasso estimator is tuned to perform conservative model selection, we derive formulas for computing the minimal coverage probability over the entire parameter space for a large class of shapes for the confidence sets, thus enabling the construction of valid confidence sets based on the Lasso estimator in these settings. The choice of shape for the confidence sets and comparison with the confidence ellipse based on the least-squares estimator is also discussed. Moreover, in the case where the Lasso estimator is tuned to enable consistent model selection, we give a simple confidence set with minimal coverage probability converging to one.

1 Introduction

The Lasso estimator as introduced in [Tibshirani (1996)] as well as many variants thereof have gained strong interest in the statistics community and in applied areas over the past two decades. As is well known, the main attraction of the Lasso estimator lies in its ability to perform model selection and parameter estimation at very low computational cost, see for instance [Alliney & Ruzinski (1994), Efron et al. (2004) and Rosset & Zhu (2007)], and in the fact that the estimator can be used in high-dimensional settings where the number of variables $p$ exceeds the number of observations $n$ ($p \gg n$).

Literature on distributional properties of the Lasso estimator in the low-dimensional setting ($p \leq n$) include the often-cited paper by [Knight & Fu (2000)] who derive the asymptotic distribution when the estimator is tuned to perform conservative model selection. [Pötscher & Leeb (2009)] give a detailed analysis in the framework of a linear regression model with orthogonal design and derive the distribution of the Lasso estimator in finite samples as well as in the two asymptotic regimes of consistent and conservative tuning. Implications of these results for confidence intervals are analyzed in [Pötscher & Schneider (2010)] and generalizations to a moderate-dimensional setting where $p \leq n$ but $p$ diverging with $n$ are contained in [Pötscher & Schneider (2011) and Schneider (2015)].

In a high-dimensional setting with $p \gg n$, confidence regions and confidence intervals in connection with the Lasso estimator have recently been treated in a number of papers including [Zhang & Zhang (2014), Van de Geer et al. (2014), Javanmard & Montanari (2014), Caner & Kock (2014) and Van de Geer (2015)]. All these papers use the idea of “de-sparsifying” the Lasso estimator which in the case of $p \leq n$ essentially reduces to using the least-squares (LS) estimator for inference. In that sense this theory leaves a gap on how to construct confidence regions based on the Lasso estimator in a low-dimensional framework to provide uncertainty quantification for the Lasso estimator also in this case.

[Lee et al. (2013)] consider finite-sample results for confidence intervals in connection with the Lasso estimator yet these authors take a different route in that their intervals are not set to cover the true parameter, but a pseudo-true value that depends on the selected model and coincides with the true parameter if the selected model is correct. All inference is conditional on the selected model. Their method is in line with the general proposal of [Berk et al. (2015)] who discuss an intricate procedure for obtaining confidence regions for this pseudo-true parameter after a model selection step.

In this paper, we construct confidence sets based on the Lasso estimator for the entire unknown parameter vector. One of the challenges of this task lies in the well-known fact that the finite-sample
distribution of the Lasso estimator depends on the unknown parameter in a complicated manner. This phenomenon does not vanish for large samples as can be seen within a so-called moving-asymptotic framework (see Pötscher & Leeb (2009) for a detailed analysis in orthogonal design) and also occurs for related estimators. In order to construct valid confidence sets, we need to know the smallest coverage probability occurring over the whole parameter space. Pötscher & Schneider (2010) derive a formula for the minimal coverage probability of fixed-width confidence intervals based on the Lasso estimator in one dimension using knowledge of its finite-sample distribution. In the general case, this finite-sample distribution is not known, so it is not clear how to obtain an expression for the coverage probability in more than one dimension. Additionally, this coverage probability clearly depends on the shape that is used for the confidence set and it is a not clear a priori what this shape should be. We do the following.

While the finite sample distribution and therefore the coverage probability for any kind of set based on the Lasso estimator is unknown in general dimensions, we show that computing the minimal coverage probability can actually be carried out without this explicit knowledge. We obtain an explicit formula for the minimal coverage probability by, in a way, deferring the minimization problem into the objective function that defines the estimator, as is depicted in Section 3. For the confidence regions, we consider a large class of shapes that is determined by a condition involving the regressor matrix. This class encompasses the elliptic shape one would use if the confidence region was based on the LS estimator, thus enabling comparisons with the LS confidence ellipse. Analogously to the fixed-width intervals in Pötscher & Schneider (2010), the confidence regions we consider are random only through their centering at the Lasso estimator (which is also in line with the setup in the literature for high-dimensional settings, see for instance Van de Geer et al. (2014)). Asymptotically, we distinguish between two regimes for the tuning parameters which we call conservative and consistent tuning. As suggested from the results in Pötscher & Schneider (2010), our results from finite samples essentially carry over asymptotically when the estimator is tuned conservatively. In the case of consistent tuning, the uniform convergence rate of the estimator is slower than \( n^{-1/2} \) and we give the asymptotic distribution of the Lasso estimator when scaled by the appropriate factor corresponding to the uniform convergence rate, as well as suggesting a simple construction for a confidence set in that case.

The remaining paper is organized as follows. In Section 2 we set the framework by stating the model, defining the estimator and introducing some notation. The main result giving the formula for the minimal coverage probability is presented in Section 3 and subsequently Section 4 is devoted to discussing how to concretely construct the corresponding confidence sets, as well as their relationship to the confidence ellipse based on the LS estimator. In Section 5 we derive asymptotic results both for the case of conservative and the case of consistent model selection. Section 6 concludes. All proofs are deferred to Section 7.

2 Setting and Assumptions

Consider the linear model

\[ y = X\beta + \varepsilon, \]

where \( y \) is the observed \( n \times 1 \) data vector, \( X \) the \( n \times p \) regressor matrix which is assumed to be non-stochastic with full column rank \( p \), \( \beta \in \mathbb{R}^p \) is the true parameter vector and \( \varepsilon \) the unobserved error term defined on some probability space \( (\Omega, \mathcal{A}, P) \) and consisting of independent and identically distributed components with mean 0 and finite variance \( \sigma^2 \). We consider a componentwise tuned Lasso estimator \( \hat{\beta}_c \), defined as the unique solution to the minimization problem

\[ \min_{\beta \in \mathbb{R}^p} L_n(\beta) = \min_{\beta \in \mathbb{R}^p} \| y - X\beta \|^2 + 2 \sum_{j=1}^p \lambda_{n,j}|\beta_j|, \]

where \( \lambda_{n,j} \) are non-negative and non-random componentwise tuning parameters that allow to exclude parameters from penalization. Note that if \( \lambda_{n,j} = 0 \) for all \( j \), this estimator is equal to the ordinary least-squares (LS) estimator \( \hat{\beta}_{\text{LS}} \) and that \( \lambda_{n,j} = c > 0 \) for all \( j \) corresponds to the ”classical” Lasso estimator as proposed by Tibshirani (1996). For later use, let \( \lambda_n = (\lambda_{n,1}, \ldots, \lambda_{n,p})' \) and \( \Lambda_n = \text{diag}(\lambda_n) \), the diagonal matrix whose diagonal elements are given by the components of \( \lambda_n \). We use \( 1_{\{\cdot\}} \) for the indicator function and make the following obvious definitions. For \( a \in \mathbb{R}^p \) and \( B \subseteq \mathbb{R}^p \), the set
For a $p \times p$ matrix $C$ and a scalar $c$, the sets $CB$ and $cB$ in $\mathbb{R}^p$ are $\{Cb : b \in B\} \subseteq \mathbb{R}^p$ and $\{cb : b \in B\} \subseteq \mathbb{R}^p$, respectively. Finally, for $k \in \mathbb{N}$, $I_k$ stands for the $k \times k$ identity matrix and $\bar{\mathbb{R}}$ denotes the extended real line $\mathbb{R} \cup \{-\infty, \infty\}$.

### 3 Finite Sample Results

We aim to construct confidence sets for the entire parameter vector $\beta$ based on the Lasso estimator $\hat{\beta}_L$. That means that for a non-random set $M \subseteq \mathbb{R}^p$, we consider sets of the form

$$\hat{\beta}_L - M = \{\hat{\beta}_L - m : m \in M\},$$

which have to satisfy that the probability of actually covering the unknown parameter $\beta$ never (for no value of $\beta$) falls below a prescribed level $1 - \alpha$ with $\alpha \in [0, 1]$. In other words, we need $P_\beta(\beta \in \hat{\beta}_L - M) \geq 1 - \alpha$ for all $\beta \in \mathbb{R}^p$ (where we stress the dependence of the probability measure on $\beta$ whenever it occurs), so that

$$\inf_{\beta \in \mathbb{R}^p} P_\beta(\beta \in \hat{\beta}_L - M) \geq 1 - \alpha.$$  

In order to achieve this, we need to be able to compute this “infimal” (minimal) coverage probability. Throughout this and the subsequent section we suppose that the errors as normally distributed

$$\varepsilon \sim N(0, \sigma^2 I_n),$$

an assumption that will be removed for asymptotic results in Section $\text{[]}$. We will show that the minimum occurs when the components of the unknown parameter become large in absolute value by essentially doing the following. We reparametrize the objective function defining the Lasso estimator so that the dependence on the unknown parameter becomes more transparent and easier to handle. We then consider the limiting cases of the objective functions when the components of the unknown parameter become large in absolute value (that is, tend to $+\infty$ or $-\infty$). We will see that it is possible to minimize the resulting objective functions explicitly, with minimizers that follow a shifted normal distribution that has the same variance-covariance matrix as the LS estimator and by construction do not depend on the unknown parameter. Finally, we will show that the infimal coverage probability of the proposed sets is indeed “achieved” for one of these finitely many limiting cases.

