TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

DISSERTATION

# Understanding the Lasso: Distribution, Model Selection Properties and Confidence Sets

Ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Assoc. Prof. Ulrike Schneider
E105
Institut für Stochastik und Wirtschaftsmathematik

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Karl Ewald

Matr.Nr. 0648596
[redacted]

Wien, am 10.5.2021

"The number of engagements that go on seems to me considerably above the proper average that statistics have laid down for our guidance."

Oscar Wilde

# Contents

# List of Figures

# List of Tables

## 0.1 Acknowledgements

The author would like to thank everyone who helped or supported him in the creation of this thesis. In particular, he wants to extend his sincerest thanks to his supervisor, Ulrike Schneider, for her guidance and support as well as for the occasional push. He would also like to thank the colleagues from his research group at TU Vienna as well as his colleagues from University of Vienna's Institute of Statistics and Decision support for their stimulating exchange of ideas and company. Finally, the author thanks his friends and family for their backing and encouragement.

The author also gratefully acknowledges support from DFG grant FOR 916.

## 0.2 Abstract

Within this thesis we analyze the Lasso estimator and its properties, such as its distribution as well as methods of inference that are based on a Lasso estimate.

In the first chapter we give a lower bound for the coverage probability of the Lasso's estimation error for a large class of sets that satisfy a rather weak condition which depends on the design matrix. This enables the construction of uniformly valid confidence sets for the entire parameter vector based on a Lasso estimate in finite samples as well as in an asymptotic setup where the estimator is tuned to perform conservative model selection. Additionally, we give an asymptotic probability one confidence set in the case where the Lasso is tuned to perform consistent model selection.

The second chapter deals with the estimator's properties, in particular providing its distribution in high and low dimensions. In the low-dimensional case the distribution is given explicitly by the cumulative distribution function, but can also be specified by a number of conditional densities (given the corresponding active sets). We also describe the unique relationship between the Lasso and the Least-squares estimator. We additionally give insight into the estimator's model selection properties and in particular show that in a high-dimensional setting the estimator may not select certain variables at all, independent of the response.

The final chapter again turns to the topic of Lasso-based confidence sets, extending the concepts developed in the first chapter: We consider the unknown variance case and explore the question of how to construct uniformly valid confidence sets for single components. The latter case is also considered in the case where a partial Lasso is applied (i.e., if some components are not penalized at all). We also investigate the validity of a procedure that is designed only to cover the estimator's non-zero components in an optimal way, meaning that the choice of the confidence set's shape depends on the selected model. Finally, we consider an adaptive procedure, where the confidence set's shape is optimized for a given sign of the parameter. Hereby the sign is estimated by a conservative estimation procedure. However, it is shown that this approach does not yield uniformly valid confidence sets.

## 0.3 Deutsche Kurzfassung

Diese Dissertation befasst sich mit dem Lasso Schätzer und seinen Eigenschaften, insbesondere mit der Verteilung des Schätzers und der Konstruktion von gleichmäßig gültigen Konfidenzmengen (welche auf dem Lasso basieren).

Im ersten Kapitel wird eine Methode entwickelt, um die minimale Überdeckungswahrscheinlichkeit des Schätzfehlers für eine Klasse von Mengen zu bestimmen. Die zulässigen Mengen müssen hierbei lediglich eine recht schwache Bedingung erfüllen, welche von der Regressormatrix abhängt. Mithilfe dieses Resultats werden sodann auf dem Lasso basierende Konfidenzmengen konstruiert. Diese sind sowohl für endliche Stichproben, als auch asymptotisch, im Fall eines konservativ eingestellten Lassos, anwendbar. Für den Fall eines konsistent eingestellten Schätzers wird eine asymptotisch gültige Konfidenzmenge mit Überdeckungswahrscheinlichkeit eins angegeben.

Das zweite Kapitel behandelt die Verteilungseigenschaften des Lasso Schätzers. Hierbei wird zunächst die Verteilungsfunktion des Schätzers hergeleitet. Im niedrigdimensionalen Fall kann die Verteilung aber auch mittels (auf die jeweils aktiven Komponenten) bedingte Dichtefunktionen beschrieben werden. Anschließend wird eine spezielle Beziehung zwischen Lasso und Kleinstquadrateschätzer hergeleitet. In dem Kapitel wird ebenfalls herausgearbeitet, dass der Schätzer in gewissen hochdimensionalen Situationen manche Regressoren nie, das heisst für keinen Wert der abhänigen Variable, auswählt. Diese Menge ist hierbei ausschließlich durch die Regressormatrix, sowie den Penalisierungsvektor bestimmt.

Das dritte Kapitel wendet sich wieder dem Thema der Konfidenzmengen zu und erweitert die vormals entwickelten Konzepte. Hierbei wird zunächst der Fall der unbekannten Fehlervarianz behandelt. Es folgt eine Diskussion der Konstruktion von für einzelne Komponenten optimierte Konfidenzmengen, sowie der Fall des partiellen Lassos, bei dem nicht alle Komponenten penalisiert werden. Es folgt eine Analyse, ob eine adaptive (das heißt, von den aktiven Komponenten abhängige) Wahl der Form der Konfidenzmengen und des zu überdeckenden Sub-Parameters eine valide Prozedur darstellt. Schließlich wird gezeigt, dass die vorherige Schätzung des Vorzeichen des wahren Parameters mit darauf folgender Verwendung dieses Ergebnisses in der Konstruktion von Konfidenzmengen keine asymptotisch korrekten Konfidenzmengen liefert.

## A note on the use of the word “we”

When delving into the world of academic literature one may often find oneself surprised at the use of the word “we”. While it is perfectly logical to use phrases such as “we show that . . . ” or “we see that . . . ” for articles with multiple authors, it is also common for single authors to use this structure. This might seem strange to some readers and while one may argue that the intense work on certain mathematical problems does encourage certain states of mind in which one might be unsure of some basic facts of reality, this thesis’ author uses the word “we” to refer to himself together with the reader. In other words, the author intends to include the reader in the sense of guiding them through the article.

# Chapter 1

# Introduction

Throughout the ages there have been transitions to new eras of human history which were mostly driven by advances in science and technology. New inventions and technologies, such as the inventions of the wheel, pottery, writing, metal works and so on, have drastically changed the way humans live, think and work. One of the most notable changes was without doubt the industrial revolution at the beginning of the 20th century. Of course, the development did not stop there. Quite to the contrary, where in recent decades we have seen the transition into a new era commonly referred to as the "digital revolution". The rapidly advancing digitalization of science as well as everyday life poses a challenge to the field of statistics in particular. This is due to the wide availability of vast amounts of data which are being generated and collected in all sorts of settings, be it website usage, shopping behavior, traffic data... (the completion of this list is left to the reader's imagination). Given the availability of this so-called "big data", it is only natural that scientists as well as governments and coperations want to make good use of the available information. However, every coin has two sides and the blessing of vast amounts of data being available for many applications can be a curse in situations where existing methods may not be appropriate, or even cease to be feasible at all.

In the context of linear regression models, the huge amount of possible explanatory variables poses a challenge to users who are often interested in rather simple and easily interpretable models. Since one usually also does not want to choose one's model completely arbitrarily, numerous so-called model selection procedures, such as evaluating each possible sub-model[1] by some selection-, or fit-criterion, like the *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC), or the iterative addition of variables that decrease one of these criteria the most[2], are employed to enable users to choose their working models in reasonable, data-driven ways. The above-mentioned procedures, however, come at great computational cost, making their application

[1] Usually referred to as *Full subset selection*.
[2] Usually referred to as *Forward selection*.

infeasible if the number of possible models is large[3]. Moreover, the presence of a model selection step prior to parameter estimation has been shown to have adverse effects on classical inferential procedures that do not take the presence of model selection into account. Indeed, the actual coverage probability[4] of these "naive" confidence sets has been shown to fall well below their nominal levels in such situations (see, for example, Pötscher, 1991 or Kabaila & Leeb, 2006) thus further complicating the procedure of estimation and conducting inference.

The belief that so-called *sparsity* may be present in many models, i.e., that the response may only depend on a few of many, or even infinitely many, possible regressors has lead to a rising interest in so-called shrinkage estimators. As the name suggests, these estimators shrink the Least-squares estimate towards, and some components even exactly to zero, thus providing model selection as well as parameter estimation in a single step. One quite notable example of these estimators is the *Least Absolute Shrinkage and Selection Operator* which is better known by its acronym, the *Lasso estimator*. It was introduced in Tibshirani (1996) and is defined as the solution to an $l_1$-penalized version of the Least-squares objective function. The procedure provides many desirable properties, such as low computational cost[5] as well as the ability to still yield unique[6] solutions. This holds even in high-dimensional settings, in which the number of explanatory variables $p$ is larger than the number of observations $n$. In such cases the Least-squares estimator, for example, is no longer unique. Not all properties of this procedure are, however, fully understood yet, in particular and probably most importantly, its distribution.

Understanding how the Lasso behaves both as an estimator as well as a model selection method is a vital piece of information a user should consider, or at least be aware of, when applying the procedure. For a wide range of applications it will be of great importance to know about the precision of the estimates obtained from using this procedure. This leads to the question of how confidence regions can be constructed based on the Lasso estimator and how their size can be kept as small as possible. At this point the issue is best divided into the high- and low-dimensional frameworks, respectively, where the number of explanatory variables either exceeds the number of observations or not. For the former case more recent contributions to the academic literature (e.g. Zhang & Zhang, 2014; Van de Geer et al., 2014) have proposed confidence sets that rely on "de-biasing" the Lasso. These procedures, however, turn out to be essentially equivalent to using the Least-squares confidence sets if applied in a low-dimensional framework.

In recent years a number of contributions to academic literature have analyzed the topic of *post-selection inference*, a concept that was first proposed by Berk et al. (2013) and aims at constructing confidence sets that seek to cover a pseudo-true value that is referred to as *the parameter given*

[3]Indeed, for a Full subset selection, there are $2^p$ possible submodels for $p$ available explanatory variables.
[4]At least when considering the classical target for inference.
[5]For more details on this see for instance Alliney & Ruzinsky (1994); Efron et al. (2004) and Rosset & Zhu (2007).
[6]Note, however, that even the Lasso estimator is not always unique, as noted, for example, in Tibshirani (2013), Ewald & Schneider (2020) and Chapter 3 of this thesis.

*the model*[7]. The articles that specifically deal with such confidence sets in relation with the Lasso estimator include Lee et al. (2016); Tibshirani et al. (2016); Meir & Drton (2017); Zhao et al. (2017); Kivaranovic & Leeb (2018); Tibshirani et al. (2018); Liu et al. (2018); Zhou et al. (2019) and Min & Zhou (2019).

However, rather little research has been conducted on the coverage of the *true* parameter in a low-dimensional setting. Pötscher & Schneider (2010) have given the optimal[8] choice of confidence interval in an orthogonal regressor setting. This article gives insights into some properties these confidence sets will have, such as the fact that the Lasso-based confidence intervals will have to be larger than the confidence intervals that are based on the Least-squares estimator. However, the "classical" low-dimensional case with correlated regressors has remained a gap in the literature so far. This thesis' first major component will be to close this gap.

Turning to the issue of the distribution, Pötscher & Leeb (2009) have provided the distribution of the Lasso in the orthogonal-regressor (low-dimensional) setup, while Rosset & Zhu (2007); Tibshirani (2013) and Zhou (2014) have all provided valuable information on the properties of the estimator. Quite notably, Zhou (2014) has, in a way, produced the distribution of the Lasso by analyzing the *augmented Lasso* which enables the reader to recover the distribution of the actual estimator. However and despite all these contributions, the existing literature has not yet provided a comprehensive and easy to grasp picture of the estimator's behavior, i.e., its distribution, and this shall be the second major component of this thesis. To that end, the cumulative distribution funcion (cdf) of the Lasso is provided for low-dimensional setups and a new characterization of the estimator's distribution is provided for high-dimensional frameworks. Also, the distribution is presented in more intuitive forms, for example in terms of conditional densities on certain submodels (so-called active sets). Moreover, facts about the model selection properties in high-dimensional settings as well as the Lasso's relationship to the Least-squares estimator in low-dimensional settings are provided. Finally, this thesis gives some insights into the Lasso estimator's model selection properties while also touching upon the subject of its uniqueness, a topic that has been covered by Tibshirani (2013), Ali & Tibshirani (2019) and Ewald & Schneider (2020).

## 1.1 Contents and publications

To deal with the issues described above, this thesis is organized into three parts: Chapter 2 gives a procedure on how to conduct inference on the true parameter based on a Lasso estimate in a low-dimensional setup, i.e., how to construct confidence sets that are based on the Lasso estimator. Chapter 3 is dedicated to the distribution of the estimator, lending greater understanding of how this procedure behaves both as an estimator as well as a model selection procedure. Finally, Chapter

[7] For a brief description refer to Section 2.2.
[8] In terms of component-wise length.

4 discusses the question of how to use and extend the theory developed in Chapter 2 to a number of different settings and applications.

The work within this thesis has been published within two papers. Ewald & Schneider (2018) deals with the topic of confidence sets that are based on the Lasso estimator and the results presented there are based on Chapter 2 as well as sections 4.2 and 4.3. Ewald & Schneider (2020) is largely based on Chapter 3, but also contains additional results regarding the Lasso's uniqueness.

## 1.2 General setting and notation

Throughout the thesis we will consider the following linear regression model:

$$y = X\beta + \varepsilon, \tag{1.1}$$

where $y$ is an (observed) $n \times 1$ response vector, $X$ an $n \times p$ (non-random) regressor matrix with rows $x_{i\cdot}$ and columns $x_{\cdot j}$. Moreover, $\beta$ denotes the unknown parameter we are interested in and $\varepsilon$ is an $n \times 1$ error vector that is defined on some probability space $(\Omega, \mathscr{A}, P)$. The error, $\varepsilon$, is assumed to be zero in expectation and its variance-covariance matrix is of the form $\sigma^2 I_n$ with $\sigma > 0$ and $I_n$ denoting the $n$-dimensional identity matrix.

As several different settings will be discussed in this thesis, such as high- and low-dimensional models, or finite-sample as well as asymptotic results, assumptions on $X$, $\beta$ and $\varepsilon$ will differ between the sections and are thus not discussed at this point. Note that in the settings considered in this thesis, the parameter as well as the other quantities will depend on the sample size $n$. Since the focus of the thesis lies mostly on finite-sample results, we shall suppress this dependence in the notation for the most part.

The (*weighted* or *generalized*) Lasso estimator, $\hat{\beta}_{\mathrm{L}}$, is defined as the solution to the Lasso objective function, a penalized version of the Least-squares objective function:

$$\hat{\beta}_{\mathrm{L}} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} L(\beta),$$

where the Lasso objective function, $L(\beta)$, is given by

$$L(\beta) = \|y - X\beta\|_2^2 + 2\sum_{j=1}^{p} \lambda_j |\beta_j|, \tag{1.2}$$

with $\lambda_j \geq 0$ being component-specific penalization weights which usually depend on the sample size $n$. For later use, we shall denote by $\lambda$ the vector of penalization weights[9], i.e., $\lambda = (\lambda_1, \ldots, \lambda_p)'$ and by $\Lambda$ the diagonal matrix containing $\lambda$ as its diagonal: $\Lambda = \mathrm{diag}(\lambda)$. Note that this definition

[9] This vector is sometimes also referred to as *tuning vector*.

of the Lasso includes the Least-squares estimator[10] in the sense that these estimators are equal when $\lambda_j = 0$ for each $j \in \{1, \dots, p\}$. Also note that the *classical* Lasso estimator, as proposed by Tibshirani (1996), is defined with equal penalization-weights $\lambda_j$. However, it may prove beneficial to users to choose those weights differently. This may be to, for example, account for different interest in the effects of certain regressors determined by the application at hand. A user may, for example, choose not to penalize some components, considered to be of particular interest, at all, thus excluding them from model selection[11]. This special case will be referred to as the *partial* Lasso estimator. The main focus of the analysis will, however, be the case in which all components underly some level of penalization and despite the theory encompassing the partial case, we shall mostly think of the penalization weights as being strictly greater than zero. Also note that the factor two in front of the penalization term in (1.2) is completely arbitrary and could be integrated into the penalization vector $\lambda$. However, defining the problem in this way will simplify some of the formulae later on.

Next, we define the so-called *active set* of the Lasso estimator as

$$\mathcal{A} = \mathcal{A}(\hat{\beta}_{\mathrm{L}}) = \{j \in \{1, \dots, p\} : \hat{\beta}_{\mathrm{L},j} \neq 0\},$$

i.e., all components of the estimator that are non-zero. Note that this active set can also be viewed as the model that is selected by the Lasso when thinking of the estimator as a model selection procedure.

To analyze the estimator, we will for the most part consider the Lasso's estimation error, $\hat{u} = \hat{\beta}_{\mathrm{L}} - \beta$, by looking at a re-parameterized version of the Lasso objective function. Setting as short-hand notation $C = X'X$ and $W = X'\varepsilon$, we define

$$\begin{aligned} V(u) &= L(u + \beta) - L(\beta). \\ &= u'Cu - u'W + 2\sum_{j=1}^{p} \lambda_j \left[|u_j + \beta_j| - |\beta_j|\right]. \end{aligned} \tag{1.3}$$

Note that $V$ is minimized at $\hat{u}$.

In case $C$ is invertible, the Least-squares estimator is given by $\hat{\beta}_{\mathrm{LS}} = (X'X)^{-1}X'y$ and its corresponding estimation error equals $\hat{u}_{\mathrm{LS}} = \hat{\beta}_{\mathrm{LS}} - \beta$. Next, we will use $\iota$ for the vector containing 1's in all components. $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ is used to denote the extended real line and $\mathbb{1}_{\{\cdot\}}$ for the indicator function. The sup-norm on $\mathbb{R}^p$ is denoted by $\|\cdot\|_\infty$. Let $e_j$ denote the $j$-th unit vector in $\mathbb{R}^p$ and let $\operatorname{sgn}(\cdot)$ denote the sign function with the function being defined component-wise, if applied to vectors. The empty set and Carthesian product are denoted by $\varnothing$ and $\Pi$, respectively.

[10] That is $\arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2$.
[11] To see that this is actually the case, see Chapter 3.

For a set $B \subseteq \mathbb{R}^p$ and $a \in \mathbb{R}^p$, we define

$$a + B = \{a + b : b \in B\},$$

and for any matrix $Z \in \mathbb{R}^{p \times p}$, we define

$$ZB = \{Zb : b \in B\}$$

with an analogous definition for set-multiplication by a scalar.

For any index set $\mathcal{I} \subseteq \{1, \ldots, p\}$, and any vector $v \in \mathbb{R}^p$ we define $v_{\mathcal{I}}$ to be the $|\mathcal{I}|$-dimensional sub-vector of components with indices contained in $\mathcal{I}$. Similarly, for a matrix $T \in \mathbb{R}^{n \times p}$, let $T_{\mathcal{I}}$ denote the $n \times |\mathcal{I}|$ sub-matrix consisting of the columns of $T$ with indices contained in $\mathcal{I}$. We denote the matrix' column-space, kernel and rank by $\text{span}(Z)$, $\ker(Z)$ and $\text{rank}(Z)$, respectively. Also, for $Z \in \mathbb{R}^{\{p \times p\}}$, let $\det(Z)$ denote the matrix' determinant.

For $d \in \{-1, 1\}^p$ let $\mathcal{O}^d = \{z \in \mathbb{R}^p : d_j z_j \geq 0 \quad \forall j \in \{1, \ldots, p\}\}$ denote the corresponding orthant of $\mathbb{R}^p$. By $\mathcal{O}^d_{\text{int}}$ we denote the orthant with strictly positive components only, that is, $\mathcal{O}^d_{\text{int}} = \{z \in \mathbb{R}^p : d_j z_j > 0\}$. And for $d \in \{1, -1, 0\}^p$, let $\mathcal{Q}^d = \{z \in \mathbb{R}^p : \text{sgn}(z_j) = d_j \quad \forall j \in \{1, \ldots, p\}\}$.