To state the main theorem, we need several definitions. First we define the reparametrized objective function $Q_n(u) = L_n(\beta + n^{-1/2}u) - L_n(\beta)$ so that $Q_n$ is uniquely minimized at $\tilde{u}_n = n^{1/2}(\hat{\beta}_L - \beta)$, the estimation error scaled by $n^{1/2}$. Of course, this scaling factor is arbitrary in finite samples, but proves to be of advantage when considering the problem in large samples in Section $\text{[5.]}$. We can write $Q_n$ as

$$Q_n(u) = u' C_n u - 2 u' W_n + 2n^{-1/2} \sum_{j=1}^p \lambda_{n,j} \left[ |u_j + n^{1/2} \beta_j| - |n^{1/2} \beta_j| \right],$$

where $C_n = X'X/n$ and $W_n = n^{-1/2}X'\varepsilon \sim N(0, \sigma^2 C_n)$. Note that for a set $M \subseteq \mathbb{R}^p$ we then have

$$P_\beta(\beta \in \hat{\beta}_L - n^{-1/2}M) = P_\beta(\tilde{u}_n \in M).$$

The above mentioned limiting cases of the objective function that we consider are defined as

$$Q_n^d(u) = u' C_n u - 2 u' W_n + 2n^{-1/2} \sum_{j=1}^p \lambda_{n,j} d_j u_j,$$

where $d = (d_1, \ldots, d_p)' \in \{-1, 1\}^p$. Holding $W_n$ fixed for a moment, we indeed see that

$$Q_n^d(u) = \lim_{j_d, j_2, \ldots, j_p \to \infty} Q_n(u).$$

As shorthand notation, we write $\tilde{u}_n^d$ for the unique minimizer of $Q_n^d$. To define the shape that we want to consider for the confidence regions, we introduce the following notation. For $m \in \mathbb{R}^p$, a vector
Figure 1: The set $A^{-\bar{\iota}}_C(m)$ with $m = (1.5, 2)'$ and $\bar{C} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$ along with the hyperplanes defining the set. The point $m = (1.5, 2)'$ is displayed as a dot.

For $d \in \{-1, 1\}^p$ and a matrix $\bar{C} \in \mathbb{R}^{p \times p}$, we define

$$A^d_C(m) = \bigcap_{j=1}^p \{z \in \mathbb{R}^p : d_j(\bar{C}m)_j \leq d_j(\bar{C}z)_j, d_jz_j \leq 0\}.$$  

The set $A^d_C(m)$ is an intersection of $2p$ half-spaces, $p$ of which determine the orthant the set is located in via the parameter $d$. The other $p$ half-spaces are defined by hyperplanes that intersect at the point $m$. Figure 1 shows one example of such a set. Note that in general, $A^d_C(m)$ could be non-empty also for $\text{sgn}(m) \neq -d$. The sets we consider are determined by the following condition.

**Condition A.** Let $\bar{C} \in \mathbb{R}^{p \times p}$ be given. We say that a set $M \subseteq \mathbb{R}^p$ satisfies Condition A with matrix $\bar{C}$ if

$$A^d_C(m) \subseteq M$$

for all $d \in \{-1, 1\}^p$ and for all $m \in M$.

The above condition will be discussed in more detail in Section 4. Using this notation, we can now state the main theorem.

**Theorem 1.** If $M_n \subseteq \mathbb{R}^p$ is non-random and satisfies Condition A with $\bar{C} = C_n$, then

$$\inf_{\beta \in \mathbb{R}^p} P_\beta(\hat{u}_n \in M_n) = \min_{d \in \{-1, 1\}^p} P(\hat{u}^d_n \in M_n),$$

where $\hat{u}^d_n \sim N(-n^{-1/2}C_n^{-1}\Lambda_n d, \sigma^2 C_n^{-1})$.

The distributions of $\hat{u}^d_n$ determining the formula for the infimal coverage probability are shifted normal distributions with the same variance-covariance matrix as the corresponding (shifted and scaled) LS estimator $\hat{u}_{\text{LS}} = n^{1/2}(\hat{\beta}_{\text{LS}} - \beta)$ and mean that depends on the regressors and the vector of tuning parameters. Since Condition A for $p = 1$ simply requires the corresponding set $M_n$ to be an interval containing zero, Theorem 1 is indeed a generalization of the formula in Theorem 5(a) in Pötscher & Schneider (2010), as discussed in the introduction. (To make the connection, note that the tuning parameter $\eta_n$ in that reference corresponds to a component $n^{-1/2}\lambda_{n,j}$ of the vector of tuning parameters in our paper.) The following obvious corollary specifies the resulting valid confidence region based on the Lasso estimator.
Corollary 2. Let $0 < \alpha < 1$. If $M_n \subseteq \mathbb{R}^p$ is non-random and satisfies Condition $\mathcal{A}$ with $\bar{C} = C_n$, as well as $\min_{d \in \{-1,1\}^p} P(\hat{u}_n^d \in M_n) = 1 - \alpha$ with $\hat{u}_n^d \sim N(-n^{-1/2}C_n^{-1}\Lambda_n d, \sigma^2 C_n^{-1})$, then
\[
\inf_{\beta \in \mathbb{R}^p} P(\beta \in \hat{M}_n - n^{-1/2}M_n) = 1 - \alpha.
\]

4 Constructing the Confidence Set

We now turn to discussing the important matter of how to choose an appropriate set $M_n \subseteq \mathbb{R}^p$ for some desired level of confidence $1 - \alpha$ by discussing concrete shapes for the confidence regions as well as their size and relation to confidence sets based on the LS estimator. As mentioned in the previous section, we need to find a set $M_n \subseteq \mathbb{R}^p$ that satisfies Condition $\mathcal{A}$ with $\bar{C} = C_n$ and such that $\min_{d \in \{-1,1\}^p} P(\hat{u}_n^d \in M_n) = 1 - \alpha$ where
\[
\hat{u}_n^d \sim N(-n^{-1/2}C_n^{-1}\Lambda_n d, \sigma^2 C_n^{-1}).
\]
The resulting confidence set for $\beta$ is then the scaled and shifted set $\hat{M}_n - M_n/n^{1/2}$. If we would base the set on the LS estimator $\hat{M}_n$, the canonical and best choice for $M_n$ in terms of volume is an ellipse determined by the contour lines of a $N(0, \sigma^2 C_n^{-1})$-distribution, the $C_n$-ellipse. Given the fact that the variance-covariance matrix of the distributions of $\hat{u}_n^d$ is in fact $\sigma^2 C_n^{-1}$, in addition to the fact that the means of the distributions average to 0, it is reasonable to consider the $C_n$-ellipse as a shape in connection with the Lasso estimator also. As stated in the following proposition, this shape complies with Condition $\mathcal{A}$.

Proposition 3. The $C_n$-ellipse given by
\[
E_{C_n}(k) = \{z \in \mathbb{R}^p : z' C_n z \leq k\}
\]
satisfies Condition $\mathcal{A}$ with $\bar{C} = C_n$ for any $k > 0$.

How to choose the parameter $k$ for a given level of coverage $1 - \alpha$ is stated in the next proposition.

Proposition 4. For any $k > 0$, we have that
\[
\arg\min_{d \in \{-1,1\}^p} P(\hat{u}_n^d \in E_{C_n}(k)) = \arg\max_{d \in \{-1,1\}^p} \|C_n^{-1/2}\Lambda_n d\|.
\]

Note that if $d^*$ solves the above optimization problem, so does $-d^*$. To finally obtain the confidence ellipse based on the Lasso estimator, pick any such optimizer $d^*$ and compute $k^* > 0$ so that $P(u_n^{d^*} \in E_{C_n}(k^*)) = 1 - \alpha$, which is easily done numerically. Note that Proposition 4 also shows that the ellipse $E_{C_n}(k^*)$, and therefore the resulting confidence set based on the Lasso estimator, is larger in volume than the one based on the LS estimator, since $E_{C_n}(k^*)$ needs to be large enough as to have mass $1 - \alpha$ with respect to the $N(-n^{1/2}C_n^{-1}\Lambda_n d^*, \sigma^2 C_n^{-1})$-measure whereas for the ellipse corresponding to the LS estimator, it suffices to have mass $1 - \alpha$ with respect to the $N(0, \sigma^2 C_n^{-1})$-measure. Clearly, the difference in size will increase as the tuning parameters become large. These observations are in line with the findings in Pötscher & Schneider (2010) who show that a confidence interval based on the Lasso estimator is larger than a confidence interval based on the LS estimator with the same coverage probability. When comparing the two confidence sets, we emphasize that since the ellipses are centered at different values, the smaller ellipse based on the LS estimator is in general not contained in the ellipse based on the Lasso estimator. This, as well as the difference in volume between the two ellipses, will also be illustrated in the example below.

It is quite obvious that the $C_n$-ellipse is not optimal as a shape for confidence sets based on the Lasso estimator since we can get higher coverage with a set of the same volume by adjusting the ellipse “towards” the contour lines of the $N(-n^{1/2}C_n^{-1}\Lambda_n d^*, \sigma^2 C_n^{-1})$-distributions (in such a way that Condition $\mathcal{A}$ is preserved). To find the best shape possible, one would have to minimize the volume of the set over all possible shapes satisfying Condition $\mathcal{A}$ subject to the constraint of holding the prescribed minimal coverage probability. This is a highly complex optimization problem and we do not dwell further on this subject here, but illustrate possible ways to construct “good” sets, as shown in the example below. Before discussing this further, note that the following proposition shows that it is easy to find the closure of an arbitrary subset of $\mathbb{R}^p$ with respect to Condition $\mathcal{A}$. 

5
Figure 2: The confidence ellipses based on and centered at the Lasso estimator $\hat{\beta}_L = (1.15, 0)'$ (red) and the smaller one based on and centered at the LS estimator $\hat{\beta}_{LS} = (1.35, 0.17)'$ (blue), respectively.

**Proposition 5.** For any $M \subseteq \mathbb{R}^p$, the set

$$\bigcup_{m \in M} \bigcup_{d \in \{-1, 1\}^p} A^d_C(m)$$

is the smallest set containing $M$ that satisfies Condition [A].

We now provide an example for $p = 2$ illustrating the difference between the confidence ellipse based on the LS estimator and the one based on the Lasso, as well as how to choose a better shape in terms of volume for the confidence set based on the Lasso estimator. The simulations and calculations were carried out using the statistical software package R. The example is set up in the following way. We let $n = 20$ and generate the $(n \times 2)$-matrix $X$ using independent and identically distributed standard normal entries that are transformed row-wise by an appropriate $(2 \times 2)$-matrix in order to get

$$C_n = X'X_n = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}.$$ 

We generate the data vector $y$ from the corresponding linear model with $\sigma^2 = 1$ (so that $\varepsilon \sim N(0, I_n)$) and true parameter chosen as $\beta = (1, 0)'$. We compute the Lasso estimator using the glmnet-package and tuning parameters $\lambda_{n,1} = \lambda_{n,2} = \sqrt{n}/2$ (asymptotically corresponding to what we will refer to as conservative model selection in the subsequent section). We also considered estimators where the tuning parameters were chosen by 10-fold cross-validation (as provided in the glmnet-package) which ended up yielding comparable results for the estimator.

We then constructed confidence ellipses with level $\alpha = 0.05$ based on both the LS and the Lasso estimator in the manner described earlier in this section. The resulting sets are shown in Figure 2. The plot clearly illustrates the above described fact that the confidence ellipse based on the Lasso estimator is larger than the confidence ellipse that is based on the LS estimator. Also, the two sets are overlapping by a large amount (in fact, the maximal distance between the two estimators is controlled by Proposition 13 in the Appendix). However, the LS ellipse is not entirely contained in the one based on the Lasso, stressing the fact the Theorem 1 yields non-trivial sets.

The above comparison between the two ellipses, however, is somewhat unfair in the sense that the shape used for both confidence sets is the optimal one (in terms of volume) for the LS estimator, but, as discussed above, not for the Lasso estimator. With the optimal shape for a Lasso confidence set being
unknown, we at least want to find a shape that improves upon the ellipse. As a basis for this, we consider the union of the contour sets corresponding to the distributions of $\check{u}_n^d$, that is, the $2^p$ shifted $C_n$-ellipses

$$U_n(k) = \bigcup_{d \in \{-1,1\}^p} E_{C_n}(k) - n^{-1/2}C_{n}^{-1}\Lambda_n d,$$

where each set in the union is of optimal shape for the corresponding distribution of $\check{u}_n^d$. As a starting point, we choose $k$ so that $P(\check{u}_n^d \in E_{C_n}(k) - n^{-1/2}C_{n}^{-1}\Lambda_n d) = 1 - \alpha$ (note that $k$ is then simply the parameter of the $C_n$-ellipse used for the LS estimator, but any $k > 0$ such that $U_n(k)$ satisfies $P(\check{u}_n^d \in U_n(k)) \geq 1 - \alpha$ works). Clearly, this set is still too large and will not satisfy Condition A, so we need to address these two issues. First, we add all points necessary so that the resulting set satisfies Condition A. Proposition 5 ensures that

$$\bigcup_{m \in U_n} \bigcup_{d \in \{-1,1\}^p} A_{C_n}(m)$$

fulfills the desired condition. Note that in this particular case, it is fairly straightforward to see that this set is simply given by the convex hull of the shifted ellipses $U_n(k)$. Finally, to get the smallest set with this shape that still holds the prescribed level of coverage, we iteratively adjust the set by reducing the parameter $k$ and re-calculate the minimal coverage probability of the resulting set until the desired minimal coverage probability is reached (up to an arbitrary level of precision). The resulting alternatively shaped set is depicted in Figure 3 (a) showing the midpoints of the $2^p = 4$ ellipses used in the construction and (b) displaying the new confidence set on top of the elliptic confidence region based on the Lasso as devised before. It is obvious that the new shape has slightly less volume than the ellipse.

5 Asymptotic Framework

We now derive asymptotic results that hold without assuming normality of the errors. Additionally to the assumptions in Section 2, for all asymptotic considerations, we assume that $X = (x_1', \ldots, x_n')'$ where
$x_i' \in \mathbb{R}^p$, meaning that the regressor matrix $X$ changes with $n$ only by appending rows, and that

$$C_n = \frac{X'X}{n} \rightarrow C$$

as $n \rightarrow \infty$, where $C$ is finite and positive definite. This setting assures consistency and asymptotic normality of the LS estimator. We will consider two different regimes of the asymptotic behavior of the tuning parameter $\lambda_n$ and start with the regime we call conservative tuning.

### 5.1 Conservative Tuning

In this regime and throughout this subsection, we require that

$$\frac{\lambda_n}{n^{1/2}} \rightarrow \lambda \in [0, \infty)^p$$

as $n \rightarrow \infty$. This implies that $\lambda_n/n \rightarrow 0$ for all $j = 1, \ldots, p$, which in turn implies consistency of $\hat{\beta}_n$ (see Theorem 1 in Knight & Fu (2000) with the slight modification that in our paper we allow for componentwise tuning parameters). We let $\Lambda = \text{diag}(\lambda)$.

**Remark 1.** Such a choice of tuning parameters indeed yields a conservative model selection procedure in the sense that

$$\limsup_{n \rightarrow \infty} \sup_{\beta \in \mathbb{R}^p} P_\beta \left( \hat{\beta}_j = 0 \right) < 1$$

for each $j = 1, \ldots, p$. In particular, if $\beta_j = 0$, we have

$$\lim_{n \rightarrow \infty} \sup_{\beta \in \mathbb{R}^p} P_\beta \left( \hat{\beta}_j = 0 \right) < 1.$$  

The latter statement was also noted by Zou (2006) in Proposition 1.

The following proposition implicitly states the asymptotic distribution of the estimator in a so-called moving-parameter framework. This proposition essentially is Theorem 5 from Knight & Fu (2000) and can be proven in the same manner simply by adjusting for componentwise tuning.

**Proposition 6.** Assume that $n^{1/2} \beta_n \rightarrow t \in \mathbb{R}^p$. Then $n^{1/2}(\hat{\beta}_n - \beta_n) \overset{d}{\rightarrow} \hat{u} = \arg \min_{u \in \mathbb{R}^p} Q(u)$, where

$$Q(u) = u'Cu - 2W'u + 2 \sum_{j=1}^p \lambda_j \left[ I_{(t_j \in \mathbb{R}^p)}(|t_j + u_j| - |t_j|) + I_{(|t_j| = \infty)} \text{sgn}(t_j)u_j \right]$$

and $W \sim N(0, \sigma^2 C)$.