Let $\phi_{(\mu,\Sigma)}$ and $\Phi_{(\mu,\Sigma)}$ denote the probability density function (pdf) and cumulative distribution function (cdf) of a random variable following a normal distribution with mean $\mu$ and variance-covariance matrix $\Sigma$. By $\phi$ and $\Phi$ we denote the pdf and cdf of a standard-normal random variable.

Finally, note that the directional derivative of a function $g : \mathbb{R}^p \to \mathbb{R}$ at a point $t \in \mathbb{R}^p$ in direction $r \in \mathbb{R}^p$ with $\|r\|_2 = 1$ is defined as and denoted by

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

Having laid out the general notation we can now proceed to the thesis' first topic which is confidence sets.

# Chapter 2

# Confidence sets based on the Lasso estimator

## 2.1 Introduction

The first chapter of the thesis deals with the topic of constructing confidence sets based on the Lasso estimator. Considering the low-dimensional ($p \leq n$) framework, Pötscher & Schneider (2010) revealed implications of the Lasso estimator's distributional properties for confidence sets in orthogonal regressor settings. In particular they provide formulae for calculating the minimal coverage probabilities as well as the optimal[1] choice of a Lasso-based confidence set in such a case. Generalizations to a moderate-dimensional setting where $p \leq n$ but $p$ diverging with $n$ are contained in Pötscher & Schneider (2011) and Schneider (2016). The former reference also contains a way of constructing asymptotic probability one confidence sets in the orthogonal regressor setting when the Lasso is tuned to perform consistent model selection. A similar approach has been taken by Amann & Schneider (2018) to obtain such confidence sets for the adaptive Lasso[2] without assuming orthogonality of the regressors.

In a high-dimensional setting with $p \gg n$, confidence regions and confidence intervals in connection with the Lasso estimator have been treated in a number of papers including Zhang & Zhang (2014); Van de Geer et al. (2014); Javanmard & Montanari (2014a,b); Van de Geer & Stucky (2016); Van de Geer (2017); Cai & Guo (2017) and Caner & Kock (2018). 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 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 in order to provide uncertainty quantification for the Lasso estimator in this case.

Dealing with a related topic, Lockhart et al. (2014) and Tian & Taylor (2017) consider signifi-

[1] With respect to length.
[2] For details on that procedure see Zou (2006).

cance tests for additional variables within a Lasso path.

Returning to the topic of confidence regions, Lee et al. (2016) 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[3] 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. (2013) who discuss an intricate procedure for obtaining confidence regions for this pseudo-true parameter after a model selection step. Articles that deal with this so-called *post-selection inference* (PoSI), after a Lasso-selection step include Lee et al. (2016); Tibshirani et al. (2016); Meir & Drton (2017); Zhao et al. (2017); Kivaranovic & Leeb (2018); Tibshirani et al. (2018); Liu et al. (2018); Zhou et al. (2019) and Min & Zhou (2019). Finally, in a slightly different setting of a mean model with independent Gaussian errors Hyun et al. (2016) propose tools for inference that are based on a generalization of the Lasso estimator.

This chapter's goal is to close the gap between the afore-mentioned results by providing confidence sets that are based on the Lasso estimator for the entire parameter vector in general low-dimensional regression settings without requiring any restrictive assumptions on either the regressor matrix, $X$, or the true parameter vector, $\beta$.

As the Lasso estimator's distribution depends on the true parameter (c.f. Pötscher & Leeb, 2009, or Chapter 3) which is, by definition unknown, we will have to construct confidence sets whose coverage probabilities cover the entire[4] true parameter *for each of its possible values* with (at least) our desired nominal coverage probability. Hence, we need to know the smallest coverage probability over the whole parameter space for a given prospective confidence set in order to construct valid confidence regions. Because the Lasso estimator's distribution's dependence on the true parameter is rather complicated (c.f. Chapter 3), the task is not straightforward. Also, this problem cannot be overcome by considering large samples, as this dependence does not vanish in such settings, at least when considering moving-parameter frameworks in which the finite-sample distribution is approximated best. Aside from answering the question of how to calculate minimal coverage probabilities for given sets we will have to ponder on how to choose a confidence set's shape in a reasonable way: While it seems straightforward to use an interval in the one-dimensional case, the matter becomes more intricate when moving to higher dimensions. In this case one may prioritize certain components over others in the sense that the marginal coverage-probabilities of a confidence set may differ, or choose to optimize the confidence sets' shape with respect to its volume. In case the estimation error's distribution is independent of the true parameter and has a Lebesgue density it is easily seen that the volume can be minimized[5] by taking an appropriate

[3] For a brief discussion of this concept, see, for example, Section 2.2.

[4] I.e., all of its components simultaneously.

[5] Indeed, to minimize the confidence set's volume, one will want to add those regions to the set that contain the true parameter with the highest probability. Clearly, these areas can be identified by searching for those areas where

contour set of the pdf and centering it at the point-estimator. While this is the case for the Least-squares estimator, for example, where the volume-optimized confidence sets turn out to be elliptic, the situation is quite different for the Lasso estimator in case of which the distribution is neither absolutely continuous, nor independent of the true parameter and hence, the task of finding (near-) optimal[6] shapes requires a bit more thought. Note that even though both finite-sample, as well as asymptotic distributions of the Lasso are known explicitly (c.f. Chapter 3), this knowledge is not used in the development of the procedure in this chapter. This is due to, on the one hand, historic reasons, as the results that are presented in the first chapter precede the ones on the estimator's distribution and, on the other hand, due to the quite complicated dependence and non-standard form of the Lasso's distribution. However, it is not even necessary to know the full distribution, as we provide a sharp lower bound for a set's coverage probability using only properties of the underlying minimization problem.

Essentially, in this chapter we do the following: It is shown that the minimal coverage probability of a set satisfying some rather mild conditions occurs when the true parameter's components are all large in absolute value. Indeed, it is shown that the minimal coverage probability[7] can be calculated by essentially deferring the minimization into the objective function that defines the estimation error. We in effect "bound"[8] the area the true estimation error can lie in by the minimizers of the objective functions when the components of the true parameter are large in absolute value with the different (stochastic) "bounding" functions only depending on the unknown parameter via its components' signs. The minimizers of these "bounding" functions turn out to have quite well-behaved distributions, thus allowing us to obtain an explicit formula for the coverage probability of a large class of sets that satisfy a condition depending on the regressor matrix.

The class of sets which satisfy the necessary requirements encompasses the elliptic shape one would use if the confidence region was based on the Least-squares estimator, thus enabling comparisons with the Least-squares confidence ellipse. In analogy 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 by the results in Pötscher & Schneider (2010), our finite-sample results 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 $\frac{1}{\sqrt{n}}$ 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 an asymptotic probability one confidence set

the pdf takes on the largest values.

[6] With respect to size.

[7] For a given sign of the true parameter.

[8] Note that these bounds are constructed for a given $\omega \in \Omega$ and are thus themselves stochastic.

in that case.

The remaining chapter is organized as follows. In Section 2.2 we will briefly discuss possible coverage targets, as a further interesting option (other than the "classical" true parameter) has been the subject of recent academic discussions. In Section 2.3 we set the framework by re-stating the model, stating the assumptions used in this chapter as well as introducing some notation. The main result providing the formula for the minimal coverage probability is presented in Section 2.4 and subsequently Section 2.5 is devoted to discussing how to concretely construct the corresponding confidence sets, as well as their relationship to the confidence ellipse based on the Least-squares estimator. In Section 2.6 we derive asymptotic results for both the cases of conservative and consistent model selection. Section 2.7 concludes.

## 2.2 On inference after model selection and coverage targets

In this section we discuss the choice of target for the confidence sets. Given the above setup it may at first seem strange that one may be interested in anything but the true parameter $\beta$ in the model (1.1) (or a sub-parameter thereof). However, the topic of inference after model selection turns out not to be trivial at all and we will thus shortly discuss the topic. This discussion will show that also the choice of coverage target may not be as obvious as it may seem initially either.

In practice users often neglect to take effects of model selection on inferential procedures into account, and as a result use "naive" tests and confidence sets. The effects of this are potentially disastrous in terms of actual parameter-coverage in the context of confidence sets (see, for example Pötscher, 1991 or Kabaila & Leeb, 2006) and type-1 error in the context of testing[9] (c.f. Taylor & Tibshirani, 2015), respectively. Hereby several effects are at play.

The first problem lies in model misspecification, as any of the true model's left-out variables that are correlated with some variable contained in selected model will cause a bias in the estimation. This in turn leads to the confidence sets being centered at an incorrect (i.e., biased) estimate, a fact that is not taken into account by naive procedures.

The second issue is the selection procedure's own stochastic properties. As certain models will be selected based on the response $y$ and hence the random error $\varepsilon$, the traditional assumptions used to construct tests and confidence sets are severely violated. Indeed, both a variable's selection probability as well as a test's rejection probability will typically depend on the estimated effect size and are thus far from independent. As a consequence, the conditional probability of falsely rejecting the null-hypothesis[10] given that a variable has been selected by a procedure is inflated. This leads to an overall inflation of the type-1 error, a fact that has been pointed out by Taylor & Tibshirani (2015).

[9] Note that while in this thesis we only discuss confidence sets, the two concepts are closely related as a confidence set may be obtained by inverting a test and the other way around.

[10] Of there being no effect for the variable, i.e., $\beta_j = 0$.

Given these properties, one may be tempted to conclude that any meaningful kind of inference is nearly impossible in the presence of large numbers of available explanatory variables, as a test's power will be very low[11] and component-wise confidence intervals quite long[12] in the largest possible model. On the other hand, the exclusion of important variables within a model selection step will render the resulting naive confidence sets and tests unusable.

However, a new view on the matter has arisen which eradicates at least the problems induced by misspecification biases. To cite George Box in saying that *"all models are wrong, but some are useful"*, Berk et al. (2013) argue that, since there is no way of determining whether one's working model, however large, is correctly specified in the sense that all regressors of the true model (should there even be one) are contained in the working model, the coverage target should not be the true parameter from the unknown data-generating model, but the *parameter given the model*. This means that the coefficients' interpretation depends on the selected model. More specifically, let the index set $\mathcal{I} \subseteq \{1, \dots, p\}$ represent the selection of regressors in a submodel. If the corresponding sub-matrix containing only those regressors, $X_\mathcal{I}$, has full rank, then one may define the so-called *post-selection inference* (PoSI) target as

$$\tilde{\beta}_\mathcal{I} = (X_\mathcal{I}' X_\mathcal{I})^{-1} X_\mathcal{I}' X \beta.$$

Note that in general $\tilde{\beta}_\mathcal{I} \neq \beta_\mathcal{I}$, the sub-vector containing the corresponding selection of components of the parameter vector in the true model. (Clearly also $\tilde{\beta}_\mathcal{I} \neq \tilde{\beta}_\mathcal{J}$ for $\mathcal{I} \neq \mathcal{J}$ in general.) The quantity $\tilde{\beta}_\mathcal{I}$ is the true parameter in the model

$$y = X_\mathcal{I} \tilde{\beta}_\mathcal{I} + \epsilon$$

with error-term $\epsilon$ such that $E(\epsilon) = 0$, in the sense that $X_\mathcal{I} \tilde{\beta}_\mathcal{I}$ is the expectation of the projection of $y$ onto the space spanned by the regressors that are contained in the set $\mathcal{I}$. Note that in general $\epsilon$ differs from $\varepsilon$, the error in the true model.

Conducting inference after model selection for this type of coverage target has been shown to yield much better results in terms minimal coverage probability when using naive procedures to construct confidence sets (c.f. Leeb et al., 2015), at least in simple settings. On the other hand, a method of constructing honest confidence sets that accounts for the presence of *any* kind of model selection has been proposed in Berk et al. (2013). However, one would expect methods that are tailored to the specific kind of model selection to be more efficient[13]. Thus the idea of covering the PoSI target has been taken up by a number of authors. Contributions that deal specifically with PoSI confidence sets after a Lasso-based model selection step include the afore-mentioned Lee et al. (2016); Tibshirani et al. (2016); Meir & Drton (2017); Zhao et al. (2017); Kivaranovic & Leeb

[11] If corrected for the global type-1 error.
[12] When covering the entire parameter vector, i.e., all components simultaneously.
[13] In the sense that they yield smaller confidence sets, or more powerful tests.

(2018) and Min & Zhou (2019). Tibshirani et al. (2018) provide a test for a PoSI target after a Lasso-, Forward-, or Least Angle Regression selection step.

While the topic of covering the PoSI target has been quite actively researched, the one of inference after model selection for the "classic" coverage target $\beta$ after a Lasso-selection has not been covered in this much detail[14] with notable contributions being Taylor & Tibshirani (2015) and Lockhart et al. (2014) who have provided a significance test for each additional variable entering a Lasso path. The latter concept was later generalized to so-called affine selection procedures (which include the Lasso) while also removing the assumption of the errors' Gaussianity, see Tian & Taylor (2017). Liu et al. (2018) discuss the problem of inference after model selection for both[15] kinds of target.

In this thesis we consider the case of confidence sets that are based on the Lasso estimator and are designed to cover the true parameter vector, or parts thereof, thus dealing with both of the above-mentioned issues of inference after model selection, i.e., parameter misspecification in the working model and accounting for the selection procedure's stochastic properties.

## 2.3 Assumptions

Recall model (1.1):

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

We now assume that $X$, the $n \times p$ regressor matrix has full column-rank $p$ and $\varepsilon$, the unobserved error term, consists of independent and identically distributed components with mean zero and finite variance $\sigma^2 > 0$.

## 2.4 Finite-sample results

As mentioned in this chapter's introduction, we aim to construct confidence sets for the entire parameter vector $\beta$ based on the Lasso estimator $\hat{\beta}_{\mathrm{L}}$. More formally, this means that for a non-random set $M \subseteq \mathbb{R}^p$, we consider sets of the form

$$\hat{\beta}_{\mathrm{L}} - M = \{\hat{\beta}_{\mathrm{L}} - m : m \in M\}$$

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 $P_\beta(\beta \in \hat{\beta}_{\mathrm{L}} - M) \geq 1 - \alpha$ for all $\beta \in \mathbb{R}^p$ (where we stress the dependence of the probability measure

[14] At least in low-dimensional settings; also see Section 2.1.
[15] The full (true) parameter as well as the PoSI target.

on $\beta$ whenever it occurs), so that

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

In order to achieve this, we need to be able to compute this "infimal" (minimal) coverage probability. For our finite-sample considerations in this and the following subsections we suppose that the errors follow a normal distribution

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

an assumption that will be removed for asymptotic results in Section 2.6. 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 all components of the unknown parameter vector $\beta$ 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 Least-squares estimator and by construction do not depend on the unknown parameter. Finally, we will show that the minimal coverage probability of the proposed sets is indeed "achieved" for one of these finitely many limiting cases.

To state the main theorem recall the re-parametrized objective function $V$ which is uniquely minimized at the estimation error, $\hat{u} = (\hat{\beta}_{\mathrm{L}} - \beta)$. Also note that $V$ depends on $\beta$, despite this being suppressed in the notation.

$$V(u) = u'Cu - 2u'W + 2\sum_{j=1}^{p} \lambda_j \left[ |u_j + \beta_j| - |\beta_j| \right].$$

Further note that for a set $M \subseteq \mathbb{R}^p$ we have

$$P_\beta(\beta \in \hat{\beta}_{\mathrm{L}} - M) = P_\beta(\hat{u} \in M).$$

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

$$V^d(u) = u'Cu - 2u'W + 2\sum_{j=1}^{p} \lambda_j d_j u_j, \tag{2.1}$$

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

$$V^d(u) = \lim_{\substack{d_j \beta_j \to \infty \\ j=1,\ldots,p}} V(u).$$

As short-hand notation, we write $\hat{u}^d$ for the unique minimizer of $V^d$.

We commence the analysis of the estimator by proving a result quantifying the maximal distance between the Lasso and the Least-squares estimator in finite samples.

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

$$\left|(C(\hat{\beta}_{L} - \hat{\beta}_{LS}))_j\right| \leq \lambda_j.$$

*Proof.* We have $W = X'\varepsilon = C\hat{u}_{\text{LS}}$ where $\hat{u}_{\text{LS}} = \hat{\beta}_{\text{LS}} - \beta$. Consider the directional derivative of $V$ at its minimizer $\hat{u}$ in the direction of $e_j$ and $(-e_j)$, the (negative) $j$-th unit vector. We have

$$\begin{aligned} 0 \leq \frac{\partial}{\partial e_j} V(\hat{u}) &= 2(C\hat{u})_j - 2W_j + 2\lambda_j \left[\mathbb{1}_{\{\hat{u}_j \geq -\beta_j\}} - \mathbb{1}_{\{\hat{u}_j < -\beta_j\}}\right] \\ &\leq 2(C\hat{u})_j - 2(C\hat{u}_{\text{LS}})_j + 2\lambda_j, \end{aligned}$$

as well as

$$\begin{aligned} 0 \leq \frac{\partial}{\partial (-e_j)} V(\hat{u}) &= -2(C\hat{u})_j + 2W_j + 2\lambda_j \left[\mathbb{1}_{\{\hat{u}_j \leq -\beta_j\}} - \mathbb{1}_{\{\hat{u}_j > -\beta_j\}}\right] \\ &\leq -2(C\hat{u})_j + 2(C\hat{u}_{\text{LS}})_j + 2\lambda_j. \end{aligned}$$

Piecing the two displays' inequalities together yields the result. □

While primarily serving as a technical vehicle in this chapter, Proposition 1 shows an interesting fact about the relationship of the Lasso and the Least-squares estimator, namely that the distance between the two is bounded by a parallelogram whose size is determined by the size of the penalization vector's components, $\lambda_j$. This will be further explored in Chapter 3.

Next, we consider the case in which the sign of the true parameter $\beta$ is known. In reality, this may be the case due to prior knowledge about an estimation problem at hand, or follow from physical constraints, for example. However, one may also simply view this assumption as a vehicle that gets us part of the way towards a more general result[16]. To simplify things further, we assume, without loss of generality[17] that $\beta \in \mathcal{O}^\iota$, i.e., that all components of the true parameter are non-negative.

The following proposition gives a way of calculating the minimal coverage probability of all sets satisfying certain shape constraints. This proposition, in fact, is the core of most considerations in this chapter. The result hinges on the fact that given the knowledge about the parameter vector's true sign, the Lasso's true estimation error can be shown to lie within a set that is framed by a

[16] In fact, this may prove to be the best view of the assumption, since otherwise, restricting the estimator to the parameter-space considered would seem more natural.

[17] Otherwise, one may recover the general case by flipping the signs of the regressor matrix' corresponding columns.

number of linear equations that depend on $\hat{u}^\iota$ and $C$. In order to define these sets' component-wise restrictions, we first define the following sets. For $m \in \mathbb{R}^p$ and a positive definite $p \times p$ matrix $\bar{C}$, we define

$$\begin{aligned} A^{d_j}_{\bar{C},j}(m) &= \{z \in \mathbb{R}^p : d_j(\bar{C}m)_j \leq d_j(\bar{C}z)_j, d_j z_j \leq 0\} \text{ and} \\ B^{d_j}_{\bar{C},j}(m) &= \{z \in \mathbb{R}^p : (\bar{C}z)_j = (\bar{C}m)_j, d_j z_j > 0\} \end{aligned}$$

for $j = 1, \ldots, p$.