Note that the vector $t$ takes over the role of $n^{1/2} \beta$ in the finite-sample version of the function, $Q_n$, where the cases of $n^{1/2} \beta = \pm \infty$ are now included in the asymptotic setting. Also, the assumption of $n^{1/2} \beta_n$ converging in $\mathbb{R}^p$ is not a restriction in the sense that, by compactness of $\mathbb{R}^p$, Proposition 6 characterizes all accumulation points of the distributions (with respect to weak convergence) corresponding to completely arbitrary sequences of $\beta_n$.

Similarly to the finite-sample case, we define $\hat{u}$ to be the unique minimizer of $Q$, and for $d \in \{-1, 1\}^p$, we define $Q^d(u) = u'Cu - 2W'u + 2 \sum_{j=1}^p \lambda_j d_j u_j$ with unique minimizer $\hat{u}^d$. We can then formulate an asymptotic version of Theorem 1.

**Theorem 7.** If $M \subseteq \mathbb{R}^p$ satisfies Condition $A$ with $C = C$, then

$$\inf_{t \in \mathbb{R}^p} P_t (\hat{u} \in M) = \min_{d \in \{-1, 1\}^p} P (\hat{u}^d \in M),$$

where $\hat{u}^d \sim N(C^{-1} \Lambda d, \sigma^2 C^{-1})$.

Given this result we can, again, construct asymptotically valid confidence sets for the parameter $\beta$ in the following way.
Corollary 8. If $M \subseteq \mathbb{R}^p$ satisfies Condition $\mathcal{A}$ with $\bar{C} = C$ and $\min_{d \in (-1,1)} P(\hat{u}^d \in M) = 1 - \alpha$, where $\hat{u}^d \sim N(C^{-1}Ad, \sigma^2 C^{-1})$ then

$$\liminf_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P(\hat{\beta} \in \hat{\beta}_n - n^{-1/2}M) = 1 - \alpha.$$ 

We find that asymptotically in the case of conservative tuning, we essentially get the same results as in finite samples when assuming normally distributed errors. The only difference is that the minimal coverage holds asymptotically and that the quantities $C_n$ and $n^{-1/2}\Lambda_n$ have settled to their limiting values $C$ and $\Lambda$, respectively.

5.2 Consistent Tuning

In the second regime and throughout this subsection, we suppose that

$$\frac{\lambda_{n,j}}{n^{1/2}} \to \infty$$

for at least one $j$ with $1 \leq j \leq p$ as well as

$$\frac{1}{n} \lambda_{n,j} \to 0$$

for all $j = 1, \ldots, p$ as $n \to \infty$, where the latter condition ensures estimation consistency of the estimator.

We refer to this regime as consistent tuning to highlight the contrast to conservative tuning where $\lambda_{n,j}/n^{1/2}$ converges for each $j = 1, \ldots, p$. Yet we emphasize that in order to ensure $P(\hat{\beta} = 0) \to 1$ whenever $\beta_j = 0$, we would need $\lambda_{n,j}/n^{1/2} \to \infty$ for each $j = 1, \ldots, p$ as well as need additional conditions on the regressor matrix $X$. We refer the reader to Zou (2006), Zhao & Yu (2006) and Yuan & Lin (2007) for a discussion concerning necessary and sufficient conditions on $X$ in this context.

In the case of consistent tuning, the rate of the estimator is no longer $n^{-1/2}$, neither when looked at in a fixed-parameter asymptotic framework (as has been noted by Zou (2006) in Lemma 3), nor (a fortiori) within a moving-parameter asymptotic framework, as discussed in in Pötscher & Leeb (2009) in Theorem 2. The latter reference shows that the correct (uniform) convergence rate depends on the sequence of tuning parameters $\lambda_n$. Since we allow for componentwise tuning, in fact, the rate depends on the largest component of the vector of tuning parameters, as can be seen from the following proposition.

We define

$$\lambda^*_n = \max_{1 \leq j \leq p} \lambda_{n,j}$$

and $\lambda_0 = (\lambda_{0,1}, \ldots, \lambda_{0,p})'$ by

$$\lambda_{n,j}/\lambda^*_n \to \lambda_{0,j} \in [0,1]$$

for each $j = 1, \ldots, p$ as $n \to \infty$. Note that $\lambda_{0,j} = 1$ for all $j$ in case all components are equally tuned.

Proposition 9. Assume that $n(\hat{\beta}_n - \beta)/\lambda^* \to \zeta \in \mathbb{R}^p$. Then $n(\hat{\beta}_n - \beta)/\lambda_n \overset{p}{\to} m = \arg \min_{u \in \mathbb{R}^p} V^\zeta(u)$, where

$$V^\zeta(u) = u'C u + 2 \sum_{j=1}^p \lambda_{0,j} \left[ \mathbb{I}_{\{\zeta_j \in \mathbb{R}\}} (|u_j + \zeta_j| - |\zeta_j|) + \mathbb{I}_{\{|\zeta_j| = \infty\}} \text{sgn}(\zeta_j) u_j \right].$$

(In contrast to the finite-sample and the conservative case, we make the dependence of the objective function $V^\zeta$ on the unknown parameter $\zeta \in \mathbb{R}^p$ apparent in the notation to clarify what we do in the following). Proposition 9 shows that $\lambda^*_n/n$ is indeed the correct (uniform) convergence rate as the limit of $n(\hat{\beta}_n - \beta)/\lambda_n$ is not 0 in general. The proposition also reveals that in the consistently tuned case, when scaled according the correct convergence rate, the limit of the sequence of estimators is always non-random, a fact that in a moving-parameter asymptotic framework has already been noted in the one-dimensional case in Pötscher & Leeb (2009). This fact allows us to construct very simple confidence sets in the case of consistent tuning by first observing that the limit of $n(\hat{\beta}_n - \beta)/\lambda_n$ is always contained in a bounded set which is described in Proposition 10. To this end, define the set

$$\mathcal{M} = \bigcup_{\zeta \in \mathbb{R}^p} \arg \min_{u \in \mathbb{R}^p} V^\zeta(u)$$

(4)

and note that the following can be shown.
Proposition 10. The set $\mathcal{M}$ can be written as
\[
\{ m \in \mathbb{R}^p : |(Cm)_j| \leq \lambda_{0,j}, 1 \leq j \leq p \} = C^{-1} \{ z \in \mathbb{R}^p : |z_j| \leq \lambda_{0,j}, 1 \leq j \leq p \}.
\]

Thus $\mathcal{M}$ can be viewed as a box distorted by the linear function $C^{-1}$, a bounded set in $\mathbb{R}^p$. In fact, this turns out to be a parallelogram whose corner points are given by the set $\{ C^{-1} \Lambda_0 d : d \in \{-1, 1\}^p \}$, where $\Lambda_0 = \text{diag}(\lambda_0)$. Note that fittingly, these corner points can be viewed as the equivalent of the means in the normal distributions (determining the minimal coverage probability) in the conservative case in Theorem 7 appearing without randomness in the limit in the consistently tuned case. Using Proposition 10, a simple asymptotic confidence set can now be constructed as is done in the following corollary.

Corollary 11. We have
\[
\lim_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_{\beta} \left( \beta \in \hat{\beta}_L - d \frac{\lambda_n}{n} \mathcal{M} \right) = 1
\]
for any $d > 1$ and
\[
\lim_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_{\beta} \left( \beta \in \hat{\beta}_L - d \frac{\lambda_n}{n} \mathcal{M} \right) = 0
\]
for any $d < 1$.

Note that nothing can be said about the boundary case $d = 1$. This corollary is a generalization of the simple confidence interval given in Proposition 6 in [Pötscher & Schneider (2010)]. Finally, also note the set $\mathcal{M}$ is not required to satisfy Condition $\mathfrak{A}$ and, in fact, will not comply with this condition for certain matrices $C$.

6 Conclusion

We consider confidence regions based on the Lasso estimator covering the entire unknown parameter vector thereby quantifying estimation uncertainty of this estimator. We provide exact formulas for the minimal coverage probability of these regions in finite samples and asymptotically in a low-dimensional framework when the estimator is tuned to perform conservative model selection. We do this without explicit knowledge of the distribution but by carefully exploiting the structure of the optimization problem that defines the estimator. The sets we consider as confidence regions need to satisfy certain shape constraints which apply to the regular confidence ellipse based on the LS estimator. We show that the LS confidence ellipse is always smaller than the one based on the Lasso estimator, but not contained in the Lasso ellipse in general. An ellipse is not the optimal shape for the confidence region based on the Lasso estimator in terms of volume. We give some guidelines on how to construct regions of smaller volume. We show how a set can be minimally enlarged in order to comply with the imposed shape condition, allowing to start the construction with sets of arbitrary shapes.

In the consistently tuned case, we give a simple asymptotic confidence regions in the shape of a parallelogram that is determined by the regressor matrix.