**Proposition 2.** *If $M \subseteq \mathbb{R}^p$ satisfies that*

$$\bigcap_{j=1}^{p} A^{\iota_j}_{C,j}(m) \cup B^{\iota_j}_{C,j}(m) \subseteq M$$

*for all $m \in M$, then*

$$\inf_{\beta \in \mathcal{O}^\iota} P_\beta(\hat{u} \in M) = P(\hat{u}^\iota \in M).$$

*Proof.* We first show that $\inf_{\beta \in \mathcal{O}^\iota} P_\beta(\hat{u} \in M) \geq P(\hat{u}^\iota \in M)$ by showing that for each fixed $\omega \in \Omega$, $\hat{u}^\iota \in M$ implies that $\hat{u} \in M$ as long as $\beta_j \geq 0$ for all $j$. For this, we first show the following two facts.

(a) $(C\hat{u}^\iota)_j \leq (C\hat{u})_j$ for all $j = 1, \ldots, p$.

Suppose there exists a $j_0$ such that $(C\hat{u}^\iota)_{j_0} > (C\hat{u})_{j_0}$ and note that since all partial derivatives of $V^\iota$ must be zero at its minimizer $\hat{u}^\iota$, we have $(C\hat{u}^\iota)_j = W_j - \lambda_j$ for each $j = 1, \ldots, p$ . Now consider the directional derivative of $V$ at its minimizer $\hat{u}$ in direction $e_{j_0}$,

$$\begin{aligned} \frac{\partial V(\hat{u})}{\partial e_{j_0}} &= 2(C\hat{u})_{j_0} - 2W_{j_0} + 2\lambda_{j_0}\left[\mathbb{1}_{\{\hat{u}_{j_0} \geq -\beta_{j_0}\}} - \mathbb{1}_{\{\hat{u}_{j_0} < -\beta_{j_0}\}}\right] \\ &\leq 2(C\hat{u})_{j_0} - 2W_{j_0} + 2\lambda_{j_0} \\ &= 2(C\hat{u})_{j_0} - 2(C\hat{u}^\iota)_{j_0} < 0, \end{aligned}$$

which is a contradiction to $\hat{u}$ minimizing $V$.

(b) $\hat{u}_j > 0$ implies $(C\hat{u})_j = (C\hat{u}^\iota)_j$ for any $1 \leq j \leq p$.

If $\hat{u}_j > 0$ (and hence $\hat{u}_j + \beta_j > 0$ when $\beta_j \geq 0$), then $V$ is partially differentiable at $\hat{u}$ with respect to the $j^{\text{th}}$ component. Therefore, we have

$$\begin{aligned} \frac{\partial V(\hat{u})}{\partial u_j} &= 2(C\hat{u})_j - 2W_j + 2\lambda_j \\ &= 2(C\hat{u})_j - 2(C\hat{u}^\iota)_j = 0. \end{aligned}$$

Now, by facts $(a)$ and $(b)$ we have that $\hat{u} \in A^{\iota}_C(\hat{u}^{\iota}) \cup B^{\iota}_C(\hat{u}^{\iota})$. So, by assumption, $\hat{u}^{\iota} \in M$ implies $\hat{u} \in M$ as long as $\beta_j \geq 0$ for all $j$. We have therefore shown that

$$\inf_{\beta \in \mathcal{O}^{\iota}} P_{\beta}(\hat{u} \in M) \geq P(\hat{u}^{\iota} \in M).$$

To see the reverse inequality, note that *if* $\hat{u}_j + \beta_j > 0$ for all $j$, then $V$ is differentiable at $\hat{u}$ and

$$\frac{\partial V(\hat{u})}{\partial u} = 2C\hat{u} - 2W + 2\lambda = 2C\hat{u} - 2C\hat{u}^{\iota} = 0,$$

implying that $\hat{u} = \hat{u}^{\iota}$. Also note that $\hat{u}_j + \beta_j > 0$ for each $j$ is equivalent to all of the Lasso's components being strictly positive, i.e., $\hat{\beta}_{\mathrm{L}} \in \mathcal{O}^{\iota}_{\mathrm{int}}$, so that

$$\{\hat{u} \in M\} \subseteq \{\hat{u}^{\iota} \in M\} \cup \{\hat{\beta}_{\mathrm{L}} \notin \mathcal{O}^{\iota}_{\mathrm{int}}\}.$$

Now let $\kappa$ be a bound in the sup-norm on the set $\{z \in \mathbb{R}^p : \|Cz\|_{\infty} \leq \|\lambda\|_{\infty}\}$ and for an arbitrary $\varepsilon > 0$, pick $\beta^* \in \mathbb{R}^p$ such that $P(\hat{u}_{\mathrm{LS}} \leq \kappa\iota - \beta^*) \leq \varepsilon$, where $\hat{u}_{\mathrm{LS}} = (\hat{\beta}_{\mathrm{LS}} - \beta^*) \sim N(0, \sigma^2 C^{-1})$. Note that by Proposition 1, this implies that

$$P_{\beta^*}(\hat{\beta}_{\mathrm{L}} \leq 0) = P_{\beta^*}(\hat{u} - \hat{u}_{\mathrm{LS}} + \hat{u}_{\mathrm{LS}} \leq -\beta^*) \leq P_{\beta^*}(-\kappa\iota + \hat{u}_{\mathrm{LS}} \leq -\beta^*) \leq \varepsilon,$$

yielding

$$\inf_{\beta \in \mathcal{O}^{\iota}} P_{\beta}(\hat{u} \in M) \leq P_{\beta^*}(\hat{u} \in M) \leq P(\hat{u}^{\iota} \in M) + \varepsilon.$$

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

Using Proposition 2, we can obtain a result on the minimal coverage probability of a large class of sets for the general case (i.e., the case where the signs of $\beta$ are unknown) by applying the proposition to all possible orthants using a sign-flipping argument. To that end, it will prove convenient to first describe the sets for which Proposition 2 can be applied orthant-wise.

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

$$\begin{aligned} A^d_{\bar{C}}(m) &= \bigcap_{j=1}^{p} A^{d_j}_{\bar{C},j}(m) \\ &= \bigcap_{j=1}^{p} \{z \in \mathbb{R}^p : d_j(\bar{C}m)_j \leq d_j(\bar{C}z)_j, d_j z_j \leq 0\}. \end{aligned}$$

The set $A^d_{\bar{C}}(m)$ is an intersection of $2p$ half-spaces, $p$ of which determine the orthant the set is located in via the sign-vector $d$. The other $p$ half-spaces are defined by hyperplanes that intersect at the point $m$. Figure 2.1 shows one example of such a set. Note that in general $A^d_{\bar{C}}(m)$ could be non-empty also for $\mathrm{sgn}(m) \neq -d$.

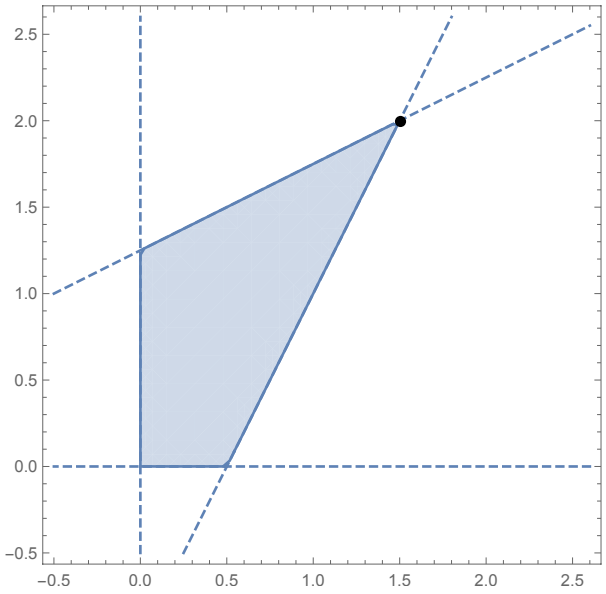



Figure 2.1: The set $A^{-\iota}_{\bar{C}}(m)$ with $\iota = (1,1)'$, $m = (1.5,2)'$ and $\bar{C} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$ along with the hyperplanes defining the set. The point $m = (1.5,2)'$ is displayed as a dot.

Before stating this chapter's main theorem note the following property which shows how applying the requirements of Proposition 2 to all orthants can be simplified to a relatively compact-looking condition.

**Lemma 3.**

$$\bigcup_{d\in\{-1,1\}^p} \bigcap_{j=1}^{p} A^{d_j}_{\bar{C},j}(m) = \bigcup_{d\in\{-1,1\}^p} \bigcap_{j=1}^{p} A^{d_j}_{\bar{C},j}(m) \cup B^{d_j}_{\bar{C},j}(m)$$

*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 point $z$ from the set on right-hand side. Then there exists a $d \in \{-1,1\}^d$ such that for each $j = 1,\dots,p$, $z$ is either contained in $A^{d_j}_j$ or in $B^{d_j}_j$. We pick $l \in \{-1,1\}^p$ in the following way: if $z \in A^{d_j}_j$, set $l_j = d_j$ and if $z \in B^{d_j}_j$, set $l_j = -d_j$. Then, by construction, $z \in A^{l}_j$ for all $j = 1,\dots,p$ and therefore $z \in \bigcap_j A^l_j$ so that $z$ is contained in the set on the left-hand side of the equation. □

The sets we consider can now be 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*

$$A^d_{\bar{C}}(m) \subseteq M$$

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

Using this notation, we can now state the main theorem which enables us to calculate minimal coverage probabilities for a large class of sets.

**Theorem 4.** *If $M \subseteq \mathbb{R}^p$ is non-random and satisfies Condition A with $\bar{C} = C$, then*

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

*where $\hat{u}^d \sim N(-C^{-1}\Lambda d, \sigma^2 C^{-1})$.*[18]

*Proof.* First note that

$$\inf_{\beta \in \mathbb{R}^p} P_\beta(\hat{u} \in M) = \min_{d \in \{-1,1\}^p} \inf_{\beta \in \mathcal{O}^d} P_\beta(\hat{u} \in M).$$

Thus, if we can show that

$$\inf_{\beta \in \mathcal{O}^d} P_\beta(\hat{u} \in M) = P(\hat{u}^d \in M)$$

for each $d \in \{-1,1\}^p$, the proof is done. Now, fix $d$ and set $D = \operatorname{diag}(d)$. We consider the function

$$\begin{aligned} \tilde{V}(u) = V(Du) &= u'DCDu - 2u'DW + 2\sum_{j=1}^{p} \lambda_j \left[|d_j u_j + \beta_j| - |\beta_j|\right] \\ &= u'\tilde{C}u - 2u'\tilde{W} + 2\sum_{j=1}^{p} \lambda_j \left[|u_j + d_j\beta_j| - |d_j\beta_j|\right], \end{aligned}$$

where $\tilde{C} = DCD$, $\tilde{W} = DW \sim N(0, \sigma^2\tilde{C})$. We write $\tilde{u}$ for the minimizer of $\tilde{V}$, and, analogously to Section 2.4, we define $\tilde{u}^\iota$ to be the minimizer of the function $u'\tilde{C}u - 2u'\tilde{W} + 2\sum_{j=1}^{p} \lambda_j u_j$.

If we can show that the set $DM$ satisfies the requirement of Proposition 2 with the matrix $\tilde{C}$ in place of $C$, we may conclude that

$$\inf_{\beta : d_j\beta_j \geq 0} P_\beta(\tilde{u} \in DM) = P(\tilde{u}^\iota \in DM).$$

Note that $\hat{u} = D\tilde{u}$, $\hat{u}^d = D\tilde{u}^\iota$ and $D^{-1} = D$, so that

$$\inf_{\beta \in \mathcal{O}^d} P(\hat{u} \in M) = \inf_{\beta \in \mathcal{O}^d} P(\tilde{u} \in DM) = P(\tilde{u}^\iota \in DM) = P(\hat{u}^d \in M),$$

which proves the formula for the infimal coverage probability. We now show that the set $DM$ satisfies that

$$\bigcap_{j=1}^{p} A^\iota_{\tilde{C},j}(Dm) \cup B^\iota_{\tilde{C},j}(Dm) \subseteq DM$$

for all $m \in M$. A straightforward calculation shows that this is equivalent to

$$\bigcap_{j=1}^{p} A^{d_j}_{C,j}(m) \cup B^{d_j}_{C,j}(m) \subseteq M$$

[18] Recall that $\Lambda = \operatorname{diag}(\lambda)$.

for each $m \in M$ which clearly holds by Condition A and Proposition 3. As $V^d$ is strictly convex, the distributional result on $\hat{u}^d$ immediately follows by solving

$$\frac{\partial V^d(u)}{\partial u} = 0$$

and observing that

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

□

**Remark 5.** *Inspection of the preceding proof will show that the characterization of the minimal coverage probability in Theorem 4 does not depend on the distribution of $\hat{u}$ and, in principle, holds for any error distribution, if $\hat{u}^d$, are viewed as the minimizers of $V^d$. Do note, however, that the distributions of our stochastic bounds for the true estimation error, $\hat{u}^d$ will very much depend on the distribution of $\varepsilon$. Since the assumption of Gaussian errors is a standard one, non-Gaussian cases will not be discussed in this thesis.*

The distributions of $\hat{u}^d$ that determine the formula for the infimal coverage probability are shifted normal distributions with the same variance-covariance matrix as the estimation error of the corresponding Least-squares estimator, $\hat{u}_{\text{LS}} = \hat{\beta}_{\text{LS}} - \beta$. The distribution's mean depends on the regressors and the vector of tuning parameters. This fact is quite useful for the calculation of minimal coverage probabilities, as the normal distribution probably is one of the best-understood distributions and easy to calculate numerically using software. Since Condition A for $p = 1$ simply requires the corresponding set $M$ to be an interval containing zero, Theorem 4 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 $\frac{1}{\sqrt{n}}\lambda_j$ of the vector of tuning parameters in this thesis.) The following obvious corollary specifies the resulting valid confidence region based on the Lasso estimator.

**Corollary 6.** *Let $0 < \alpha < 1$. If $M \subseteq \mathbb{R}^p$ is non-random and satisfies Condition A with $\bar{C} = C$, as well as $\min_{d \in \{-1,1\}^p} P(\hat{u}^d \in M) = 1 - \alpha$ with $\hat{u}^d \sim N(-C^{-1}\Lambda d, \sigma^2 C^{-1})$, then*

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

## 2.5 Constructing a confidence set

We now turn to a discussion of the important matter of how to choose an appropriate set $M \subseteq \mathbb{R}^p$ 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 Least-squares estimator. As mentioned in the previous section, we need to find a set $M \subseteq \mathbb{R}^p$ that satisfies Condition A with $\bar{C} = C$ and

such that $\min_{d\in\{-1,1\}^p} P(\hat{u}^d \in M) = 1-\alpha$ where

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

The resulting confidence set for $\beta$ is then the shifted set $\hat{\beta}_{\mathrm{L}} - M$. If we based the set on the Least-squares estimator $\hat{\beta}_{\mathrm{LS}}$ instead of $\hat{\beta}_{\mathrm{L}}$, the canonical and best choice for $M$ in terms of volume is an ellipse determined by the contour lines of a $N(0, \sigma^2 C^{-1})$-distribution, the *C-ellipse*. Given the fact that the variance-covariance matrix of the distributions of $\hat{u}^d$ is in fact $\sigma^2 C^{-1}$, in addition to the fact that the means of the distributions add up to zero, it seems reasonable to consider the $C$-ellipse as a shape in connection with the Lasso estimator also. Indeed, it turns out that the $C$-ellipse does have some convenient properties. First, it does comply with Condition A which is shown in the following proposition.

**Proposition 7.** *For any $k > 0$ the $C$-ellipse given by*

$$E_C(k) = \{z \in \mathbb{R}^p : z'Cz \leq k\}$$

*satisfies Condition A with $\bar{C} = C$ for any $k > 0$.*

*Proof.* Let $m \in E_C(k)$ and $t \in A_C^d(m)$. We show that $t \in E_C(k)$. Remember that $D = \mathrm{diag}(d)$ satisfies $DD = I_p$. Since $t \in A_C^d(m)$ we have $-Dt \in \mathcal{O}^\iota$ and $-DC(m-t) \in \mathcal{O}^\iota$ implying that

$$t'C(m-t) = (Dt)'DC(m-t) \geq 0.$$

Furthermore, since $(m-t)'C(m-t) \geq 0$, we have

$$m'C(m-t) \geq y'C(m-t) \geq 0$$

which in turn yields

$$m'Cm \geq m'Ct \geq t'Ct \geq 0.$$

But this means that $k \geq m'Cm \geq m'Ct \geq t'Ct$ and therefore $t \in E_C(k)$. □

Given that the $C$-ellipse satisfies Condition A, we still have to calculate the coverage probabilities of $2^p$ Gaussian random variables in order to construct a valid confidence set, perhaps even iteratively for several sizes. Conveniently, however, it turns out that one only has to consider one of these $2^p$ $\hat{u}^d$'s. Indeed, it is sufficient to consider the vertexes of the parallelogram[19] $\{C^{-1/2}\Lambda d : d \in \{-1,1\}^p\}$ that have the largest Euclidean distance from the origin. This is shown in the next proposition.

[19] This set can be viewed as a box around the origin that is distorted by the linear Function $C^{-1/2}\Lambda$ which yields a parallelogram.

**Proposition 8.** *For any $k > 0$, we have that*

$$\underset{d\in\{-1,1\}^p}{\arg\min}\ P\left(\hat{u}^d \in E_C(k)\right) = \underset{d\in\{-1,1\}^p}{\arg\max}\ \|C^{-1/2}\Lambda d\|_2.$$

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

$$P\left(\hat{u}^d \in E_C(k)\right) = P\left(C^{1/2}\hat{u}^d \in C^{1/2}E_C(k)\right),$$

where $C^{1/2}\hat{u}^d \sim N(-C^{-1/2}\Lambda d, \sigma^2 I_p)$ and $C^{1/2}E_C(k) = \{z \in \mathbb{R}^p : \|z\|_2^2 \leq k\}$. So clearly, the smallest probability will be achieved for the distribution with mean furthest away from the origin, which is any $d^*$ maximizing $\|C^{-1/2}\Lambda d\|_2$ over all $d \in \{-1,1\}^p$. □

Note that if $d^* \in \{-1,1\}^p$ 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 $P(\hat{u}^{d^*} \in E_C(k^*)) = 1-\alpha$, which is easily done numerically. Note that Proposition 8 also shows that the ellipse $E_C(k^*)$, and therefore the resulting confidence set based on the Lasso estimator, is larger in volume than the one based on the Least-squares estimator, since $E_C(k^*)$ needs to be large enough as to have mass $1-\alpha$ with respect to the $N(-C^{-1}\Lambda d^*, \sigma^2 C^{-1})$-measure whereas for the ellipse corresponding to the Least-squares estimator, it suffices to have mass $1-\alpha$ with respect to the $N(0, \sigma^2 C^{-1})$-measure. Clearly, the difference in size will increase as the tuning parameters become larger. 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 Least-squares estimator with the same coverage probability. When comparing the two confidence sets, we emphasize that since the ellipses are centered at different values[20], the smaller ellipse based on the Least-squares 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 below example.

It is quite obvious that the $C$-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 $N(-C^{-1}\Lambda d^*, \sigma^2 C^{-1})$-distributions (in such a way that Condition A is preserved). To find the best possible shape, one would have to minimize the volume of the set over all possible shapes satisfying Condition 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 A,

[20] I.e., $\hat{\beta}_{\mathrm{L}}$ and $\hat{\beta}_{\mathrm{LS}}$.

by simply "adding" the required points.