Acknowledgements

The authors gratefully acknowledge support from DFG grant FOR 916.

7 Proofs

We start the proof section with introducing some notation that will be used throughout this section. Let $e_j$ denote the $j^{th}$ unit vector in $\mathbb{R}^p$ and let $\iota = (1, \ldots, 1)^T \in \mathbb{R}^p$. For a vector $d \in \{-1, 1\}^p$, we define $\mathcal{O}^d$ to be the corresponding orthant of $\mathbb{R}^p$, that is, $\mathcal{O}^d = \{ z \in \mathbb{R}^p : d_j z_j \geq 0 \}$ and $\bar{\mathcal{O}}^d$ to be the corresponding orthant of $\mathbb{R}^p$, that is, $\bar{\mathcal{O}}^d = \{ z \in \mathbb{R}^p : d_j z_j \geq 0 \}$. By $\mathcal{O}_{\text{int}}^d$ we denote the orthant with strictly positive components only, that is, $\mathcal{O}_{\text{int}}^d = \{ z \in \mathbb{R}^p : z_j > 0 \}$. The sup-norm on $\mathbb{R}^p$ is denoted by $\| \cdot \|_\infty$.

To remind the reader of some notation relevant for the following proofs that was introduced previously throughout the paper, note that $\hat{u}_n = n^{1/2}(\hat{\beta}_L - \beta)$, where $\hat{u}_n$ is the minimizer of $Q_n$, and $\hat{u}_{\text{LS}} =
$n^{1/2}(\hat{\beta}_{LS} - \beta)$. The minimizer of $Q_n^d$ was labeled $\hat{u}_n^d$. The asymptotic versions in the conservatively tuned case were labeled $\hat{u}$ and $Q$, as well as $\hat{u}^d$ and $Q^d$, respectively.

The directional derivative of a function $g : \mathbb{R}^p \to \mathbb{R}$ at $u$ in the direction of $r \in \mathbb{R}^p \setminus \{0\}$ is defined as

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

7.1 Proofs for Section 3

In order to prove the main theorem, we start by re-writing Condition A. For $m \in \mathbb{R}^p$ and a $p \times p$ matrix $\bar{C}$, we define

$$A_{C,j}^d(m) = \{ z \in \mathbb{R}^p : d_j(\bar{C}m)_j \leq d_j(\bar{C}z)_j, d_jz_j \leq 0 \}$$

and

$$B_{C,j}^d(m) = \{ z \in \mathbb{R}^p : (\bar{C}z)_j = (\bar{C}m)_j, d_jz_j > 0 \}$$

for $j = 1, \ldots, p$. Note that clearly we have

$$A_C^d(m) = \bigcap_{j=1}^p A_{C,j}^d(m),$$

and that, in fact, also the following lemma holds.

**Lemma 12.**

$$\bigcup_{d \in \{-1,1\}^p} \bigcap_{j=1}^p A_{C,j}^d(m) = \bigcup_{d \in \{-1,1\}^p} A_{C,j}^d(m) \cup B_{C,j}^d(m)$$

**Proof.** We fix $m$ and $\bar{C}$, drop the corresponding subscripts and show that the set on the left-hand side of the equation contains the set on the right-hand side of the equation. To this end, take any $z$ from the set on right-hand side. Then there exists a $d \in \{-1,1\}^p$ such that for each $j = 1, \ldots, p$, $z$ is either contained in $A_{j}^d$ or in $B_{j}^d$. We pick $f \in \{-1,1\}^p$ in the following way: if $z \in A_{j}^d$, set $f_j = d_j$ and if $z \in B_{j}^d$, set $f_j = -d_j$. Then, by construction, $z \in A_j^f$ for all $j = 1, \ldots, p$ and therefore $z \in \bigcap_{j=1}^p A_j^f$ so that $z$ is contained in the set on the left-hand side of the equation. 

Since needed later on, we also prove the following proposition which quantifies the maximal distance between the Lasso and the LS estimator in finite samples.

**Proposition 13.** For each $j = 1, \ldots, p$, we have

$$\left| (X'X(\hat{\beta}_{L} - \hat{\beta}_{LS}))_j \right| \leq \lambda_{n,j},$$

or, equivalently,

$$\left| (C_n(\hat{u}_n - \hat{u}_{LS}))_j \right| \leq n^{-1/2}\lambda_{n,j},$$

where $\hat{u}_{LS} = n^{1/2}(\hat{\beta}_{LS} - \beta)$.

**Proof.** The two inequalities above just differ by a scaling factor. We show the latter one. We have $W_n = n^{-1/2}X'\epsilon = C_n\hat{u}_{LS}$. Consider the directional derivative of $Q_n$ at its minimizer $\hat{u}_n$ in the direction of $e_j$ and $-e_j$. We have

$$0 \leq \frac{\partial}{\partial e_j} Q_n(\hat{u}_n) = 2(C_n\hat{u}_n)_j - 2W_n,j + 2n^{-1/2}\lambda_{n,j} \left[ \mathbf{1}_{\{\hat{u}_j \geq -n^{1/2}\beta_j\}} - \mathbf{1}_{\{\hat{u}_j < -n^{1/2}\beta_j\}} \right]$$

$$\leq 2(C_n\hat{u}_n)_j - 2(C_n\hat{u}_{LS})_j + 2n^{-1/2}\lambda_{n,j},$$

as well as

$$0 \leq \frac{\partial}{\partial (-e_j)} Q_n(\hat{u}_n) = -2(C_n\hat{u}_n)_j + 2W_n,j + 2n^{-1/2}\lambda_{n,j} \left[ \mathbf{1}_{\{\hat{u}_j \leq -n^{1/2}\beta_j\}} - \mathbf{1}_{\{\hat{u}_j > -n^{1/2}\beta_j\}} \right]$$

$$\leq -2(C_n\hat{u}_n)_j + 2(C_n\hat{u}_{LS})_j + 2n^{-1/2}\lambda_{n,j}.$$ 

Piecing the two displays above together yields the second inequality in the proposition. 

11
To proceed note that \( Q_n^d \) as defined in [1] is a simple quadratic and strictly convex function in \( u \) with unique minimizer \( \hat{u}_n^d \) given by
\[
\hat{u}_n^d = C_n^{-1}(W_n - n^{-1/2}A_n d),
\] (5)
where \( W_n \sim N(0, \sigma^2 C_n) \). We first show Theorem [1] for one orthant of the parameter space \( \mathbb{R}^p \), as is formulated in Proposition [14].

**Proposition 14.** If \( M_n \subseteq \mathbb{R}^p \) satisfies that
\[
\bigcap_{j=1}^p A_{C_n,\beta}^{i,j}(m) \cup B_{C_n,\beta}^{i,j}(m) \subseteq M_n
\]
for all \( m \in M_n \), then
\[
\inf_{\beta \in \Omega} P_\beta(\hat{u}_n \in M_n) = P(\hat{u}_n^i \in M_n).
\]

In essence, Proposition [14] states Theorem [1] for the orthant of the parameter space where all components of \( \beta \) are non-negative. The condition in Proposition [14] takes the role of Condition [4] for the corresponding orthant, as will become apparent later on in the proof of Theorem [1].

**Proof of Proposition [14].** We first show that \( \inf_{\beta \in \Omega} P_\beta(\hat{u}_n \in M_n) \geq P(\hat{u}_n^i \in M_n) \) by showing that for each fixed \( \omega \in \Omega \), \( \hat{u}_n^i \in M_n \) implies that \( \hat{u}_n \in M_n \) as long as \( \beta_j \geq 0 \) for all \( j \). For this, we first show the following two facts.

(a) \((C_n\hat{u}_n^i)_j \leq (C_n\hat{u}_n)_j \) for all \( j = 1, \ldots, p \).

Suppose there exists a \( j_0 \) with such that \((C_n\hat{u}_n^i)_j_0 > (C_n\hat{u}_n)_j_0 \) and note that by [6] we have \((C_n\hat{u}_n^i)_j = W_{n,j} - n^{-1/2}\lambda_{n,j} \) for each \( j = 1, \ldots, p \). Now consider the directional derivative of \( Q_n \) at its minimizer \( \hat{u}_n \) in direction \( e_{j_0} \),
\[
\frac{\partial Q_n(\hat{u}_n)}{\partial e_{j_0}} = 2(C_n\hat{u}_n)_j - 2W_{n,j_0} + 2n^{-1/2}\lambda_{n,j_0} \left[ 1\{\hat{u}_n,j_0 \geq -n^{-1/2}\beta_{j_0}\} - 1\{\hat{u}_n,j_0 < -n^{-1/2}\beta_{j_0}\} \right]
\]
\[
\leq 2(C_n\hat{u}_n)_j - 2W_{n,j_0} + 2n^{-1/2}\lambda_{n,j_0}
\]
\[
= 2(C_n\hat{u}_n)_j - 2(C_n\hat{u}_n^i)_j < 0,
\]
which is a contradiction to \( \hat{u}_n \) minimizing \( Q_n \).

(b) \( \hat{u}_{n,j} > 0 \) implies \((C_n\hat{u}_n^i)_j = (C_n\hat{u}_n)_j \) for any \( 1 \leq j \leq p \).

If \( \hat{u}_{n,j} > 0 \) (and hence \( \hat{u}_{n,j} + n^{1/2}\beta_j > 0 \) when \( \beta_j \geq 0 \)), then \( Q_n \) is partially differentiable at \( \hat{u}_n \) with respect to the \( j \)th component. Therefore, we have
\[
\frac{\partial Q_n(\hat{u}_n)}{\partial u_j} = 2(C_n\hat{u}_n)_j - 2W_{n,j} + 2n^{1/2}\lambda_{n,j}
\]
\[
= 2(C_n\hat{u}_n)_j - 2(C_n\hat{u}_n^i)_j = 0.
\]

Now, by Facts [a] and [b] we clearly have that \( \hat{u}_n \in A_{C_n,\beta}^{i,j}(\hat{u}_n^i) \cup B_{C_n,\beta}^{i,j}(\hat{u}_n^i) \). So, by assumption, \( \hat{u}_n^i \in M_n \) clearly implies \( \hat{u}_n \in M_n \) as long as \( \beta_j \geq 0 \) for all \( j \). We have therefore shown that
\[
\inf_{\beta \in \Omega} P_\beta(\hat{u}_n \in M_n) \geq P(\hat{u}_n^i \in M_n).
\]

To see the reverse inequality, note that if \( \hat{u}_{n,j} + n^{1/2}\beta_j > 0 \) for all \( j \), then \( Q_n \) is differentiable at \( \hat{u}_n \) and
\[
\frac{\partial Q_n(\hat{u}_n)}{\partial u} = 2C_n\hat{u}_n - 2W_n + 2n^{-1/2}\lambda_n = 2C_n\hat{u}_n - 2C_n\hat{u}_n^i = 0,
\]
implying that \( \hat{u}_n = \hat{u}_n^i \). Also note that \( \hat{u}_{n,j} + n^{1/2}\beta_j > 0 \) for each \( j \) is equivalent to \( \hat{\beta}_L \in \mathcal{O}_{int} \), so that
\[
\{\hat{u}_n \in M_n\} \subseteq \{\hat{u}_n^i \in M_n\} \cup \{\hat{\beta}_L \notin \mathcal{O}_{int}\}.\]
Now let $\kappa$ be a bound in the sup-norm on the set $\{z \in \mathbb{R}^p : \|C_n z\|_\infty \leq n^{-1/2}\|\lambda_n\|_\infty\}$ and for an arbitrary $\varepsilon > 0$, pick $\beta^* \in \mathbb{R}^p$ such that $P(\hat{u}_{LS} \leq \kappa - n^{1/2}\beta^*) \leq \varepsilon$, where $\hat{u}_{LS} = n^{1/2}(\hat{\beta}_{LS} - \beta^*) \sim N(0, \sigma^2 C_n^{-1})$.

Note that by Proposition 13, this implies that
$$P_{\beta^*}(\hat{\beta}_L \leq 0) = P_{\beta^*}(\hat{u}_n - \tilde{u}_{LS} + \tilde{u}_{LS} \leq -n^{1/2}\beta^*) \leq P_{\beta^*}(-\kappa + \tilde{u}_{LS} \leq -n^{1/2}\beta^*) \leq \varepsilon,$$
yielding
$$\inf_{\beta \in \mathbb{R}^p} P_{\beta}(\hat{u}_n \in M_n) \leq P_{\beta^*}(\hat{u}_n \in M_n) \leq P(\tilde{u}_n' \in M_n) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows the desired inequality.

Essentially, we have now shown the main theorem for one part of the parameter space $\mathbb{R}^p$. By flipping signs, we can apply Proposition 14 to each orthant $O^d$, thus obtaining the formula for the infimal coverage over the whole space.

**Proof of Theorem**

First note that
$$\inf_{\beta \in \mathbb{R}^p} P_{\beta}(\hat{u}_n \in M_n) = \min_{d \in \{-1, 1\}^p} \inf_{\beta \in O^d} P_{\beta}(\hat{u}_n \in M_n).$$

Thus, if we can show that
$$\inf_{\beta \in O^d} P_{\beta}(\hat{u}_n \in M_n) = P(\tilde{u}_n^d \in M_n)$$
for each $d \in \{-1, 1\}^p$, the proof is done. Now, fix $d$ and set $D = \text{diag}(d)$. We consider the function
$$\tilde{Q}_n(u) = Q_n(Du) = u'DC_n Du - 2u'DW_n + 2n^{-1/2} \sum_{j=1}^p \lambda_{n,j} \left[|d_j u_j + n^{1/2} \beta_j| - |n^{1/2} \beta_j|\right],$$
where $\tilde{\beta}_n = DC_n D, \tilde{W}_n = DW_n \sim N(0, \sigma^2 \tilde{C}_n)$. We write $\tilde{u}_n$ for the minimizer of $\tilde{Q}_n$, and, analogously to Section 3, we define $\tilde{u}_n^d$ to be the minimizer of the function $u'\tilde{\beta}_n u - 2u'\tilde{W}_n + 2n^{-1/2} \sum_{j=1}^p \lambda_{n,j} u_j$.

If we can show that the set $DM_n$ satisfies the requirement of Proposition 14 with the matrix $\tilde{C}_n^*$ in place of $C_n$, we may conclude that
$$\inf_{\beta, d, \beta_j \geq 0} P_{\beta}(\hat{u}_n \in DM_n) = P(\tilde{u}_n^d \in DM_n).$$

Note that $\hat{u}_n = D\tilde{u}_n, \tilde{u}_n^d = D\tilde{u}_n^d$ and $D^{-1} = D$, so that
$$\inf_{\beta \in O^d} P(\hat{u}_n \in M_n) = \inf_{\beta \in O^d} P(\tilde{u}_n \in DM_n) = P(\tilde{u}_n^d \in DM_n) = P(\tilde{u}_n^d \in M_n),$$
which proves the formula for the infimal coverage probability. We now show that the set $DM_n$ satisfies that
$$\bigcap_{j=1}^p A_{\tilde{C}_n,j}^d(Dm) \cup B_{\tilde{C}_n,j}^d(Dm) \subseteq DM_n$$
for all $m \in M_n$. A straightforward calculation shows that this is equivalent to
$$\bigcap_{j=1}^p A_{C_n,j}^d(m) \cup B_{C_n,j}^d(m) \subseteq M_n$$
for each $m \in M$ which clearly holds by Condition A and Proposition 12.

The distributional result on $\tilde{u}_n^d$ immediately follows by [5].
7.2 Proofs for Section 4

Proof of Proposition 4. Let \( m \in E_{C_n}(k) \) and \( y \in A^d_{C_n}(m) \). We show that \( y \in E_{C_n}(k) \). Remember that \( D = \text{diag}(d) \) satisfies \( DD = I_p \). Since \( y \in A^d_{C_n}(m) \) we have \( -Dy \in \mathcal{O}^c \) and \( -DC(m - y) \in \mathcal{O}^c \) implying that
\[
y'C(m - y) = (Dy)'DC(m - y) \geq 0.
\]
Furthermore, since \( (m - y)'C(m - y) \geq 0 \), we have
\[
m'C(m - y) \geq y'C(m - y) \geq 0,
\]
which in turn yields
\[
m'Cm \geq m'Cy \geq y'Cy \geq 0.
\]
But this means that \( k \geq m'Cm \geq m'Cy \geq y'Cy \) and therefore \( y \in E_{C_n}(k) \). \( \square \)

Proof of Proposition 5. We transform the ellipse to a sphere and the corresponding normal distribution to have independent components with equal variances.

\[
P \left( \tilde{u}^d_n \in E_{C_n}(k) \right) = P \left( \hat{C}_{n}^{1/2} \tilde{u}^d_n \in \mathcal{C}_{n}^{1/2} E_{C_n}(k) \right),
\]
where \( \hat{C}_{n}^{1/2} \tilde{u}^d_n \sim N(-\hat{C}_{n}^{-1/2} \Lambda_d, \sigma^2 I_p) \) and \( \mathcal{C}_{n}^{1/2} E_{C_n}(k) = \{ z \in \mathbb{R}^p : \| z \| \leq k \} \). So clearly, the smallest probability will be achieved for the distribution with mean furthest away from the origin, which is any \( d^* \) maximizing \( \| \hat{C}_{n}^{-1/2} \Lambda_d d \| \) over all \( d \in \{-1, 1\}^p \). \( \square \)