**Proposition 9.** *For any $M \subseteq \mathbb{R}^p$, the set*

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

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

*Proof.* We start by showing that for any $m \in \mathbb{R}^p$, $d \in \{-1,1\}^p$, we have

$$A^d_{\bar{C}}(t) \subseteq A^d_{\bar{C}}(m) \quad \text{for all } t \in A^d_{\bar{C}}(m). \tag{2.2}$$

Let $z \in A^d_{\bar{C}}(t)$. Then $d_j z_j \leq 0$ and $(\bar{C}t)_j \leq (\bar{C}z)_j$ for all $j$. But since $t \in A^d_{\bar{C}}(m)$, we also have $(\bar{C}m)_j \leq (\bar{C}t)_j$ for all $j$ so that that $(\bar{C}m)_j \leq (\bar{C}z)_j$ for all $j$ and therefore $z \in A^d_{\bar{C}}(m)$, thus proving (2.2). So clearly, the set

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

satisfies Condition A. For each $m \in M$, choose $d \in \{-1,1\}^p$ in such a way that $d_j = 1$ if $m_j = 0$ and $d_j = -\operatorname{sgn}(m_j)$ for $m_j \neq 0$. We then get $m \in A^d_{\bar{C}}(m)$, implying that the set in the display above actually contains $M$. □

We now take a look at an example for $p = 2$ illustrating the difference between the confidence ellipse based on the Least-squares 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

$$C = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}.$$

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 $\lambda_1 = \lambda_2 = \frac{1}{2}$. 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 construct confidence ellipses with level $\alpha = 0.05$ based on both the Least-squares and the Lasso estimator in the manner described earlier in this section. The resulting sets are shown in Figure 2.2.

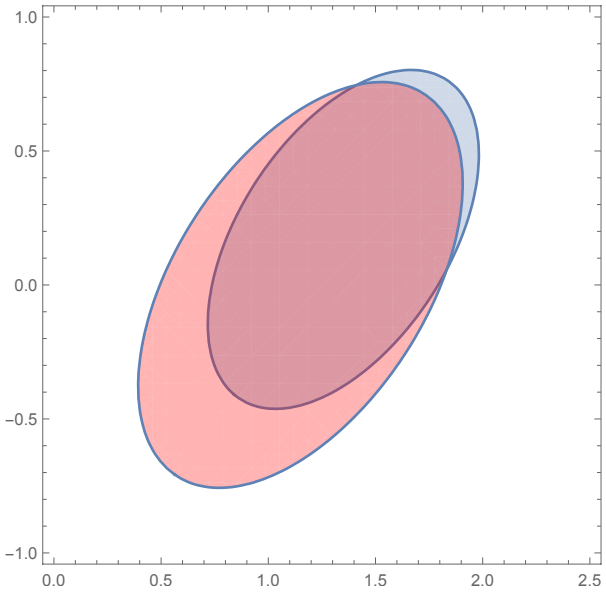



Figure 2.2: The confidence ellipses based on and centered at the Lasso estimator $\hat{\beta}_{\mathrm{L}} = (1.15, 0)'$ (red) and the smaller one based on and centered at the Least-squares estimator $\hat{\beta}_{\mathrm{LS}} = (1.35, 0.17)'$ (blue), respectively.

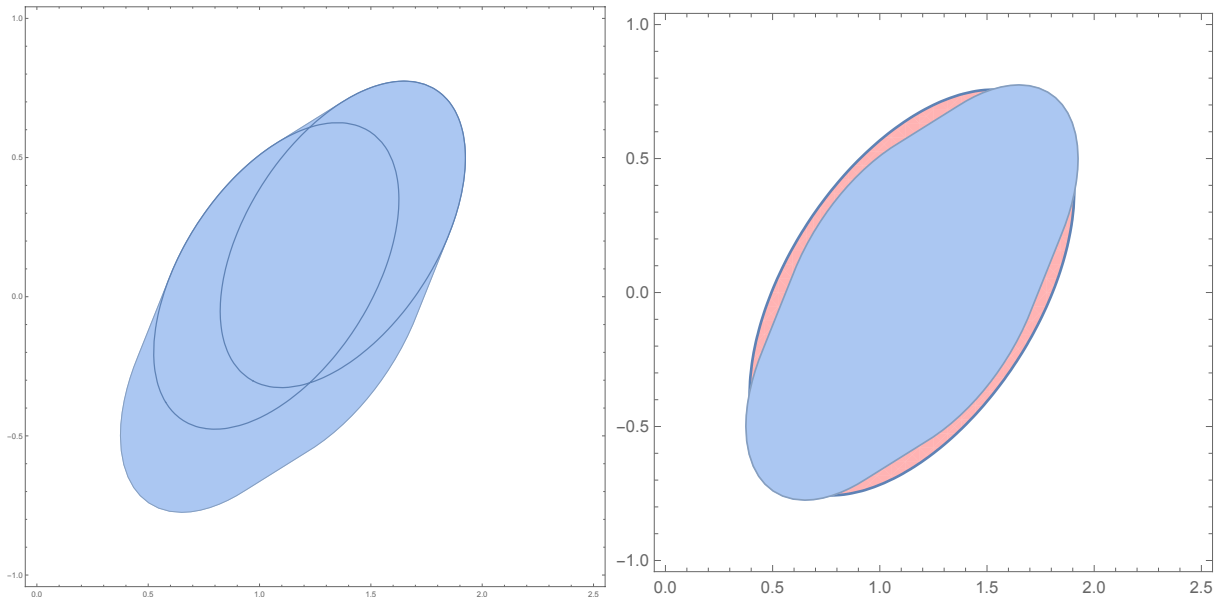
The plot clearly illustrates the above-mentioned fact that the confidence ellipse based on the Lasso estimator is larger than the confidence ellipse that is based on the Least-squares estimator. Also, the two sets are overlapping by a large amount. This is not surprising, as the maximal distance between the two estimators is bounded, c.f. Proposition 1. However, the Least-squares ellipse is not entirely contained in the one based on the Lasso, stressing the fact the Theorem 4 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[21] one for the Least-squares 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 $\hat{u}^d$, that is, the $2^p$ shifted $C$-ellipses

$$U(k) = \bigcup_{d \in \{-1,1\}^p} E_C(k) - n^{-1/2} C^{-1} \Lambda d,$$

where each set in the union is of optimal shape for the corresponding distribution of $\hat{u}^d$. As a starting point we choose the set's size parameter $k$ so that $P(\hat{u}^d \in E_C(k) - C^{-1}\Lambda d) = 1 - \alpha$. (Note that $k$ is then simply the parameter of the $C$-ellipse used for the Least-squares estimator, but any $k > 0$ such that $U(k)$ satisfies $P(\hat{u}^d \in U(k)) \geq 1 - \alpha$ 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

[21] In terms of volume.




(a) (b)

Figure 2.3: (a) Construction of the alternative shape based on $2^p = 4$ ellipses with two of the Least-squares-ellipses displayed within the set. (b) The resulting improved confidence set with the alternative shape (blue) and the previous elliptic shape (red), both based on at the Lasso estimator $\hat{\beta}_{\mathrm{L}} = (1.15, 0)'$.

necessary so that the resulting set satisfies Condition A. Proposition 9 ensures that

$$\bigcup_{m \in U(k)} \bigcup_{d \in \{-1,1\}^p} A_C^d(m)$$

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 $U(k)$. 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 in our example is depicted in Figure 2.3, (a) showing 2 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 quite obvious that the new shape has slightly less volume than the ellipse.

## 2.6 Asymptotic framework

We now derive asymptotic results that hold without assuming normality of the errors. Naturally, most quantities in our setup depend on the sample size $n$. While this has been suppressed in

the notation used so far, we make this dependence explicit in this subsection. Thus, $X = X_n$, $C = C_n = X_n'X_n$, $\varepsilon = \varepsilon_n$ and $W_n = X_n'\varepsilon_n$. Regarding the design matrix we assume that, in addition to the assumptions in Section 2.3, and *for all asymptotic considerations*, $X_n = (x_{1\cdot}', \ldots, x_{n\cdot}')'$ where $x_{i\cdot} \in \mathbb{R}^p$, meaning that the regressor matrix $X_n$ changes with $n$ only by appending rows.

We also need to make assumptions about the asymptotic behavior of the design matrix:

$$\frac{C_n}{n} \longrightarrow C_\infty$$

as $n \to \infty$, where $C_\infty$ is finite and positive definite. (Note that this setting assures consistency and asymptotic normality of the Least-squares estimator, c.f. Theorem 58 in Appendix C, for example.)

In the subsequent analysis we consider a so-called *moving-parameter* framework, i.e., we also allow the parameter $\beta$ to depend on the sample size: $\beta = \beta_n$. In practice, also the tuning vector $\lambda$ will typically (be chosen to) depend on the sample size, so that $\lambda = \lambda_n$. In line with the general notation, let $\Lambda_n = \operatorname{diag}(\lambda_n)$. While not important in the finite-sample analysis, the tuning parameter's asymptotic behavior is of vital importance when considering the limiting behavior of the estimator. We will consider two different regimes of the tuning parameter's behavior as $n$ gets large and start with the regime we refer to as *conservative tuning*.

Before proceeding to the corresponding subsection note that also the estimators $\hat{\beta}_{\mathrm{L}}$ and $\hat{\beta}_{\mathrm{LS}}$ do depend on the sample size. However, we will continue to suppress this dependence in order to ease notation.

### 2.6.1 Conservative tuning

In this regime and *throughout this subsection*, we require that

$$\frac{\lambda_n}{\sqrt{n}} \longrightarrow \lambda_\infty \in [0, \infty)^p$$

as $n \to \infty$. This implies that $\frac{\lambda_{n,j}}{n} \to 0$ for all $j = 1, \ldots, p$ , which in turn implies consistency of $\hat{\beta}_{\mathrm{L}}$ (see Theorem 1 in Knight & Fu, 2000 with the slight modification that in this thesis we allow for component-wise defined tuning parameters). Similarly to the previous sections let $\Lambda_\infty = \operatorname{diag}(\lambda_\infty)$.

**Remark 10.** *Such a choice of tuning parameters indeed yields a conservative model selection procedure in the sense that*

$$\limsup_{n\to\infty} \sup_{\beta\in\mathbb{R}^p} P_\beta\left(\hat{\beta}_j = 0\right) < 1 \tag{2.3}$$

*for each $j = 1, \ldots, p$. In particular, if $\beta_{n,j} = 0$ for each $n$, we have*

$$\limsup_{n\to\infty} P_{\beta_n}\left(\hat{\beta}_j = 0\right) < 1.$$

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

*Proof.* We show (2.3). Note that Proposition 1 entails that

$$\hat{\beta}_{\mathrm{L}} \in \hat{\beta}_{\mathrm{LS}} - \frac{1}{\sqrt{n}} B_n,$$

where

$$B_n = \{z \in \mathbb{R}^p : |\tfrac{1}{n}(C_n z)_j| \leq \tfrac{1}{\sqrt{n}} \lambda_{n,j} \text{ for } j = 1, \ldots, p\}.$$

Since $\lambda_n$ converges, we have $B_n \subseteq n C_n^{-1} \bar{B}_\delta$ with $\bar{B}_\delta = \{x \in \mathbb{R}^p : \|x\|_\infty \leq \delta\}$ for some $\delta > 0$. Since $nC_n^{-1} \to C_\infty^{-1}$, the set $\{nC_n^{-1} : n \in \mathbb{N}\}$ is bounded in operator sup-norm by Banach-Steinhaus[22], 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 $\liminf_{n\to\infty} \inf_{\beta \in \mathbb{R}^p} P_\beta(\hat{\beta}_{\mathrm{L},j} \neq 0) > 0$. To this end, define $\Psi_j = \mathbb{R}^{j-1} \times \{0\} \times \mathbb{R}^{p-j}$. Let $\xi_{j,n}^2$ and $\xi_{j,\infty}^2$ be the positive $j^{\text{th}}$ diagonal element of $C_n^{-1}$ and $C_\infty^{-1}$, respectively. Observe that

$$\begin{aligned}
\inf_{\beta \in \mathbb{R}^p} P_\beta(\hat{\beta}_{\mathrm{L},j} \neq 0) &\geq \inf_{\beta \in \mathbb{R}^p} P_\beta\left((\hat{\beta}_{\mathrm{LS}} - \frac{1}{\sqrt{n}} B_n) \cap \Psi_j = \varnothing\right) \\
&\geq \inf_{\beta \in \mathbb{R}^p} P_\beta\left(\sqrt{n}\hat{\beta}_{\mathrm{LS},j} + \gamma < 0 \text{ or } \sqrt{n}\hat{\beta}_{\mathrm{LS},j} - \gamma > 0\right) \\
&= 2\Phi(-\tfrac{\gamma}{\xi_{i,n}}) \longrightarrow 2\Phi(-\tfrac{\gamma}{\xi_{i,\infty}}) > 0
\end{aligned}$$

□

To derive asymptotically valid confidence sets for the general case of independent errors we will consider an appropriately scaled version of the Lasso estimation error. To that end let

$$\begin{aligned}
Q_n(u) &= L\left(\tfrac{u}{\sqrt{n}} + \beta_n\right) - L(\beta_n) \\
&= u' \tfrac{C_n}{n} u - 2\tfrac{u' W_n}{\sqrt{n}} + \tfrac{2}{\sqrt{n}} \sum_{j=1}^{p} \lambda_j \left[|u_j + \sqrt{n}\beta_{n,j}| - |\sqrt{n}\beta_{n,j}|\right].
\end{aligned}$$

Note that this function is minimized at $\sqrt{n}(\hat{\beta}_{\mathrm{L}} - \beta_n)$, the scaled estimation error. To implicitly state the asymptotic distribution of the estimator by considering the asymptotic behavior of the function $Q_n$ in the following proposition. This proposition essentially is Theorem 5 from Knight & Fu (2000) and can be proven in the same manner simply by adjusting for component-wise tuning.

**Proposition 11.** *Assume that* $\sqrt{n}\beta_n \to \beta_\infty \in \overline{\mathbb{R}}^p$. *Then* $\sqrt{n}(\hat{\beta}_L - \beta_n) \xrightarrow{d} \hat{u}_\infty = \arg\min_{u \in \mathbb{R}^p} Q_\infty(u)$, *where*

$$Q_\infty(u) = u' C_\infty u - 2W_\infty' u + 2\sum_{j=1}^{p} \lambda_j \left[\mathbb{1}_{\{\beta_{\infty,j} \in \mathbb{R}\}}(|\beta_{\infty,j} + u_j| - |\beta_{\infty,j}|) + \mathbb{1}_{\{|\beta_{\infty,j}| = \infty\}} \operatorname{sgn}(\beta_{\infty,j}) u_j\right] \tag{2.4}$$

*and* $W_\infty \sim N(0, \sigma^2 C_\infty)$.

[22] C.f., for example, Theorem 57 in Appendix C.

Note that the vector $\beta_\infty$ takes over the role of $\sqrt{n}\beta_n$ in the finite-sample version of the function, $Q_n$, where the cases of $\sqrt{n}\beta_{n,j} = \pm\infty$ are now included in the asymptotic setting. Also, the assumption of $\sqrt{n}\beta_n$ converging in $\overline{\mathbb{R}}^p$ is not a restriction in the sense that, by compactness of $\overline{\mathbb{R}}^p$, Proposition 11 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 now define $\hat{u}_\infty$ to be the unique minimizer of $Q_\infty$, and for $d \in \{-1, 1\}^p$, we define $Q^d_\infty(u) = u'C_\infty u - 2W'_\infty u + 2\sum_{j=1}^p \lambda_{\infty,j} d_j u_j$ with unique minimizer $\hat{u}^d_\infty$.

**Proposition 12.** *If $M \subseteq \mathbb{R}^p$ satisfies that*

$$\bigcap_{j=1}^{p} A^{\iota j}_{C_\infty,j}(m) \cup B^{\iota j}_{C_\infty,j}(m) \subseteq M$$

*for all $m \in M$, then*

$$\inf_{\beta_\infty \in \bar{\mathcal{O}}^\iota} P_{\beta_\infty}(\hat{u}_\infty \in M) = P(\hat{u}^\iota_\infty \in M).$$

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

$$\inf_{\beta_\infty \in \bar{\mathcal{O}}^\iota} P_{\beta_\infty}(\hat{u}_\infty \in M) \leq P_{\beta^*_\infty}(\hat{u}_\infty \in M) = P(\hat{u}^\iota_\infty \in M).$$

□

We can now formulate an asymptotic version of Theorem 4.

**Theorem 13.** *If $M \subseteq \mathbb{R}^p$ satisfies Condition A with $\bar{C} = C_\infty$, then*

$$\inf_{\beta_\infty \in \overline{\mathbb{R}}^p} P_{\beta_\infty}\left(\hat{u}_\infty \in M\right) = \min_{d \in \{-1,1\}^p} P\left(\hat{u}^d_\infty \in M\right),$$

*where $\hat{u}^d_\infty \sim N(C_\infty^{-1}\Lambda_\infty d, \sigma^2 C_\infty^{-1})$.*

*Proof.* The proof again is completely analogous to the proof of Theorem 4 after identifying $\sqrt{n}\beta_n$ with $\beta_\infty$, dropping the subscript $n$ everywhere and using Proposition 12 instead of Proposition 2.

Also, replace the orthant $\mathcal{O}^d$ by its closure $\bar{\mathcal{O}}^d$, let again $D = \text{diag}(d)$ and note that

$$\begin{aligned}
&V_\infty^d(u) = V_\infty(Du) \\
&= u'DC_\infty Du - 2u'DW_\infty + 2\sum_{i=1}^{p} \lambda_j \left[\mathbb{1}_{\{\beta_{\infty,j}\in\mathbb{R}\}}(|\beta_{\infty,j} + d_j u_j| - |\beta_{\infty,j}|) + \mathbb{1}_{\{|\beta_{\infty,j}|=\infty\}} \operatorname{sgn}(\beta_{\infty,j}) d_j u_j\right] \\
&= u'\tilde{C}_\infty u - 2u'\tilde{W}_\infty + 2\sum_{i=1}^{p} \lambda_j \left[\mathbb{1}_{\{d_j\beta_{\infty,j}\in\mathbb{R}\}}(|u_j + d_j\beta_{\infty,j}| - |d_j t_j|) + \mathbb{1}_{\{|d_j\beta_{\infty,j}|=\infty\}} \operatorname{sgn}(d_j\beta_{\infty,j}) u_j\right],
\end{aligned}$$

where $\tilde{C}_\infty = DC_\infty D$ and $\tilde{W}_\infty = DW_\infty$. □

Given this result we can again construct asymptotically valid confidence sets for the parameter $\beta_n$ in the following way.

**Corollary 14.** *If $M \subseteq \mathbb{R}^p$ satisfies Condition A with $\bar{C} = C_\infty$ and $\min_{d\in\{-1,1\}^p} P\left(\hat{u}^d \in M\right) = 1-\alpha$, where $\hat{u}^d \sim N(C_\infty^{-1}\Lambda_\infty d, \sigma^2 C_\infty^{-1})$ then*

$$\liminf_{n\to\infty} \inf_{\beta_n\in\mathbb{R}^p} P\left(\beta_n \in \hat{\beta}_L - \tfrac{1}{\sqrt{n}}M\right) = 1-\alpha.$$

*Proof.* Let $c = \liminf_{n\to\infty} \inf_{\beta_n\in\mathbb{R}^p} P_{\beta_n}(\beta \in \hat{\beta}_{\text{L}} - \frac{1}{\sqrt{n}}M)$. Then there exists a sequence $\beta_n$ in $\mathbb{R}^p$ such that $P_{\beta_n}(\beta_n \in \hat{\beta}_{\text{L}} - \frac{1}{\sqrt{n}}M) \to c$. Assume that $\sqrt{n}\beta_n \to \beta_\infty \in \overline{\mathbb{R}}^p$ (if the sequence does not converge, pass to subsequences). Since

$$P_{\beta_n}(\beta_n \in \hat{\beta}_{\text{L}} - \tfrac{1}{\sqrt{n}}M) = P_{\beta_n}(\sqrt{n}(\hat{\beta}_{\text{L}} - \beta_n) \in M) \longrightarrow c = P_{\beta_\infty}(\hat{u}_\infty \in M)$$

as $n \to \infty$ in the notation of Proposition 11. Theorem 13 then yields $c \geq \min_{d\in\{-1,1\}^p} P(\hat{u}_\infty^d \in M) = 1-\alpha$. To see the reverse inequality, let $\beta_n = d \in \{-1,1\}^p$ and note that for this sequence, we have

$$P_{\beta_n}(\beta_n \in \hat{\beta}_{\text{L}} - \tfrac{1}{\sqrt{n}}M) = P_{\beta_n}(\sqrt{n}(\hat{\beta}_{\text{L}} - \beta_n) \in M) \longrightarrow P_{\beta_\infty}(\hat{u}_\infty \in M)$$

as $n \to \infty$, where $\beta_\infty = (d_1\infty, \ldots, d_p\infty)' \in \overline{\mathbb{R}}^p$. Note that for this choice of $\beta_\infty$, $P_{\beta_\infty}(\hat{u}_\infty \in M) = P(\hat{u}_\infty^d \in M)$. Since $d \in \{-1,1\}^p$ was arbitrary, $c \leq \min_{d\in\{-1,1\}^p} P(\hat{u}_\infty^d \in M) = 1-\alpha$ follows. □

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 $\frac{1}{n}C_n$ and $\frac{1}{\sqrt{n}}\Lambda_n$ have settled to their limiting values $C_\infty$ and $\Lambda_\infty$, respectively.

### 2.6.2 Consistent tuning

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

$$\tfrac{1}{\sqrt{n}}\lambda_{n,j} \longrightarrow \infty$$

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

$$\tfrac{1}{n}\lambda_{n,j} \longrightarrow 0$$

*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 $\lambda_{n,j}$ converges for each $j = 1, \ldots, p$. Yet we emphasize that in order to ensure $P_{\beta_n}(\hat{\beta}_{\mathrm{L},j} = 0) \to 1$ whenever $\beta_{n,j} = 0$, we would need $\lambda_{n,j} \to \infty$ for each $j = 1, \ldots, p$ as well as need additional conditions on the regressor matrix $X_n$. For a discussion concerning necessary and sufficient conditions on $X_n$ in this context see Zou (2006), Zhao & Yu (2006) and Yuan & Lin (2007).

In the case of consistent tuning, the rate of the estimator is no longer $\frac{1}{\sqrt{n}}$, 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 component-wise 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

$$\lambda_n^{\max} = \max_{1 \leq j \leq p} \lambda_{n,j}$$

and $\lambda^* = (\lambda_1^*, \ldots, \lambda_p^*)'$ by

$$\frac{\lambda_{n,j}}{\lambda_n^{\max}} \longrightarrow \lambda_j^* \in [0, 1]$$

as $n \to \infty$ for each $j = 1, \ldots, p$. Note that $\lambda_j^* = 1$ for all $j$ in case all components are equally tuned.

**Proposition 15.** *Assume that* $n\beta_n/\lambda_n^{\max} \to \zeta \in \overline{\mathbb{R}}^p$. *Then* $n(\hat{\beta}_L - \beta)/\lambda_n^{\max} \xrightarrow{p} m = \arg\min_{u \in \mathbb{R}^p} G_\infty^\zeta(u)$, *where*

$$G_\infty^\zeta(u) = u' C_\infty u + 2 \sum_{j=1}^{p} \lambda_j^* \left[ \mathbb{1}_{\{\zeta_j \in \mathbb{R}\}} (|u_j + \zeta_j| - |\zeta_j|) + \mathbb{1}_{\{|\zeta_j| = \infty\}} \operatorname{sgn}(\zeta_j) u_j \right].$$

*Proof.* Define the function $G_n(u) = n[L(\beta_n + \frac{\lambda_n^{\max}}{n} u) - L(\beta_n)]/(\lambda_n^{\max})^2$ and note that $G_n$ is minimized at $n(\hat{\beta}_{\mathrm{L}} - \beta)/\lambda_n^{\max}$. The function $G_n$ is then given by

$$G_n(u) = u' \frac{C_n}{n} u - 2 \frac{1}{\lambda_n^{\max}} u' X' \varepsilon + 2 \sum_{j=1}^{p} \frac{\lambda_{n,j}}{\lambda_n^{\max}} \left[ \left| u_j + \frac{n}{\lambda_n^{\max}} \beta_{n,j} \right| - \left| \frac{n}{\lambda_n^{\max}} \beta_{n,j} \right| \right].$$

Clearly $\frac{1}{n} u' C_n u \to u' C_\infty u$ by assumption. Since $X_n' \varepsilon_n / \lambda_n^{\max} = (\sqrt{n}/\lambda_n^{\max}) X_n' \varepsilon_n / \sqrt{n}$ and $\lambda_n^{\max}/\sqrt{n} \to \infty$ as well as $X_n' \varepsilon_n / \sqrt{n} = O_P(1)$, the second term in the above display vanishes in probability. To treat the third term, simply note that $\lambda_{n,j}/\lambda_n^{\max} \to \lambda_j^* \in [0, 1]$ and $n\beta_{n,j}/\lambda_n^{\max} \to \zeta_j \in \overline{\mathbb{R}}$ by

assumption. Piecing this together yields

$$G_n(u) \xrightarrow{p} u'C_\infty u + 2\sum_{j=1}^{p} \lambda_j^* \left[ \mathbb{1}_{\{\zeta_j \in \mathbb{R}\}} (|u_j + \zeta_j| - |\zeta_j|) + \mathbb{1}_{\{|\zeta_j| = \infty\}} \operatorname{sgn}(\zeta_j) u_j \right] = G_\infty^\zeta(u)$$

as $n \to \infty$. Since $G_n$ and $G_\infty^\zeta$ are strictly convex and $G_\infty^\zeta$ is non-random, it follows by Geyer (1996)[23] that also the corresponding minimizers converge in probability to the minimizer of the limiting function. □

In contrast to the finite-sample and the conservative-tuning case, we make the dependence of the objective function $G_\infty^\zeta$ on the unknown parameter $\zeta \in \overline{\mathbb{R}}^p$ apparent in the notation to clarify what is done in the following. Proposition 15 shows that $\lambda_n^{\max}/n$ is indeed the correct (uniform) convergence rate as the limit of $n(\hat{\beta}_{\mathrm{L}} - \beta_n)/\lambda_n^{\max}$ is not zero 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 in a moving-parameter asymptotic framework. This fact has already been noted for 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 $n(\hat{\beta}_{\mathrm{L}} - \beta_n)/\lambda_n^{\max}$ is always contained in a bounded set which is described in Proposition 16. To this end, define the set of all possible minimizers to the asymptotic objective function as

$$\mathcal{M} = \bigcup_{\zeta \in \overline{\mathbb{R}}^p} \operatorname*{arg\,min}_{u \in \mathbb{R}^p} G_\infty^\zeta(u). \tag{2.5}$$

It turns out that this set can be given explicitly:

**Proposition 16.** *The set $\mathcal{M}$ can be written as*

$$\left\{ z \in \mathbb{R}^p : |(C_\infty z)_j| \le \lambda_j^*, 1 \le j \le p \right\} = C_\infty^{-1} \left\{ z \in \mathbb{R}^p : |z_j| \le \lambda_j^*, 1 \le p \right\}.$$

*Proof.* The equality of the two sets given in the above display is trivial. We show that the set $\mathcal{M}$ as defined in (2.5) 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 $\zeta \in \overline{\mathbb{R}}^p$ such that $m$ is the minimizer of $G_\infty^\zeta$. We need to show that $|(C_\infty m)_j| \le \lambda_j^*$ for all $j$. Assume that $|(C_\infty m)_{j_0}| > \lambda_j^*$ for some $1 \le j_0 \le p$. If $(C_\infty m)_{j_0} > \lambda_{j_0}^*$ we consider the directional derivative of $G_\infty^\zeta$ at its minimizer $m$ in the direction of $-e_{j_0}$ to get

$$\begin{aligned} \frac{\partial G_\infty^\zeta(m)}{\partial(-e_{j_0})} &= -2(C_\infty m)_j + 2\lambda_{j_0}^* \left[ \mathbb{1}_{\{m_j + \zeta_j \le 0\}} - \mathbb{1}_{\{m_j + \zeta_j > 0\}} \right] \\ &\le -2(C_\infty m)_j + 2\lambda_{j_0}^* < 0, \end{aligned}$$

[23] The corresponding result is stated as Theorem 56 in Appendix C.

which is a contradiction to $m$ minimizing $G_\infty^\zeta$. If $(Cm)_{j_0} < -\lambda_{j_0}^*$, then consider the directional derivative of $G_\infty^\zeta$ at $m$ in the direction of $e_{j_0}$ to arrive at

$$\begin{aligned}\frac{\partial G_\infty^\zeta(m)}{\partial e_{j_0}} &= 2(C_\infty m)_j + 2\lambda_{j_0}^* \left[\mathbb{1}_{\{m_j+\zeta_j \geq 0\}} - \mathbb{1}_{\{m_j+\zeta_j<0\}}\right] \\ &\leq -2(C_\infty m)_j + 2\lambda_{j_0}^* < 0,\end{aligned}$$

yielding a contradiction also.

To see the reverse set-inclusion, we need to show that for any $m \in \mathbb{R}^p$ satisfying $|(C_\infty m)_j| \leq \lambda_j^*$ for all $j = 1, \ldots, p$, there exists a $\zeta \in \overline{\mathbb{R}}^p$ such that $m$ is the minimizer of $G_\infty^\zeta$. Let $\zeta = -m \in \mathbb{R}^p$ and consider the directional derivative of $G_\infty^\zeta$ at $m$ in any direction $r \in \mathbb{R}^p$ with $\|r\|_2 = 1$.

$$\frac{\partial G_\infty^\zeta(m)}{\partial r} = 2r'C_\infty m + 2\sum_{j=1}^p \lambda_j^*|r_j| \geq \sum_{j=1}^p -2|(C_\infty m)_j r_j| + 2\lambda_j^*|r_j| = 2\sum_{j=1}^p \left[-|(C_\infty m)_j| + \lambda_j^*\right]|r_j| \geq 0.$$

Since the directional derivative is non-negative in any direction $r \in \mathbb{R}^p : \|r\|_2 = 1$ and $G_\infty^\zeta$ is (strictly) convex, $m$ must be the minimizer. □

Thus $\mathcal{M}$ can be viewed as a box that is distorted by the linear function $C_\infty^{-1}$, a bounded set in $\mathbb{R}^p$. In fact, this turns out to be a parallelogram whose corner points are given by the set $\{C_\infty^{-1}\Lambda^* d : d \in \{-1, 1\}^p\}$, where $\Lambda^* = \text{diag}(\lambda^*)$. 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 13, appearing without randomness in the limit in the consistently tuned case. Using Proposition 16, a simple asymptotic confidence set can now be constructed as is done in the following corollary.

**Corollary 17.** *We have*

$$\lim_{n\to\infty} \inf_{\beta\in\mathbb{R}^p} P_\beta\left(\beta \in \hat{\beta}_L - \psi\frac{\lambda_n^{\max}}{n}\mathcal{M}\right) = 1$$

*for any* $\psi > 1$ *and*

$$\lim_{n\to\infty} \inf_{\beta\in\mathbb{R}^p} P_\beta\left(\beta \in \hat{\beta}_L - \psi\frac{\lambda_n^{\max}}{n}\mathcal{M}\right) = 0$$

*for any* $\psi < 1$.

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

$$P_{\beta_{n_k}}\left(\beta_{n_k} \in \hat{\beta}_\text{L} - \psi\frac{\lambda_{n_k}^{\max}}{n_k}\mathcal{M}\right) = P_{\beta_{n_k}}\left(\frac{n_k}{\lambda_{n_k}^{\max}}(\hat{\beta}_\text{L} - \beta_{n_k}) \in \psi\mathcal{M}\right) \longrightarrow c$$

as $k \to \infty$. Note that $\psi\mathcal{M} = \{m \in \mathbb{R}^p : |(C_\infty m)_j| \leq \psi\lambda_j^*, 1 \leq j \leq p\}$. Now, pick a further subsequence $n_{k_l}$ such that $\lambda_{n_{k_l}}^{\max}\beta_{n_{k_l}}/n_{k_l}$ converges in $\overline{\mathbb{R}}^p$ to, say, $\zeta$. Proposition 15 then shows that

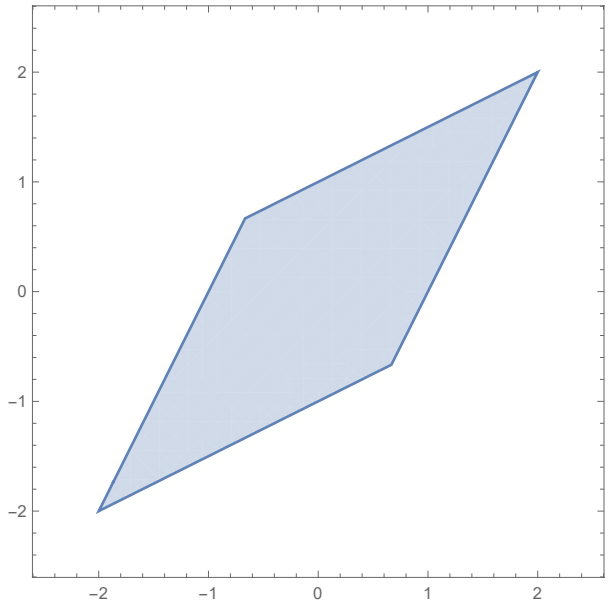



Figure 2.4: The set $\mathcal{M}$ for $C_\infty = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$ and $\lambda_j^* = 1$ for $j = 1, 2$.

$n_{k_l}(\hat{\beta}_{\mathrm{L}} - \beta_{n_{k_l}})/\lambda_{n_{k_l}}^{\max}$ converges in probability to the unique minimizer of $G_\infty^\zeta$ as $l \to \infty$. Finally, Proposition 16 implies that $c = 1$.

We next look the case where $\psi < 1$. Let $m = C_\infty^{-1}\lambda^*$ so that $m \in \mathcal{M} \setminus \psi\mathcal{M}$. From the proof of Proposition 16, we know that for $\zeta = -m$ we have $m = \arg\min_{u \in \mathbb{R}^p} G_\infty^\zeta(u)$. Let $\beta_n = n\zeta/\lambda_n^{\max}$. By Proposition 15, $n(\hat{\beta}_{\mathrm{L}} - \beta_n)/\lambda_n^{\max}$ converges to $m$ in $P_{\beta_n}$-probability, so that $P_{\beta_n}(n(\hat{\beta}_{\mathrm{L}} - \beta_n)/\lambda_n^{\max} \in d\mathcal{M}) \to 0$. □

Note that nothing can be said about the boundary case $\psi = 1$. This corollary is a generalization of the simple confidence interval given in Proposition 6 in Pötscher & Schneider (2010). The shape of the set $\mathcal{M}$ is displayed in Figure 2.4 for $C_\infty = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$ and $\lambda_j^* = 1$ for each $j \in \{1, 2\}$. Note the set $\mathcal{M}$ is not required to satisfy Condition A and, in fact, will not comply with this condition for certain matrices $C_\infty$.

A few properties should be noted here: First, the size of these sets decreases much more slowly at the rate $\frac{1}{n}\lambda_n^{\max}$ compared to the case of conservatively tuned model selection, where the size decreases at rate[24] $\frac{1}{\sqrt{n}}$. This is in line with Pötscher & Schneider (2010) and shows that one should refrain from using consistently tuned Lasso estimators if one is interested in inference.

Moreover, it should be noted that the concept of an asymptotic probability one confidence set is quite non-standard and hard to interpret, as it is questionable in which sense the asymptotic

[24]However, one should also note here that the confidence regions based on the consistently tuned Lasso may be smaller, even contained in those based on the conservatively tuned estimator for small sample sizes $n$. This reflects the fact that these are merely asymptotically valid, rendering them rather useless for quite small samples.

distribution of the appropriately scaled estimation error, which is simply point-mass, represents the finite-sample distribution which (while contracting to a point-mass) has full support on $\mathbb{R}^p$ for any finite $n$ and thus a probability one confidence region would have to be all of $\mathbb{R}^p$ (up to Lebesgue null-sets) for any finite sample. It is thus probably best to view this asymptotic confidence set as an interesting theoretical construct, rather than a method to conduct meaningful inference and use a conservatively tuned Lasso estimator in case the latter is desired.

## 2.7 Discussion and conclusion

This chapter provides a method of constructing confidence sets for the entire parameter vector that are based on the Lasso estimator. This is done by, for each $\omega \in \Omega$, bounding the estimation error (whose distribution depends on the true parameter in an intricate way) by the minimizers of functions that only depend on the unknown parameter via its sign-vector. These functions' minimizers have quite well-behaved distributions and can be used to lower-bound the coverage probability of sets satisfying rather mild conditions that depend on the regressor matrix. Moreover, it is shown that the resulting bound for the coverage probability is indeed attained for certain parameters, thus yielding a formula for the minimal coverage probability for such sets.

It is shown that the volumes of the resulting confidence sets are larger than the volume of the (canonical) Least-squares confidence set. Moreover, the difference in volume depends on the level of penalization that is governed by the size of the parameter vector $\lambda$, whereby smaller penalization parameters correspond to smaller confidence regions. These findings hold both in finite samples as well as in asymptotic settings where the estimator is tuned to perform conservative model selection.

When the Lasso is tuned to perform consistent model selection, an asymptotic probability one confidence region can be constructed. This takes the form of a parallelogram that is determined by the regressor matrix. The practical use of such a set is, however, debatable, since the asymptotic distribution of the estimation error, which is simply point-mass, does not reflect the finite-sample behavior well.

This chapter lays the theoretical foundations for further considerations, as the results and considerations can be used to obtain solutions for more specific and practice-oriented issues, such as the unknown variance case, single component-inference and partial Lasso inference. Some of those questions are discussed in Chapter 4.

# Chapter 3

# On the distribution of the Lasso estimator

## 3.1 Introduction

Due to the lack of a closed-form expression, the distribution of the Lasso estimator has remained unknown in the general[1] case for quite some time despite its popularity in both theoretical analysis and practice. The literature on the distributional properties of the Lasso estimator in the low-dimensional setting ($p \leq n$) includes the often-cited paper by Knight & Fu (2000) who derive the asymptotic distribution when the estimator is tuned to perform conservative model selection, albeit in an implicit form. Pötscher & Leeb (2009) give a detailed analysis in the framework of a linear regression model with orthogonal design in which case the problem can be treated component-wise and an explicit expression of the estimator is available. They derive the distribution of the Lasso estimator in finite samples as well as in the two asymptotic regimes of consistent and conservative tuning. They provide valuable insights into the behavior of the estimator by showing that the component-wise distribution takes the form of a shifted Least-squares estimator conditional on the estimate differing from zero and point-mass at the origin. However, this does not yield too much information about the case of larger numbers of regressors with non-trivial dependence structures. In more recent contributions, Jagannath & Upadhye (2018) give an approximate expression for the marginal pdfs of the one-dimensional components of a linear transformation of the Lasso estimator, while Miolane & Montanari (2018) provide concentration inequalities in a high-dimensional setup under sparsity assumptions on the true parameter.

Most notably, however, Zhou (2014) produced the distribution of the so-called *augmented Lasso*, that is, the Lasso estimator augmented by the subgradient of the penalization term $\|\beta\|_1$ of the Lasso function[2] by carefully examining the Karush Kuhn Tucker (KKT) conditions of a minimization

[1] I.e., a regression model with several non-orthogonal regressors.
[2] Evaluated at the Lasso solution.

problem equivalent to the one used to define the Lasso in this thesis[3]. While this augmentation seems redundant, it enables the construction of a bijection between the term $X'\varepsilon$ and the augmented estimator. The distribution of the Lasso estimator itself can then be obtained by calculating the marginal distribution of the augmented estimator. As the distribution of the augmented estimator is mainly used as a vehicle to obtain more efficient re-sampling algorithms, the link between the error term and the Lasso, however, is not easy to grasp when using the - quite abstract - approach presented in this reference, thus leaving room for further analysis.