Proof of Proposition 6. We start by showing that for any \( m \in \mathbb{R}^p, d \in \{-1, 1\}^p \), we have
\[
A^d_{C}(y) \subseteq A^d_{C}(m) \quad \text{for all} \quad y \in A^d_{C}(m).
\]
Let \( z \in A^d_{C}(y) \). Then \( d_j z_j \leq 0 \) and \( (\hat{C}y)_j \leq (\hat{C}z)_j \) for all \( j \). But since \( y \in A^d_{C}(m) \), we also have \( (\hat{C}m)_j \leq (\hat{C}z)_j \) for all \( j \) so that \( (\hat{C}m)_j \leq (\hat{C}z)_j \) for all \( j \) and therefore \( z \in A^d_{C}(m) \), thus proving 6. So clearly, the set
\[
\bigcup_{m \in M} \bigcup_{d \in \{-1, 1\}^p} A^d_{C}(m)
\]
satisfies Condition [X]. For each \( m \in M \), choose \( d \in \{-1, 1\}^p \) in such a way that \( d_j = 1 \) if \( m_j = 0 \) and \( d_j = -\text{sgn}(m_j) \) for \( m_j \neq 0 \). We then get \( m \in A^d_{C}(m) \), implying that the set in the display above actually contains \( M \). \( \square \)

7.3 Proofs for Section 5

Proof of Remark 7. We show 2. Note that Proposition 13 entails that
\[
\hat{\beta}_L \in \hat{\beta}_{LS} - \frac{1}{n^{1/2}} B_n,
\]
where
\[
B_n = \{ z \in \mathbb{R}^p : \| (C_n z)_j \| \leq n^{-1/2} \lambda_{n,j} \quad \text{for} \quad j = 1, \ldots, p \}.
\]
Since \( \lambda_n/n^{1/2} \) converges, we have \( B_n \subseteq C_{n}^{-1} B_{\delta} \) with \( B_{\delta} = \{ x \in \mathbb{R}^p : \| x \|_\infty \leq \delta \} \) for some \( \delta > 0 \). Since \( C_{n}^{-1} \rightarrow C^{-1} \), the set \( \{ C_{n}^{-1} : n \in \mathbb{N} \} \) is bounded in operator sup-norm by Banach-Steinhaus, so that the set \( B_n \) is uniformly bounded over \( n \) in sup-norm by, say, \( \gamma > 0 \). We now fix a component \( j \) and show that \( \liminf_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_{\beta}(\hat{\beta}_{LS,j} \neq 0) > 0 \). To this end, define \( R_j = \mathbb{R}^{p-1} \times \{ 0 \} \times \mathbb{R}^{p-j} \). Let \( \xi_{i,n}^2 \) and \( \xi_{i,j}^2 \) be the positive \( j^\text{th} \) diagonal element of \( C_{n}^{-1} \) and \( C^{-1} \), respectively. Observe that
\[
\inf_{\beta \in \mathbb{R}^p} P_{\beta}(\hat{\beta}_{LS,j} \neq 0) \geq \inf_{\beta \in \mathbb{R}^p} P_{\beta}\left( (\hat{\beta}_{LS} - \frac{1}{n^{1/2}} B_n) \cap R_j = \emptyset \right)
\geq \inf_{\beta \in \mathbb{R}^p} P_{\beta}(n^{1/2}\hat{\beta}_{LS,j} + \gamma < 0 \text{ or } n^{1/2}\hat{\beta}_{LS,j} - \gamma > 0)
= 2\Phi(-\gamma/\xi_{i,n}) \rightarrow 2\Phi(-\gamma/\xi_{i}) > 0
\]
\( \square \)
In order to prove Theorem 7 we need an asymptotic version of Proposition 14 which is formulated in the following.

**Proposition 15.** If $M \subseteq \mathbb{R}^p$ satisfies that

$$\bigcap_{j=1}^{p} A_{ij}^{2j}(m) \cup B_{ij}^{2j}(m) \subseteq M$$

for all $m \in M_n$, then

$$\inf_{t \in \mathcal{O}} P_t(\hat{u} \in M) = P(\hat{u}^i \in M).$$

**Proof of Proposition 15.** The first part of the proof is completely analogous to the first part of the proof of Proposition 14 after identifying $n^{1/2} \beta$ with $t$ and dropping the subscript $n$. To see the reverse inequality, note that for $t^* = (\infty, \ldots, \infty) \in \mathbb{R}^p$, we actually have $Q = Q^i$, so that $\hat{u} = \hat{u}^i$ in this case which already yields that

$$\inf_{t \in \mathcal{O}} P_t(\hat{u} \in M) \leq P_t(\hat{u}^i \in M) = P(\hat{u}^i \in M).$$

□

**Proof of Theorem 7.** The proof again is completely analogous to the proof of Theorem 1 after identifying $n^{1/2} \beta$ with $t$, dropping the subscript $n$ everywhere and using Proposition 15 instead of Proposition 14. Also, replace $\mathcal{O}^d$ by $\mathcal{O}^d$ and note that

$$Q^d(u) = Q(Du) = u'DCDu - 2u'DW + 2 \sum_{i=1}^{p} \lambda_j \left[ \mathbb{1}_{\{t_j \in \mathbb{R}\}}(|t_j| + d_j u_j) + \mathbb{1}_{\{|t_j| = \infty\}} \text{sgn}(t_j) d_j u_j \right]$$

$$= u'\tilde{C}u - 2u'\tilde{W} + 2 \sum_{i=1}^{p} \lambda_j \left[ \mathbb{1}_{\{|d_j t_j| \in \mathbb{R}\}}(|u_j + d_j t_j| - |d_j t_j|) + \mathbb{1}_{\{|d_j t_j| = \infty\}} \text{sgn}(d_j t_j) u_j \right],$$

where $\tilde{C} = DCD$ and $\tilde{W} = DW$. □

**Proof of Corollary 8.** Let $c = \liminf_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_\beta(\beta \in \hat{\beta}_n - n^{-1/2} M)$. Then there exists a sequence $\beta_n$ in $\mathbb{R}^p$ such that $P_{\beta_n}(\beta_n \in \hat{\beta}_n - n^{-1/2} M) \to c$. Assume that $n^{1/2} \beta_n \to t \in \mathbb{R}^p$ (if the sequence does not converge, pass to subsequences). Since

$$P_{\beta_n}(\beta_n \in \hat{\beta}_n - n^{-1/2} M) = P_{\beta_n}(n^{1/2}(\hat{\beta}_n - \beta_n) \in M) \to c = P_t(\hat{u} \in M)$$

as $n \to \infty$ in the notation of Proposition 1. Theorem 7 then yields $c \geq \min_{d \in \{-1,1\}^p} P(\hat{u}^d \in M) = 1 - \alpha$. To see the reverse inequality, let $\beta_n = d \in \{-1,1\}^p$ and note that for this sequence, we have

$$P_{\beta_n}(\beta_n \in \hat{\beta}_n - n^{-1/2} M) = P_{\beta_n}(n^{1/2}(\hat{\beta}_n - \beta_n) \in M) \to P_t(\hat{u} \in M)$$

as $n \to \infty$, where $t = (d_1 \infty, \ldots, d_p \infty)' \in \mathbb{R}^p$. Note that for this choice of $t$, $P_t(\hat{u} \in M) = P(\hat{u}^d \in M)$. Since $d \in \{-1,1\}^p$ was arbitrary, $c \leq \min_{d \in \{-1,1\}^p} P(\hat{u}^d \in M) = 1 - \alpha$ follows. □

**Proof of Proposition 2.** Define the function $V_n(u) = n[L_n(\beta_n + \lambda_n^* u/n) - L_n(\beta_n)]/(\lambda_n^*)^2$ and note that $V_n$ is minimized at $n(\beta_n - \beta_n^*)/\lambda_n^*$. The function $V_n$ is then given by

$$V_n(u) = u'X'X^{-1} - 2\frac{1}{\lambda_n} u'X' \varepsilon + 2 \sum_{j=1}^{p} \frac{\lambda_n j}{\lambda_n^*} \left[ u_j + \frac{n}{\lambda_n^*} \beta_n j \right].$$

Clearly $u'X'Xu/n \to u'Cu$ by assumption. Since $X'\varepsilon/\lambda_n^* = (n^{1/2}/\lambda_n^*)X'\varepsilon/n^{1/2}$ and $\lambda_n^*/n^{1/2} \to \infty$ as well as $X'\varepsilon/n^{1/2} = O_p(1)$, the second term in the above display vanishes in probability. To treat the
third term, simply note that \( \lambda_{n,j}/\lambda_n^* \to \lambda_{0,j} \in [0,1] \) and \( n\beta_{n,j}/\lambda_n^* \to \zeta_i \in \mathbb{R} \) by assumption. Piecing this together yields

\[
V_n(u) \xrightarrow{P} u'Cu + 2\sum_{j=1}^{p} \lambda_{0,j} \left[ \mathbb{1}_{\{|\zeta_j|\leq 0\}}(|\zeta_j| - |\zeta_j|) + \mathbb{1}_{\{|\zeta_j|=\infty\}} \text{sgn}(\zeta_j) u_j \right] = V^\zeta(u).
\]

Since \( V_n \) and \( V^\zeta \) are strictly convex and \( V^\zeta \) is non-random, it follows by Geyer (1996) that also the corresponding minimizers converge in probability to the minimizer of the limiting function.