It turns out that a much more intuitive[4] representation the Lasso's distribution can be obtained in a rather simple fashion, for example, using a fairly simple argument about convex optimization. In this way, it is possible to gain a far better understanding of the relationship between the Lasso estimator and the error-term, i.e., the random part in our model, and thus ultimately, the distribution of the estimator itself.

In essence we do the following: We consider a different minimality condition that is based on directional derivatives. It turns out that using these conditions yields quite well-behaved distributions conditional on any given active set, i.e., the estimator's non-zero components. Given these parts we can then piece together the distribution of the Lasso estimator to obtain its cdf. This line of reasoning works in both high- and low-dimensional settings, even though much more explicit results can be obtained in the latter case. Moreover, using this minimality condition it is possible to establish a unique relationship between the Lasso and the Least-squares estimator in the low-dimensional case in the form of so-called *shrinkage areas*[5]: For each possible value of the Lasso estimator, one can specify a set containing the corresponding Least-squares estimators.

The chapter also touches upon the topic of the estimator's uniqueness in high-dimensional settings and gives an intuitive illustration of the issue. An important contribution on this topic which provides, for example, sufficient conditions for uniqueness, is Tibshirani (2013), while Ali & Tibshirani (2019) extend these considerations to the generalized Lasso, which is considered in this thesis. Finally, Ewald & Schneider (2020) extend this thesis' findings and give a necessary and sufficient condition for uniqueness, while Schneider & Tardivel (2020) generalize this result to penalized Least-squares estimators whose penalties take the shape of a polytope-shaped norm (and which encompass the Lasso).

We will also see that, in certain situations, the Lasso will not include some regressors in any its models and this property is purely based on the design matrix as well as the penalty vector. This is related[6] to the concept of *SAFE* rules (El Ghaoui et al., 2012) and *STRONG* rules (Tibshirani

[3]Indeed, the Lasso can be viewed as the minimizer of a differentiable function over a restricted area by splitting the parameter vector up into positive and negative parts, c.f. Rosset & Zhu (2007).

[4]In comparison to Zhou (2014).

[5]We choose the term *shrinkage area*, since all Least-squares estimates that fall into one of these sets are "shrunk" to a corresponding Lasso solution by penalizing the estimator's objective function.

[6]To learn how these concepts differ from the one that will be introduced in this chapter, refer to Section 3.4.

et al., 2012) which allow for a removal of certain variables, thus improving the algorithm's efficiency.

The chapter is organized as follows. After introducing some more notation (Section 3.2) we deal with the low-dimensional case, obtaining the estimator's distribution in finite samples in Section 3.3. We then turn to the afore-mentioned shrinkage areas to describe the Lasso's relationship with the Least-squares estimator in Section 3.3.1. We continue by characterizing the Lasso's distribution in a high-dimensional setting in Section 3.4 including results on the estimator's model selection properties.

## 3.2 Setting and notation

Recall our linear model (1.1)

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

We now assume that $\varepsilon$, the unobserved error term to be normally distributed: $\varepsilon \sim N(0, I_n\sigma^2)$.

Let $\{D_+, D_-, D_0\}$ be a partition of $\{1, \dots, p\}$ into three sets (some of which may be empty). It will be convenient to also describe this partition by a vector $d \in \{-1, 0, 1\}^p$ with $d_j = \mathbb{1}_{\{j \in D_+\}} - \mathbb{1}_{\{j \in D_-\}}$. Recall that for such $d$, we denote by $\mathcal{Q}^d = \{z \in \mathbb{R} : \operatorname{sgn}(z_j) = d_j \text{ for } j = 1, \dots, p\} = \{z \in \mathbb{R}^p : z_j < 0 \text{ for } j \in D_-, z_j > 0 \text{ for } j \in D_+, z_j = 0 \text{ for } j \in D_0\}$. Note that for $m \in \mathbb{R}^d$, $m + \beta \in \mathcal{Q}^d$ is a short-hand notation for $m_j < -\beta_j$ for $j \in D_-$, $m_j > -\beta_j$ for $j \in D_+$ and $m_j = -\beta_j$ for $j \in D_0$. To ease notataion later on let $D_\pm = D_+ \cup D_-$ and let $C_\pm = X'_{D_\pm} X_{D_\pm}$.

## 3.3 The low-dimensional case

Throughout this section, we assume that $X$ has full column rank $p$, implying that we are considering the low-dimensional setting where $p \leq n$. For our arguments, we again use the re-parametrized version of the objective function: As in Chapter 2 we shall analyze the Lasso's distribution by again considering its estimation error $\hat{u} = \hat{\beta}_{\text{L}} - \beta$ which is the minimizer of the function

$$V(u) = L(\beta + u) - L(\beta) = u'Cu - 2u'W + 2\sum_{j=1}^{p} \lambda_j \left[|u_j + \beta_j| - |\beta_j|\right],$$

where $W = X'\varepsilon \sim N(0, \sigma^2 C)$.

The main difficulty when analyzing the Lasso estimator is that it is defined as the minimizer to a non-differentiable function and hence does in general not possess an explicit[7] form. Because of this, previous analyses of the Lasso have considered various minimality conditions for the problem defining the estimator. Most notable is perhaps the approach that has been adopted by Rosset & Zhu (2007) as well as Zhou (2014) in which the components of the Lasso are split into positive

[7] At least not one that has a rather simple form and can be obtained by simply setting the function's derivative to zero.

and negative parts and the Lasso is viewed as the minimizer of a differentiable function whose components are restricted to being non-negative. The transformed minimization problem can then be analyzed by looking at the corresponding Karush-Kuhn-Tucker conditions.

Instead of minimizing a differentiable function on a constrained space, we adopt a different characterization of the Lasso estimator. While the Lasso function is not differentiable everywhere, all of its directional derivatives exist at each point in $\mathbb{R}^p$. Note that if, at some point of a function, all directional derivatives are non-negative, the point must be a local minimizer. For strictly convex functions, a point satisfying this property will be the function's global minimizer. Conveniently, it turns out that for the function $V$ this property can be ensured by checking a much simpler condition. Indeed, it is sufficient to check the directional derivatives in all (positive and negative) directions of the coordinate axes. This characterization of the minimizer which leads up to this chapter's main theorem is stated in the following lemma.

**Lemma 18.** *Let $m \in \mathbb{R}^p$. The following are equivalent:*

*(a)* $\frac{\partial V(m)}{\partial r} \geq 0 \;\; \forall r \in \mathbb{R}^p : \|r\|_2 = 1$

*(b)* $\frac{\partial V(m)}{\partial e_j} \geq 0$ *and* $\frac{\partial V(m)}{\partial (-e_j)} \geq 0$ *for* $j = 1, \ldots, p$.

*Proof.* Only $(b) \Rightarrow (a)$ needs to be proved. Let $d \in \{-1, 0, 1\}^p$ such that $m + \beta \in \mathcal{Q}^d$ and let $\{D_-, D_+, D_0\}$ be the corresponding partition of $\{1, \ldots, p\}$. A straight-forward calculation shows that

$$\begin{aligned}
\frac{\partial V(m)}{\partial r} &= 2r'Cm - 2r'W + 2\sum_{j=1}^{p} \lambda_j \left(-\mathbb{1}_{\{j \in D_-\}} r_j + \mathbb{1}_{\{j \in D_+\}} r_j + \mathbb{1}_{\{j \in D_0\}} |r_j|\right) \\
&= \sum_{j=1}^{p} \mathbb{1}_{\{r_j \geq 0\}} \left[(2Cm - 2W)_j + 2\lambda_j(-\mathbb{1}_{\{j \in D_-\}} + \mathbb{1}_{\{j \in D_+ \cup D_0\}})\right] r_j \\
&\quad + \mathbb{1}_{\{r_j < 0\}} \left[-(2Cm - 2W)_j + 2\lambda_j(\mathbb{1}_{\{j \in D_- \cup D_0\}} - \mathbb{1}_{\{j \in D_+\}})\right] (-r_j) \\
&= \sum_{j=1}^{p} \mathbb{1}_{\{r_j \geq 0\}} \frac{\partial V(m)}{\partial e_j} r_j + \sum_{j=1}^{p} \mathbb{1}_{\{r_j < 0\}} \frac{\partial V(m)}{\partial (-e_j)} (-r_j) \geq 0.
\end{aligned}$$

□

Using this characterization, we can now state the main theorem which gives a set of probabilities that make up the whole distribution of $\hat{u} = \hat{\beta}_{\text{L}} - \beta$.

**Theorem 19.** *Let $z \in \mathbb{R}^p$. Let $d \in \{-1, 0, 1\}^p$ such that $z + \beta \in \mathcal{Q}^d$ and let $\{D_-, D_+, D_0\}$ be the*

*corresponding partition of* $\{1,\dots,p\}$. *Then*

$$P(\hat{u}_j \le z_j \text{ for } j \in D_-,\ \hat{u}_j \ge z_j \text{ for } j \in D_+,\ \hat{u}_j = z_j \text{ for } j \in D_0)$$
$$= \int_{\substack{s_j \in [-\lambda_j,\lambda_j] \\ j \in D_0}} \cdots \int \int_{\substack{m_j \ge z_j \\ j \in D_+}} \cdots \int \int_{\substack{m_j \le z_j \\ j \in D_-}} \cdots \int \phi_{(0,\sigma^2 C)}(Cm_\beta + s_\lambda)\ \det(C_\pm) dm_{D_-} dm_{D_+} ds_{D_0},$$

*where* $m_\beta$ *is given by* $(m_\beta)_{D_\pm} = m_{D_\pm}$, $(m_\beta)_{D_0} = -\beta_{D_0}$ *and* $s_\lambda \in \mathbb{R}^p$ *is given by* $(s_\lambda)_{D_-} = -\lambda_{D_-}, (s_\lambda)_{D_+} = \lambda_{D_+}$, $(s_\lambda)_{D_0} = s_{D_0}$.

**Remark 20.** *In the above display* $m_\beta$ *denotes the vector that is composed of integration variables for the estimator's non-zero components and (plus/minus) the parameter's components for all others. Conversely, the vector* $s_\lambda$ *contains integration variables for the estimator's zero-components and are otherwise identical to plus, or minus*[8]*, the components of the tuning-vector* $\lambda$. *This is due to the estimator's structure, where some components are either shifted towards zero, or exactly to zero, thus "collapsing" the corresponding dimension. This will become apparent in the theorem's proof.*

*Proof of Theorem 19.* Since the function $V$ is strictly convex, a point $m \in \mathbb{R}^p$ minimizes $V$ if and only if the directional derivatives at this point satisfy $\frac{\partial V(m)}{\partial r} \ge 0$ for all $r \in \mathbb{R}^p : \|r\| = 1$. We wish to find all minimizers $m$ satisfying $m_j \le z_j$ for $j \in D_-$, $m_j \ge z_j$ for $j \in D_+$ and $m_j = z_j$ for $j \in D_0$. Note that this implies that $m + \beta \in \mathcal{Q}^d$. By Lemma 18 together with the fact that the conditions $\frac{\partial V(m)}{\partial e_j} \ge 0$ and $\frac{\partial V(m)}{\partial (-e_j)} \ge 0$ reduce to $\frac{\partial V(m)}{\partial u_j} = 0$ if $V$ is differentiable at $m$ with respect to the $j$-th component, we get that the following necessary and sufficient conditions for such $m$ to be a minimizer of $V$:

$$\begin{cases} W_j = (Cm)_j - \lambda_j & \text{for } j \in D_- \\ W_j = (Cm)_j + \lambda_j & \text{for } j \in D_+ \\ (Cm)_j - \lambda_j \le W_j \le (Cm)_j + \lambda_j & \text{for } j \in D_0 \end{cases} \tag{3.1}$$

Therefore, $m$ satisfying $m + \beta \in \mathcal{Q}^d$ is a minimizer of $V$ if and only if $W$ lies in the set

$$\{s \in \mathbb{R}^p : \ s_j = (Cm)_j - \lambda_j \text{ for } j \in D_-, s_j = (Cm)_j + \lambda_j \text{ for } j \in D_+,$$
$$(Cm)_j - \lambda_j \le s_j \le (Cm)_j + \lambda_j \text{ for } j \in D_0\},$$

which can be written as

$$Cm + \{s_\lambda : (s_\lambda)_{D_-} = -\lambda_{D_-}, (s_\lambda)_{D_+} = \lambda_{D_+}, |s_{\lambda,j}| \le \lambda_j \text{ for } j \in D_0\}.$$

Since we are interested in all minimizers $m$ of $V$ that satisfy $m_j \le z_j$ for $j \in D_-$, $m_j \ge z_j$ for

[8] Depending on the sign.

$j \in D_+$ and $m_j = z_j$ for $j \in D_0$ (i.e., $m - z \in \mathcal{Q}^d$), we let

$$Д(z) = \{Cm : m - z \in \mathcal{Q}^d\} + \{s \in \mathbb{R}^p : s_{D_-} = -\lambda_{D_-}, s_{D_+} = \lambda_{D_+}, |s_j| \leq \lambda_j \text{ for } j \in D_0\}.$$

As $W$ follows a $N(0, \sigma^2 C)$-distribution, we have

$$P(W \in Д(z)) = \int_{Д(z)} \phi_{(0,\sigma^2 C)}(v)\, dv.$$

Applying the substitution $v = Cm_\beta + s_\lambda$ yields the result. □

Given these probabilities for the estimation error, we can now give a formula for the corresponding probabilites of the actual estimator by using a simple shifting argument.

**Corollary 21.** *Let $z \in \mathbb{R}^p$ satisfy $z \in \mathcal{Q}^d$.*

$$\begin{aligned} &P(\hat{\beta}_{L,j} \geq z_j \text{ for } j \in D_-,\ \hat{\beta}_{L,j} \leq z_j \text{ for } j \in D_+,\ \hat{\beta}_{L,j} = 0 \text{ for } j \in D_0) \\ &= \int\cdots\int_{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}} \int\cdots\int_{\substack{s_j \leq z_j - \beta_j \\ j \in D_+}} \int\cdots\int_{\substack{s_j \geq z_j - \beta_j \\ j \in D_-}} \phi_{(0,\sigma^2 C)}(Cm_\beta + s_\lambda)\, \det(C_\pm) dm_{D_-} dm_{D_+} ds_{D_0}, \end{aligned}$$

*where $m_\beta$ and $s_\lambda \in \mathbb{R}^p$ are given by $(m_\beta)_{D_\pm} = m_{D_\pm}$, $(m_\beta)_{D_0} = (-\beta)_{D_0}$ and $(s_\lambda)_{D_-} = -\lambda_{D_-}, s_{D_+} = \lambda_{D_+}$, $(s_\lambda)_{D_0} = s_{D_0}$, respectively.*

An interesting consequence of this is the fact that the probability of the estimator being equal to exactly zero can be calculated by integrating the corresponding Gaussian density over the $\lambda$-box, given by $[-\lambda_1, \lambda_1] \times \cdots \times [-\lambda_p, \lambda_p]$, whose size is clearly determined by the penalization vector $\lambda$:

**Corollary 22.**

$$P(\hat{\beta}_L = 0) = \int_{-\lambda_p}^{\lambda_p} \cdots \int_{-\lambda_1}^{\lambda_1} \phi_{(C\beta, \sigma^2 C)}(s)\, ds.$$

Theorem 19 now puts us into a position to fully specify the distribution of the Lasso estimator. In case $\lambda_j > 0$ for all $j$, one easily sees from the preceding corollary that this distribution is not absolutely continuous with respect to the $p$-dimensional Lebesgue-measure and thus no pdf exists for this distribution. One can, however, represent the distribution through Lebesgue-densities after conditioning on which components of the estimator are negative, positive, and equal to zero. Towards this end, define $\mathcal{E}_\beta^d = \mathcal{Q}^d - \beta$. Note that $\hat{\beta}_{\mathrm{L}} \in \mathcal{Q}^d$ if and only if $\hat{u} \in \mathcal{E}_\beta^d$.

**Proposition 23.** *Assume that $P(\hat{u} \in \mathcal{E}_\beta^d) > 0$. Then, the distribution of $\hat{u} = \hat{\beta}_L - \beta$, conditional on the event $\{\hat{u} \in \mathcal{E}_\beta^d\}$, can be represented by a $\|d\|_1$-dimensional Lebesgue-density given by*

$$f^d(z_{D_\pm}) = \frac{\mathbb{1}_{\{\hat{u} \in \mathcal{E}_\beta^d\}}}{P(\hat{u} \in \mathcal{E}_\beta^d)} \int\cdots\int_{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}} \phi_{(0,\sigma^2 C)}\left(Cz_\beta + s_\lambda\right) \det(C_\pm) ds_{D_0},$$

*where $z_\beta$ is defined by $(z_\beta)_{D_\pm} = z_{D_\pm}$ and $(z_\beta)_{D_0} = -\beta_{D_0}$, and $s_\lambda$ is defined by $(s_\lambda)_{D_-} = -\lambda_{D_-}$, $(s_\lambda)_{D_+} = \lambda_{D_+}$, and $(s_\lambda)_{D_0} = s_{D_0}$.*

**Remark 24.** *Note that the constants $P(\hat{\beta}_L \in \mathcal{E}^d_\beta)$ can be calculated using Corollary 21.*

*Proof of Proposition 23.* Observe that

$$f^d(z_{D_\pm}) = \left(\frac{\partial}{\partial z_j}\right)_{j\in D_\pm} P\left(\hat{u}_j \le z_j \text{ for } j \in D_\pm | \hat{u} \in \mathcal{E}^d_\beta\right),$$

and note that by Theorem 19 we have for any $z \in \mathcal{E}^d_\beta$

$$\begin{aligned}
&P\left(\hat{u}_j \le z_j \text{ for } j \in D_-, \hat{u}_j \ge z_j \text{ for } j \in D_+ | \hat{u} \in \mathcal{E}^d_\beta\right) \\
&= \frac{1}{P(\hat{u} \in \mathcal{E}^d_\beta)} P\left(\hat{u}_j \le z_j \text{ for } j \in D_-, \hat{u}_j \ge z_j \text{ for } j \in D_+, \hat{u}_j = -\beta_j \text{ for } j \in D_0\right) \\
&= \frac{1}{P(\hat{u} \in \mathcal{E}^d_\beta)} \int\limits_{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}} \cdots \int \int\limits_{\substack{m_j \le z_j \\ j \in D_+}} \cdots \int \int\limits_{\substack{m_j \ge z_j \\ j \in D_-}} \cdots \int \phi_{(0,\sigma^2 C)}(Cm_\beta + s_\lambda) \det(C_\pm) dm_{D_-} dm_{D_+} ds_{D_0},
\end{aligned}$$

where $m_\beta \in \mathbb{R}^p$ is defined by $(m_\beta)_{D_\pm} = m_{D_\pm}$, and $(m_\beta)_{D_0} = -\beta_{D_0}$ and $s_\lambda \in \mathbb{R}^p$ is defined by $(s_\lambda)_{D_-} = -\lambda_{D_-}$, $(s_\lambda)_{D_+} = \lambda_{D_+}$, and $(s_\lambda)_{D_0} = s_{D_0}$. Differentiating with respect to $z_j : j \in D_\pm$ and taking the absolute value while noting that the conditional density is zero for all values that are not contained in the event that is being conditioned on gives the density, thus completing the proof. □

Equipped with even more notation, $\mathcal{E}^d_\beta(z) = \{s \in \mathcal{E}^d_\beta : s_j \le z_j \text{ for } j = 1, \dots, p\}$, we can now give a formula for the cdf of of $\hat{u} = \hat{\beta}_L - \beta$.

**Theorem 25.** *The cdf of $\hat{u} = \hat{\beta}_L - \beta$ is given by*

$$F(z) = P(\hat{u}_1 \le z_1, \dots, \hat{u}_p \le z_p) = \sum_{d \in \{-1,0,1\}^p} \int_{\mathcal{E}^d_\beta(z)} h^d(m_{D_\pm})\, d\nu_{\|d\|_1},$$

*where $\nu_k$ denotes $k$-dimensional Lebesgue-measure and where*

$$h^d(m_{D_\pm}) = \mathbb{1}_{\{\hat{u} \in \mathcal{E}^d_\beta\}} \int\limits_{\substack{s_j \in [-\lambda_j, \lambda_j] \\ j \in D_0}} \cdots \int \phi_{(0,\sigma^2 C)}\left(Cm_\beta + s_\lambda\right) \det(C_\pm) ds_{D_0},$$

*with $m_\beta \in \mathbb{R}^p$ given by $(m_\beta)_{D_\pm} = m_{D_\pm}$, and $(m_\beta)_{D_0} = -\beta_{D_0}$ and $s_\lambda \in \mathbb{R}^p$ given by $(s_\lambda)_{D_-} = -\lambda_{D_-}$, $(s_\lambda)_{D_+} = \lambda_{D_+}$, and $(s_\lambda)_{D_0} = s_{D_0}$.*

*Proof.* It is easily seen that

$$P\left(\hat{u}_1 \leq z_1, \ldots, \hat{u}_p \leq z_p\right) = \sum_{d \in \{-1,0,1\}^p} P(\hat{u} \in \mathcal{E}_\beta^d) \int_{\mathcal{E}_\beta^d(z)} f^d(s_{D_\pm})\, d\nu_{\|d\|_1}.$$

Plugging in the formula for $f^d$ completes the proof. □

**Remark 26.** *Inspection of the proofs reveals that the assumption of Gaussian errors is non-essential. One can thus easily obtain the distribution of the estimator for any kind of error-distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$. The estimator's resulting cdf will be the same as the above formula with the density of $X'\varepsilon$ replacing $\phi_{(0,\sigma^2 C)}$, the term's pdf when $\varepsilon \sim N(0, \sigma^2 I_n)$.*

Figures 3.1 and 3.2 display an example of the distribution of $\hat{u}$ for $C = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$, $\lambda_1 = \lambda_2 = 0.75$ and $\beta = (0, -0.25)'$. One can see that the Lasso's estimation error follows a shifted normal distribution conditional on $\hat{u}_j \neq -\beta_j$ for each $j$ with the shift depending on the signs of $\hat{u} + \beta$, as is to be seen in Figure 3.1. The second figure displays the mass which lies on the set $\{z \in \mathbb{R}^2 : z_1 = -\beta_1, \beta_2 \neq 0\}$, i.e., the density functions $h^{(0,1)}$ and $h^{(0,-1)}$ on their corresponding domains. The mass on the set $\{z \in \mathbb{R}^2 : z_2 = -\beta_2, \beta_1 \neq 0\}$ looks qualitatively similar to Figure 3.2. Note that we do also have some point-mass at $(-\beta)$, as is pointed out by Corollary 22.

All in all, we can observe quite an interesting picture: Conditionally on all components differing from zero, the distribution is simply Gaussian. More precisely, the Lasso behaves like a shifted version of the Least-squares estimator, with the direction of the shift depending on the sign of the estimator.

Also, looking at the area where at least one component is equal to zero, which coincides with the axes in the p=2 case, we see that the probability mass of $X'\varepsilon$ that lies in a neighborhood of that region is "compressed" into a lower-dimensional density. Indeed, in our example, the density conditional on the active set $\mathcal{A} = \{2\}$ qualitatively looks similar to a piece-wise Gaussian one. This is due to the fact that each point of the function $h$ is essentially obtained by integrating a part of a higher-dimensional normal distribution over an interval, thus reducing the dimension, a fact that will become even more apparent in the next section.

**Remark 27.** *Note that, using the same assumptions and arguments as in Proposition 11, one can obtain the asymptotic distribution of the conservatively-tuned Lasso estimator in a moving-parameter framework for an unknown error-distribution, if taking, for example, the same set of assumptions as in Section 2.6. This can be done by, again, considering the distribution of the asymptotic objective function problem defining the Lasso*[9]*, $V_\infty(u)$. Since the minimizer of the Least-squares part of the objective function converges to a Gaussian distribution, the limiting distribution*

[9]More precisely, $V_\infty$ is minimized at the Lasso's asymptotic estimation error.

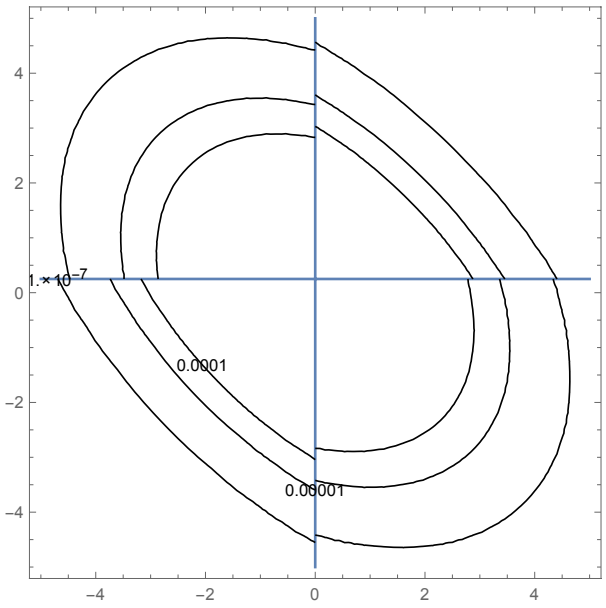



Figure 3.1: The contour lines of the the absolute continuous part of the distribution for $C = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$, $\lambda = (0.75, 0.75)'$ and $\beta = (0, -0.25)'$. Note that the area which contains probability mass that is not absolutely continuous with respect to the two-dimensional Lebesgue measure is displayed in blue.

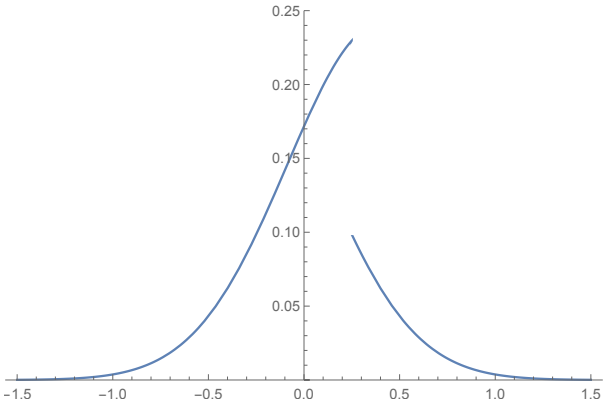



Figure 3.2: The functions $h^{(0,1)}$ and $h^{(0,-1)}$ for $C = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$, $\lambda = (0.75, 0.75)'$ and $\beta = (0, -0.25)'$.

*of the Lasso will look like the one in finite samples with Gaussian errors with the limits of the appropriately scaled parameters replacing their finite-sample counterparts.*

### 3.3.1 The shrinkage areas of the Lasso estimator

Using the conditions for minimality from the proof of Theorem 19, we can establish a direct relationship between the Least-squares estimator and the Lasso in the following sense: For any point $t \in \mathbb{R}^p$, there exists a set $S(t) \subseteq \mathbb{R}^p$, such that the Lasso estimator assumes the value $t$ if and only if the Least-squares estimator lies within the set $S(t)$. We refer to the set $S(t)$ as *shrinkage area* since the Lasso estimator can be viewed as a procedure that shrinks the Least-squares estimates from the set $S(t)$ to the point $t$. This is formalized in the following proposition.

**Proposition 28.** *For each $t \in \mathbb{R}^p$ there exists a set $S(t) \subseteq \mathbb{R}^p$, such that*

$$\hat{\beta}_{LS} \in S(t) \iff \hat{\beta}_L = t.$$

*Moreover, for $t \in \mathcal{O}^d$, the set $S(t)$ is given by*

$$S(t) = \{z \in \mathbb{R}^p : (Cz)_j = (Ct)_j + \operatorname{sgn}(t_j)\lambda_j \text{ for } j \in D_- \cup D_+, |(C(z-t)_j| \leq \lambda_j \text{ for } j \in D_0\}.$$

*Note that the sets $S(t)$ are disjoint for different $t$'s.*

*Proof.* Note that we have $\hat{\beta}_{\text{LS}} - \beta = (X'X)^{-1}X'\varepsilon = C^{-1}W$. With the minimality conditions (3.1) from the proof of Theorem 19 and the fact that $W = C(\hat{\beta}_{\text{LS}} - \beta)$ and some re-arranging we get that $m = \hat{\beta}_{\text{L}} - \beta$ minimizes $V$ if and only if $\hat{\beta}_{\text{LS}}$ satisfies

$$\begin{cases} (C\hat{\beta}_{\text{LS}})_j = (C\hat{\beta}_{\text{L}})_j - \lambda_j & \text{for } j \in D_- \\ (C\hat{\beta}_{\text{LS}})_j = (C\hat{\beta}_{\text{L}})_j + \lambda_j & \text{for } j \in D_+ \\ |C(\hat{\beta}_{\text{LS}} - \hat{\beta}_{\text{L}})_j| \leq \lambda_j & \text{for } j \in D_0, \end{cases}$$

or, $\hat{\beta}_{\text{LS}} \in S(t)$ for $\hat{\beta}_{\text{L}} = t$, as required. □

Given this result we can identify areas in which components of the Least-squares estimator are shrunk to zero by the Lasso. For $p = 2$, $C = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$ and $\lambda = (0.75, 0.75)'$ this leads to the picture displayed in Figure 3.3. Interestingly, the resulting shrinkage areas for active sets that exclude exactly one component are simply bands around the coordinate axes. They are, however, not symmetric around the axes in general. This is caused by the regressors' correlation which "distorts" the otherwise axes-parallel and symmetric sets.

The model selection probabilities for each component that are associated with the Lasso can thus also be calculated using the distribution of the Least-squares estimator (at least in the low-dimensional case), which is reflected by the following corollary.

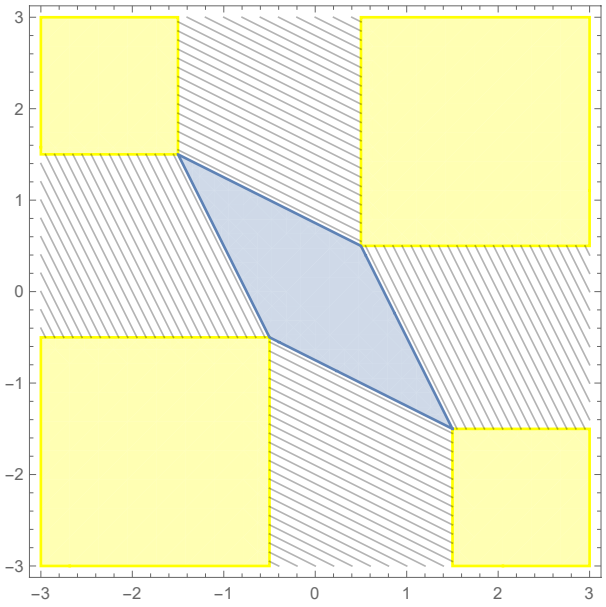



Figure 3.3: The shrinkage areas with respect to $\hat{\beta}_{\text{LS}}$ for $C = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$ and $\lambda_1 = \lambda_2 = 0.75$. The blue parallelogram equals $S((0,0)')$ and the yellow areas equal all cases where $\hat{\beta}_j \neq 0$ for each $j$. The dashed areas are the shrinkage areas for Lasso estimators with exactly one non-zero component. Note that each line-segment in the dashed area gives a shrinkage set for such a Lasso estimator.

**Corollary 29.**

$$P(\hat{\beta}_{L,j} = 0) = P(\hat{\beta}_{LS} \in \bigcup_{\substack{d \in \{-1,0,1\}^p: \\ d_j = 0}} \bigcup_{t \in \mathcal{Q}^d} S(t)).$$

Considering that the Lasso is the minimizer of a penalized version of the Least-squares objective function, it may not come as a surprise that there is a special relationship between the two procedures. However, the simplicity of the dependence might be a bit unexpected, as, after distorting the space by $C$, the components of the (transformed) Least-squares estimator, $C\hat{\beta}_{\text{LS}}$, are either shifted towards zero by a constant $\lambda_j$, or "collapse" to zero, in case the shift would exceed the distance from zero, when "translating" it to the (transformed) Lasso estimator, $C\hat{\beta}_{\text{L}}$. This exactly reflects the behavior of the Lasso in a one-dimensional setting.

While it is interesting in its own right, potential uses of this property are not immediately apparent, since the probably most obvious application would be another algorithm of computing the estimator. As any algorithm that makes use of this property would involve the calculation of the Least-squares estimator, one cannot gain anything in terms of computational efficiency, since current algorithms are not more (computationally) costly than computing the Least-squares estimator.

## 3.4 The high-dimensional case

In this setting we look at the case where $p > n$ implying that $\operatorname{rank}(X) < p$. Applying the same reasoning as in the case $p \leq n$, we can again give the distribution of the Lasso, albeit in a somewhat less explicit form. Note that in this case the true parameter is not identified without further assumptions and we denote by $\mathcal{B}_0$ the set of all $\beta \in \mathbb{R}^p$ that yield the model given in (3.2), that is, $\mathcal{B}_0 = \{\beta \in \mathbb{R}^p : X\beta = \mu\}$ where $\mu = \mathbb{E}(y)$. Furthermore, the Lasso solution need not be unique anymore without further assumptions on $X$ and $\lambda$.

Note that as the true parameter is not identified in this setting, the notion of an "estimation error" may seem quite arbitrary, as any point $\beta \in \mathcal{B}_0$ could be taken as a reference point to define the error. The distribution of $\hat{u}$ is thus only defined up to an arbitrary shift corresponding to the choice of $\beta \in \mathcal{B}_0$. It turns out, however, that the distribution of the estimator itself does not depend on the specific choice of $\beta \in \mathcal{B}_0$, but on $X\beta$ only. Indeed, for two parameters $\beta^{(1)}$ and $\beta^{(2)}$ satisfying $X\beta^{(1)} = X\beta^{(2)}$ we will get the same response $y$ for any fixed realization of the error-vector $\varepsilon$. It would thus be more natural to discuss the distribution of the Lasso estimator itself as the leading case, however, in order to maintain continuity with the previous parts of the thesis we shall proceed as in the low-dimensional case and give the distribution for any arbitrary (but fixed) $\beta \in \mathcal{B}_0$. We do not interpret this "estimation error" further, but rather view the quantity $\hat{u}$ as a technical construct that is used to obtain the result.

We start with a definition that will prove useful in the following derivations . For $m \in \mathbb{R}^p$, let

$$\mathcal{S}(m) = Cm + \Pi_{j=1}^{p} \mathcal{T}_j(m),$$

where

$$\mathcal{T}_j(m) = \begin{cases} \{\operatorname{sgn}(m_j + \beta_j)\lambda_j\} & m_j + \beta_j \neq 0 \\ [-\lambda_j, \lambda_j] & m_j + \beta_j = 0. \end{cases}$$

For a set $M \subset \mathbb{R}^p$, we define

$$\mathcal{S}(M) = \bigcup_{m \in M} \mathcal{S}(m).$$

Using these definitions, we are able to state a first high-level result on the estimator's distribution.

**Theorem 30.** *For any set $M \subseteq \mathbb{R}^p$ and any fixed $\beta \in \mathcal{B}_0$, we have*

$$P(\operatorname*{arg\,min}_{u \in \mathbb{R}^p} V(u) \cap M \neq \varnothing) = P(W \in \mathcal{S}(M)),$$

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

*Proof.* Using the same necessary and sufficient conditions for $m \in \mathbb{R}^p$ to be a minimizer of $V$ as in (3.1), we see that

$$m \in \arg\min V \iff W \in \mathcal{S}(m).$$

□

The sets $\mathcal{S}(M)$ can be viewed as shrinkage areas for the term $X'\varepsilon$ in analogy to the ones defined for the Least-squares estimator in Section 3.3.1. Also note that the set $\mathcal{S}$ depends on $\beta$, despite this being suppressed in the notation. This set can now be used to calculate the probability of a Lasso solution lying within a set.

**Corollary 31.** *Let the columns of $U \in \mathbb{R}^{p\times n}$ form a basis of* $\text{span}(X')$*. The probability that the estimation error for a fixed $\beta \in \mathcal{B}_0$ falls within a set $M$ can be calculated as*

$$P(\arg\min_{u\in\mathbb{R}^p} V(u) \cap M \neq \varnothing) = \mathbb{1}_{\{\text{span}(X')\cap\mathcal{S}(M)\neq\varnothing\}} \cdot \int_{U'\mathcal{S}(M)} \phi_{(0,\sigma^2U'CU)}(x)dx.$$

*Proof.* Take some $N \in \mathbb{R}^{p\times(p-n)}$ which satisfies $\text{rank}(N) = p - n$ and $N'X = 0$. Next, note that $N'W = 0$ almost surely, since $E(X'W)$=0 and $\text{Var}(N'W) = \sigma^2N'CN = 0$. We now have that

$$\begin{aligned} P(W \in \mathcal{S}(M)) &= P((U', N')W \in (U', N')\mathcal{S}(M)) \\ &= P(U'W \in U'\mathcal{S}(M) \text{ and } 0 \in U'\mathcal{S}(M)) \\ &= P(U'W \in U'\mathcal{S}(M)) \cdot \mathbb{1}_{\{\ker(N')\cap\mathcal{S}(M)\neq\varnothing\}} \\ &= \mathbb{1}_{\{\ker(N')\cap\mathcal{S}(M)\neq\varnothing\}} \cdot \int_{U'\mathcal{S}(M)} \phi_{(0,\sigma^2U'CU)}(x)dx. \end{aligned}$$

Noting that $\ker(N') = \text{span}(X')$ completes the proof. □

Assuming uniqueness[10] of the estimator, one can now easily state the cdf of the Lasso estimation error in the high-dimensional case.

**Corollary 32.** *If the Lasso estimator is unique and $U \in \mathbb{R}^{p\times n}$ is a matrix containing orthonormal columns in* $\text{span}(X')$ *then, for any fixed $\beta \in \mathcal{B}_0$, the cdf of $\hat{u} = \hat{\beta}_L - \beta$ is given by*

$$F(z) = \int_{U'\mathcal{S}(R(z))} \phi_{(0,\sigma^2U'CU)}(x)dx \cdot \mathbb{1}\{\text{span}(X') \cap \mathcal{S}(R(z)) \neq \varnothing\},$$

*where $R(z) = \{t \in \mathbb{R}^p : t_j \leq z_j \quad \forall j \in \{1, \ldots, p\}\}$.*

*Proof.* Note that

$$F_\beta(z) = P(\hat{u}_j \leq z_j \quad \forall j \in \{1, \ldots, p\}) = P(\hat{u} \in \mathcal{S}(R(z))),$$

and use Corollary 31. □

Given the above results we can now state the corresponding result on the distribution of the actual estimator, $\hat{\beta}_\text{L}$, by shifting this distribution by the true parameter $\beta$.

[10] Otherwise, the meaning of the term "cdf" would be rather- unclear. Necessary and sufficient conditions for uniqueness of the Lasso estimator with equal penalization weights are given, for example, in Ewald & Schneider (2020).