**Proof of Proposition 10.** The equality of the two sets given in the display of Proposition 10 is trivial. We show that the set \( \mathcal{M} \) as defined in (9) is equal to the set on the left-hand side and start by proving that \( \mathcal{M} \) is contained in that set. Take any \( m \in \mathcal{M} \), by definition, there exists a \( \zeta \in \mathbb{R}^d \) such that \( m \) is the minimizer of \( V^\zeta \). We need to show that \( |\langle Cm \rangle_j| \leq \lambda_{0,j} \) for all \( j \). Assume that \( |\langle Cm \rangle_j| > \lambda_{0,j} \) for some \( 1 \leq j_0 \leq p \). If \( \langle Cm \rangle_{j_0} > \lambda_{0,j_0} \) we consider the directional derivative of \( V^\zeta \) at its minimizer \( m \) in the direction of \(-e_{j_0}\) to get

\[
\frac{\partial V^\zeta(m)}{\partial (-e_{j_0})} = -2\langle Cm \rangle_{j_0} + 2\lambda_{0,j_0} \left[ \mathbb{1}_{\{m_j + \zeta_j \leq 0\}} - \mathbb{1}_{\{m_j + \zeta_j > 0\}} \right]
\]

which is a contradiction to \( m \) minimizing \( V^\zeta \). If \( \langle Cm \rangle_{j_0} < -\lambda_{0,j_0} \), then consider the directional derivative of \( V^\zeta \) at \( m \) in the direction of \( e_{j_0} \) to arrive at

\[
\frac{\partial V^\zeta(m)}{\partial e_{j_0}} = 2\langle Cm \rangle_{j_0} + 2\lambda_{0,j_0} \left[ \mathbb{1}_{\{m_j + \zeta_j \geq 0\}} - \mathbb{1}_{\{m_j + \zeta_j < 0\}} \right]
\]

yielding a contradiction also.

To see the reverse set-inclusion, we need to show that for any \( m \in \mathbb{R}^p \) satisfying \( |\langle Cm \rangle_j| \leq \lambda_{0,j} \) for all \( j = 1, \ldots, p \), there exists a \( \zeta \in \mathbb{R}^d \) such that \( m \) is the minimizer of \( V^\zeta \). Let \( \zeta = -m \in \mathbb{R}^d \) and consider the directional derivative of \( V^\zeta \) at \( m \) in any direction \( r \in \mathbb{R}^p \setminus \{0\} \):

\[
\frac{\partial V^\zeta(m)}{\partial r} = 2\langle Cr \rangle_{m} + 2\sum_{j=1}^{p} \lambda_{0,j} |r_j| \geq \sum_{j=1}^{p} -2|\langle Cm \rangle_j r_j| + 2\lambda_{0,j} |r_j| = 2\sum_{j=1}^{p} [\langle Cm \rangle_j + \lambda_{0,j}] |r_j| \geq 0.
\]

Since the directional derivative is non-negative in any direction \( r \in \mathbb{R}^p \setminus \{0\} \) and \( V^\zeta \) is (strictly) convex, \( m \) must be the minimizer.

**Proof of Corollary 11.** We start with the case \( d > 1 \). Let \( c = \liminf_{n \to \infty} \inf_{\beta \in \mathbb{R}^p} P_{\beta}(\beta \in \hat{\beta}_L - d\lambda_n^* \mathcal{M}/n) \). By definition, there exists a subsequence \( n_k \) and elements \( \beta_{n_k} \in \mathbb{R}^p \) such that

\[
P_{\beta_{n_k}} \left( \beta_{n_k} \in \hat{\beta}_L - d\lambda_{n_k}^* \mathcal{M} \right) = P_{\beta_{n_k}} \left( \frac{n_k}{\lambda_{n_k}^*} (\hat{\beta}_L - \beta_{n_k}) \in d\mathcal{M} \right) \to c
\]
as \( k \to \infty \). Note that \( d\mathcal{M} = \{ m \in \mathbb{R}^p : \langle Cm \rangle_j \leq d\lambda_{0,j}, 1 \leq j \leq p \} \). Now, pick a further subsequence \( n_k \) such that \( \lambda_{n_k}^* \beta_{n_k}/n_k \) converges in \( \mathbb{R}^p \) to, say, \( \zeta \). Proposition 9 then shows that \( n_k(\hat{\beta}_L - \beta_{n_k})/\lambda_{n_k}^* \) converges in probability to the unique minimizer of \( V^\zeta \) as \( l \to \infty \). Finally, Proposition 10 implies that \( c = 1 \).

We next look the case where \( d < 1 \). Let \( m = C^{-1}\lambda_0 \) so that \( m \in \mathcal{M} \setminus d\mathcal{M} \). From the proof of Proposition 10 we know that for \( \zeta = -m \) we have \( m = \arg\min_u \langle V^\zeta(u) \rangle \). Let \( \beta_n = n\zeta/\lambda_n^* \). By Proposition 9 \( n(\hat{\beta}_L - \beta_n)/\lambda_n^* \) converges to \( m \) in \( P_{\beta_n} \)-probability, so that \( P_{\beta_n}(n(\hat{\beta}_L - \beta_n)/\lambda_n^* \in d\mathcal{M}) \to 0 \).
References

Alliney S., Ruzinsky A. (1994). ‘An Algorithm for the Minimization of Mixed $l_1$ and $l_2$ Norms with Applications to Bayesian Estimation’. IEEE Transactions on Signal Processing 42:618–627.

Berk R., et al. (2013). ‘Valid post-selection inference’. Annals of Statistics 41:802–837.

Caner M., Kock A. B. (2014). ‘Asymptotically Honest Confidence Regions for High Dimensional Parameters by the Desparsified Conservative Lasso’. Tech. Rep. 1410.4208, arXiv.

Efron B., et al. (2004). ‘Least Angle Regression’. Annals of Statistics 32:407–499.

Geyer C. (1996). ‘On the Asymptotics of Convex Stochastic Optimization’. Unpublished manuscript.

Javanmard A., Montanari A. (2014). ‘Confidence Intervals and Hypothesis Testing for High-Dimensional Regression’. Journal of Machine Learning Research 15:2869–2909.

Knight K., Fu W. (2000). ‘Asymptotics of Lasso-Type Estimators’. Annals of Statistics 28:1356–1378.

Lee J. D., et al. (2013). ‘Exact Post-Selection Inference with the Lasso’. Tech. Rep. 1311.6238, arXiv.

Pötscher B. M., Leeb H. (2009). ‘On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding’. Journal of Multivariate Analysis 100:2065–2082.

Pötscher B. M., Schneider U. (2010). ‘Confidence Sets Based on Penalized Maximum Likelihood Estimators in Gaussian Regression’. Electronic Journal of Statistics 4:334–360.

Pötscher B. M., Schneider U. (2011). ‘Distributional Results for Thresholding Estimators in High-Dimensional Gaussian Regression Models’. Electronic Journal of Statistics 5:1876–1934.

Rosset S., Zhu J. (2007). ‘Piecewise Linear Regularized Solution Paths’. Annals of Statistics 35:1012–1030.

Schneider U. (2015). ‘Confidence Sets Based on Thresholding Estimators in High-Dimensional Gaussian Regression’. Econometric Reviews to appear.

Tibshirani R. (1996). ‘Regression Shrinkage and Selection via the Lasso’. Journal of the Royal Statistical Society Series B 58:267–288.

Van de Geer S. (2015). ‘$\chi^2$-Confidence Sets in High-Dimensional Regression’. Tech. Rep. 1502.07131, arXiv.

Van de Geer S., et al. (2014). ‘On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models’. Annals of Statistics 42:1166–1202.

Yuan M., Lin Y. (2007). ‘On the Non-negative Garrotte Estimator’. Journal of the Royal Statistical Society Series B 69:143–161.

Zhang C., Zhang S. S. (2014). ‘Confidence Intervals for Low Dimensional Parameters’. Journal of the Royal Statistical Society Series B 76:217–242.

Zhao P., Yu B. (2006). ‘On Model Selection Consistency of Lasso’. Journal of Machine Learning Research 7:2541–2563.

Zou H. (2006). ‘The Adaptive Lasso and Its Oracle Properties’. Journal of the American Statistical Association 101:1418–1429.