**Corollary 33.** *For any set $M \subseteq \mathbb{R}^p$ and any $\beta \in \mathcal{B}_0$, we have*

$$P(\arg\min_{\beta \in \mathbb{R}^p} L(\beta) \cap M \neq \varnothing) = P(W \in \tilde{\mathcal{S}}(M))$$

*where $W \sim N(0, \sigma^2 C)$ and $\tilde{\mathcal{S}}(M) = \bigcup_{m \in M} \tilde{\mathcal{S}}(m)$ with $\tilde{\mathcal{S}}(m) = C(m - \beta) + \Pi_{j=1}^{p} \tilde{\mathcal{T}}_j(m)$ and*

$$\tilde{\mathcal{T}}_j(m) = \begin{cases} \{\operatorname{sgn}(m_j)\lambda_j\} & m_j \neq 0 \\ [-\lambda_j, \lambda_j] & m_j = 0. \end{cases}$$

*In particular, the distribution of the estimator $\hat{\beta}_L$ does not depend on the choice of $\beta \in \mathcal{B}_0$.*

*Proof.* First note that $\arg\min_u V(u) = \arg\min_\beta L(\beta) - \beta$. We thus have that for any $t \in \mathbb{R}^p$,

$$t \in \arg\min_{\beta} L(\beta) \iff t - \beta \in \arg\min_{u} V(u).$$

Using Theorem 30, we get that

$$t - \beta \in \arg\min_{u} V(u) \iff W \in C(t - \beta) + \Pi_{j=1}^{p} \tilde{B}_j = \tilde{\mathcal{S}}.$$

Finally, note that $\tilde{\mathcal{S}}$ depends on $\beta$ only via $C\beta$ which assumes the same value for all $\beta \in \mathcal{B}_0$. □

We could now easily formulate statements on how to calculate these probabilities along the lines of Corollary 31 and on the cdf of the Lasso in case the estimator is unique. However, given that the formulae are not very explicit in any case, very little is to be gained by doing so and we shall thus refrain from that. Instead we will turn our attention to a more interesting implication of the preceding results. Using Corollary 33, we can see that some models will never be selected by the Lasso estimator.

**Remark 34.** *Let $\mathcal{B}_{\mathcal{M}} = \{\beta \in \mathbb{R}^p : \beta_j \neq 0 \iff j \in \mathcal{M}\}$ denote the parameter-space associated with some model $\mathcal{M} \subseteq \{1, \ldots, p\}$. Since $\operatorname{supp}(W) = \operatorname{span}(X')$, any model $\mathcal{M}$ which satisfies that $\tilde{\mathcal{S}}(\mathcal{B}_{\mathcal{M}}) \cap \operatorname{span}(X') = \varnothing$ will be selected with probability zero*[11]*. Indeed, since $C(m - \beta) = X'X(m - \beta) \in \operatorname{span}(X')$, we have*

$$\begin{aligned} \operatorname{span}(X') \cap \tilde{\mathcal{S}}(\mathcal{B}_{\mathcal{M}}) = \varnothing &\iff \operatorname{span}(X') \cap \bigcup_{m \in \mathcal{B}_{\mathcal{M}}} C(m - \beta) + \tilde{\mathcal{T}}(\mathcal{B}_{\mathcal{M}}) = \varnothing \\ &\iff \operatorname{span}(X') \cap \tilde{\mathcal{T}}(\mathcal{B}_{\mathcal{M}}) = \varnothing, \end{aligned}$$

[11] In fact, the corresponding active sets will never, for no $\omega \in \Omega$, be selected, since $X'\varepsilon$ will always, for each $\varepsilon \in \mathbb{R}^n$ (and hence each $\omega \in \Omega$), lie within $\operatorname{span}(X')$.

*where*

$$\tilde{\mathcal{T}}(\mathscr{B}_{\mathscr{M}}) = \bigcup_{m\in\mathscr{B}_{\mathscr{M}}} \Pi_{j=1}^{p} \begin{cases} [-\lambda_j, \lambda_j], \forall j \notin \mathscr{M} \\ \operatorname{sgn}(m_j)\lambda_j, \forall j \in \mathscr{M} \end{cases} = \Pi_{j=1}^{p} \begin{cases} [-\lambda_j, \lambda_j], \forall j \notin \mathscr{M} \\ \{-\lambda_j, \lambda_j\}, \forall j \in \mathscr{M}. \end{cases}$$

*Note that for all models containing some regressor, i.e., for all $\mathscr{M} \neq \varnothing$, $\tilde{\mathcal{T}}(\mathscr{M})$ is the union of parallel faces of the p-dimensional $\lambda$-box and for $\mathscr{M} = \varnothing$, $\tilde{\mathcal{T}}(\mathscr{M})$ is simply the $\lambda$-box itself. This implies that the model $\mathscr{M} = \varnothing$ will always have positive probability, if $\lambda_j > 0$ for all $j = 1, ..., p$.*

Remark 34 shows that, in the high-dimensional setting, model selection by the Lasso estimator is indeed not entirely data-driven in the sense that there is a *structural active set*, $\mathcal{A}^S = \{j : P(\hat{\beta}_{\mathrm{L},j} \neq 0) > 0\} \subseteq \{1, \ldots, p\}$ which is only determined by the regressor-matrix $X$ and the penalization-vector $\lambda$. In particular, the true parameter $\beta$ as well as the error $\varepsilon$ do not have any[12] influence on this set. In other words, some models are not being considered by the model selection procedure even before "looking" at the response vector $y$. Considering the problem from a different angle, this means that one can restrict, or implicitly choose, the class of models considered by the selection procedure by the choice of $\lambda$, or the scaling of the regressors. This shows that the scaling of the regressors and the choice of the penalization parameters have a great influence on the model selection properties of the Lasso and that it should thus be chosen with great care, as this means to make an implicit choice of the models under consideration. From Remark 34, we can furthermore see that the active sets which are being excluded by the model selection procedure merely on basis of $X$ and $\lambda$ are determined by observing which faces of the $\lambda$-box, $\Pi_{j=1}^{p}[-\lambda_j, \lambda_j]$, do not intersect the the span of $X'$, an $n$-dimensional linear subspace of $\mathbb{R}^p$. This will be made more transparent in the following simple example.

**Example 35.** *Suppose $\beta = (0, 0)'$, $X = (1, 2)$ (hence, $n = 1$ and $p = 2$). Moreover, assume that $\varepsilon \sim N(0, 1)$. Take $\lambda_1 = \lambda_2$. Note that $X'\varepsilon = (\varepsilon, 2\varepsilon)' \in \operatorname{span}(X')$. It is easily seen that*

$$\mathcal{S}(m) \cap \operatorname{span}(X') = \varnothing$$

*whenever $m_1 \neq 0$ implying that $\hat{\beta}_{L,1} = 0$ almost surely. Note that this property does not depend on the distribution of $\varepsilon$, but is merely determined by $\lambda$ and $X$. Next, we see that*

$$\begin{aligned} P(\hat{\beta}_L = 0) &= P((\varepsilon, 2\varepsilon)' \in [-\lambda_1, \lambda_1] \times [-\lambda_1, \lambda_1]) \\ &= P(2\varepsilon \in [-\lambda_1, \lambda_1]) \\ &= \Phi(\tfrac{\lambda_1}{2}) - \Phi(-\tfrac{\lambda_1}{2}). \end{aligned}$$

[12] At least as long as one assumes that the distribution of $\varepsilon$ has support on all of $\mathbb{R}^n$.

*Next, we have that for* $t < 0$,

$$\begin{aligned} P(\hat{u}_2 < t, \hat{u}_1 = -\beta_2) &= P(\varepsilon \leq t - \tfrac{\lambda_1}{2} \text{ and } 2\varepsilon \in \bigcup_{l<t} \{z : z_1 = 2l - \lambda_1 \text{ and } 4l - \lambda_1 \leq z_2 \leq 4l + \lambda_1\}) \\ &= P(\varepsilon \leq 2t - \tfrac{\lambda_1}{2}) \\ &= \Phi(2t - \tfrac{\lambda_1}{2}). \end{aligned}$$

*Similarly, we get that for* $t > 0$,

$$P(\hat{u}_2 > t, \hat{u}_1 = -\beta_2) = 1 - \Phi(2t + \tfrac{\lambda_1}{2}).$$

*The distribution of* $\hat{\beta}_L$ *is thus given by*

$$\hat{\beta}_{L,1} =_{as} 0$$

*and* $\hat{\beta}_{L,2}$ *follows the distribution*

$$dF(t) = \left(\Phi(2t + \tfrac{\lambda_1}{2}) - \Phi(2t - \tfrac{\lambda_1}{2})\right) \delta_0(t) + \phi(2t - \tfrac{\lambda_1}{2}) \mathbb{1}_{\{t<0\}} dt + \phi(2t + \tfrac{\lambda_1}{2}) \mathbb{1}_{\{t>0\}} dt.$$

*It is interesting to note that the distribution of* $\hat{\beta}_{L,2}$ *is the same as the one of the Lasso estimator*[13] *in the smaller model,* $y_i = 2\beta_2 + \varepsilon_i$, *where the first regressor is left out. Indeed, using the Lasso in the smaller model is quite easily seen to be equivalent to using the Lasso in the larger model in our example, since the procedure only actually considers models that do not contain the first regressor. This fact is, of course, only valid for the specific form of* $X$ *and* $\lambda$. *It illustrates the fact that the models which are being considered by the Lasso estimator with positive probability do not depend on* $\beta$ *and* $\varepsilon$, *as pointed out in Remark 34. The Lasso is thus not to be viewed as a purely data-driven model selection procedure in a high-dimensional setup. However, note that the choice between the mean-model (*$\beta_1 = \beta_2 = 0$*) and the single-regressor model (*$\beta_1 = 0$ *and* $\beta_2 \neq 0$*) does very much depend on* $\beta$ *and* $\varepsilon$. *The shrinkage areas for the estimator in this example are displayed in Figure 3.4 along with the area the probability mass of* $X'\varepsilon$ *is concentrated on. Note that a shrinkage area for* $X'\varepsilon$ *corresponding to a Lasso estimator with exactly one non-zero component is given by a line-segment that is parallel to one*[14] *of the coordinate axes. Hence, each dash in the figure actually gives a shrinkage set for such a Lasso estimate. The figure underlines the fact that this example's setup only produces models where* $\hat{\beta}_{L,1} = 0$.

Assuming the existence of such a structural active set that has cardinality at most[15] $n$, it is easily seen that, using the Lasso in the high-dimensional model is equivalent to using the Lasso with the same weights in the corresponding low-dimensional model. This entails that in such cases

[13] With the same penalization parameter.

[14] More precisely, it is parallel to the x-axis for $\hat{\beta}_{L,1} = 0$ and parallel to the y-axis for $\hat{\beta}_{L,2} = 0$.

[15] Note that this is not nessecarily the case, since an $n$-dimensional linear subspace may intersect more than $n$ faces of the $\lambda$-box.

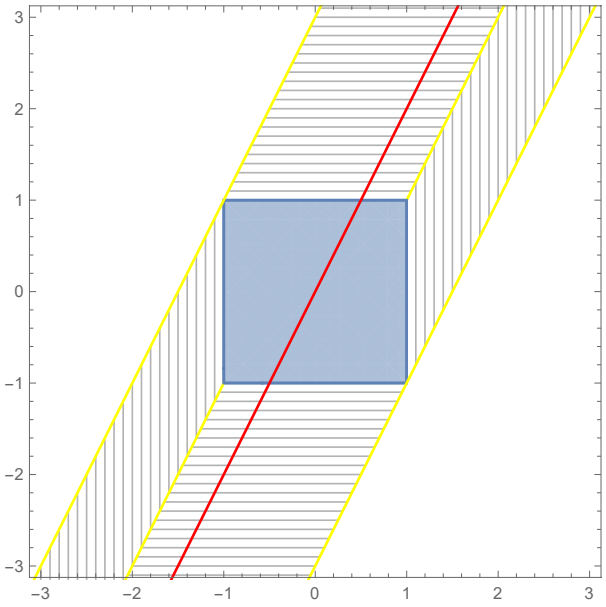



Figure 3.4: The shrinkage areas with respect to $X'\varepsilon$ with the span of $X'$, i.e., the area on which the probability mass of $X'\varepsilon$ is concentrated, being displayed in red. The area corresponding to the zero-estimate, $\hat{\beta}_{\mathrm{L}} = (0,0)'$, is displayed in blue, while the yellow areas correspond estimates with both components differing from zero. The areas corresponding to estimates with exactly one zero-component are displayed as dashed areas (horizontally for $\hat{\beta}_{\mathrm{L},1} = 0$ and vertically for $\hat{\beta}_{\mathrm{L},2} = 0$).

the Lasso solution yields a model in which the parameter of interest is identifiable despite the model's high-dimensional nature. If the cardinality of the structural active set is exactly equal to the sample size $n$, one can construct confidence sets for the parameter *given the model corresponding to the structural active set*[16] using the results from Chapter 2. Note that in this case, this target is simply one particular point from the set of true parameters $\mathcal{B}_0$. This is described in more detail in Section 3.4.1, but to first illustrate the idea, consider Example 35 again:

**Example 35** (continued)**.** *Here, $\mathcal{A}^S = \{2\}$ and the corresponding model is given by*

$$y_i = 2\beta_2 + \varepsilon_i.$$

*As mentioned before, the distribution of the Lasso estimator in this is model is equal to the distribution of the Lasso estimator's second component in the model containing both regressors. Using this property, we can construct a confidence set for $\beta_2$, the parameter in the model containing only the second regressor. In the single-regressor model we can find the optimal confidence set (in terms of size) using the results from Pötscher & Schneider (2009).*

**Remark 36.** *Inspecting the* shrinkage sets *for $X'\varepsilon$ from Theorem 30, $\mathcal{S}(m)$, more closely, we see that they are, in fact, not necessarily disjoint for different $m$'s. Indeed, we will have different*

[16] In the sense of the PoSI target, as described in Section 2.2.

*active sets corresponding to the same value of $X'\varepsilon$ in some situations. To see this note that, since $\mathcal{S}(m) = Cm + \mathcal{T}(m)$, we can take $t \in \ker(X')$ and consider $\mathcal{S}(t+m) = Cm + \mathcal{T}(t+m)$, where $\mathcal{T}(t+m)$ merely depends on the signs of $\beta + m$. This entails that any $m + t$ satisfying that $\operatorname{sgn}(\beta + m) = \operatorname{sgn}(\beta + m + t)$ will also be a solution to the Lasso objective function. Moreover, it is easily seen that these sets are, in general, not disjoint, since clearly, $\mathcal{T}_j(m) \subseteq \mathcal{T}_j(\tilde{m})$ whenever $\tilde{m}_j = 0$ and $|m_j| \geq 0$ for each $j = 1, \ldots, p$ and the inclusion being strict whenever $\tilde{m}_j = 0$ and $m_j \neq 0$ for at least one $j$.*

In Example 35 the Lasso solution is always, for each value of $X'\varepsilon$, unique, since the linear span of $X'$ does not intersect the areas that lie in the (non-empty) intersection of shrinkage areas that correspond to different active sets. It is not difficult, however, to construct an example where the Lasso solution is not unique anymore, as is seen in the following.

**Example 37.** *Again, take the model from Example 35 with $X = (1, 2)$, $\beta = (0, 0)'$ and suppose $\varepsilon \sim N(0, 1)$. Choose $\lambda = (1, 2)'$. Note that $\lambda \in span(X')$ and thus, for each $\varepsilon > 0$,*

$$X'\varepsilon = (\varepsilon, 2\varepsilon)' \in \mathcal{S}\left((\varepsilon + \lambda_1, 0)'\right) = (\varepsilon + 1, 2\varepsilon + 2)' + \{1\} \times [-2, 2],$$

*but also*

$$X'\varepsilon \in \mathcal{S}\left((0, \tfrac{\varepsilon}{2} + \tfrac{\lambda_2}{4})'\right) = (\varepsilon + 1, 2\varepsilon + 2)' + [-1, 1] \times \{2\}.$$

*Note that we now have one solution where the first component equals zero and another one where the second component is zero. By convexity of the Lasso objective function it follows that each convex combination of the above minimizers also is a minimizer, yielding a continuum of solutions for which both components differ from zero. This is reflected in the following display where for any $\psi \in [0, 1]$ we have that*

$$X'\varepsilon \in \mathcal{S}\left(\left(\psi(\varepsilon + \lambda_1), (1 - \psi)(\tfrac{\varepsilon}{2} + \tfrac{\lambda_2}{4})\right)'\right) = (\varepsilon + 1, 2\varepsilon + 2) + \{1\} \times \{2\}.$$

*Noting that choosing $\psi \notin [0, 1]$ would yield $\mathcal{T}_j = -\lambda_j$ for one $j \in \{1, 2\}$ in the previous display, we see that the convex hull of the points $(\varepsilon + \lambda_1, 0)'$ and $(0, \frac{\varepsilon}{2} + \frac{\lambda_2}{4})'$ indeed contains all solutions to the Lasso problem whenever $\varepsilon > 0$.*

*The ambiguity of the estimator in this example is also reflected in Figure 3.5 which illustrates how the span of $X'$ intersects the (non-empty) intersection of the shrinkage areas containing only the first, only the second, and the shrinkage areas corresponding to models that contain both regressors.*

Example 37 shows an already known property of the Lasso from another perspective: The solution to the Lasso problem is in general not unique. Moreover, if the solution is not unique, then, by convexity of the problem, there exists a continuum of solutions, a fact that has already been pointed out by Tibshirani (2013). The example also shows that the set of $\varepsilon$'s which give ambiguous Lasso-solutions is not necessarily a null set with respect to the distribution of $X'\varepsilon$.

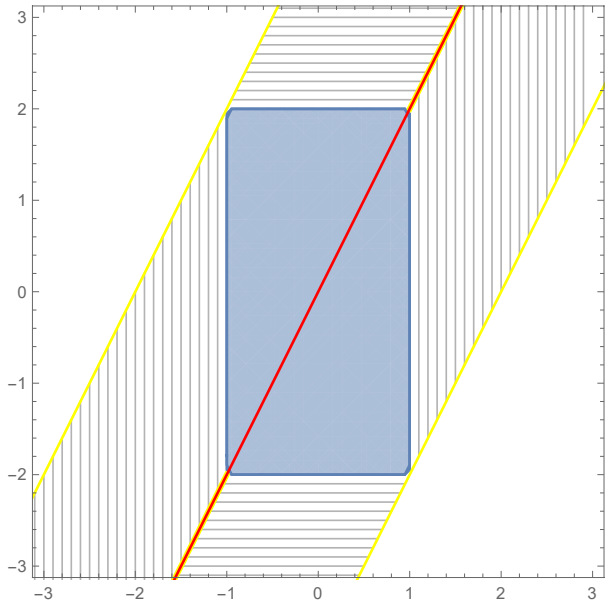



Figure 3.5: The shrinkage areas with respect to $X'\varepsilon$ in Example 37. The span of $X'$, i.e., the area on which the probability mass of $X'\varepsilon$ is concentrated, is displayed in red. The area corresponding to the zero-estimate, i.e., $\hat{\beta}_{\mathrm{L}} = (0,0)'$, is displayed in blue, while the yellow areas correspond to estimates with both components differing from zero. The areas corresponding to estimates with exactly one zero-component are displayed as dashed areas (horizontally for $\hat{\beta}_{\mathrm{L},1} = 0$ and vertically for $\hat{\beta}_{\mathrm{L},2} = 0$).

In our example the problem of a non-unique solution of the Lasso could be overcome by slightly altering the weights of the estimator. One should bear in mind, however, that this would mean to implicitly make a choice about the class of models under consideration via the structural active set, as pointed out in Remark 34.

Clearly, Example 37 shows that the structural active set may be equal to the entire set of explanatory variables. It is quite obvious that in this simple $n = 1$, $p = 2$ example, the Lasso estimator will always have a structural active set of dimension one, whenever it is unique. In this case uniqueness of the estimator and the existence of a structural active set with cardinality $n$ go hand in hand. It turns out, however, that this is not the case in general. Remark 34 shows that the structural active set can be determined by looking at how many faces of the $p$-dimensional $\lambda$-box are intersected by the $n$-dimensional linear subspace that is spanned by the columns of the design matrix. Increasing the dimensions for both $n$ and $p$ we can quite easily construct an example where the structural active set is equal to the entire set of regressors and the estimator is unique.

**Example 38.** *Take* $X = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right)$ *and* $\lambda = (1, 1, 1)'$. *Moreover, assume that* $\beta = (0, 0, 0)'$ *and that* $\varepsilon \sim N((0, 0)', \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right))$. *Note that the columns of* $X$ *are in general position*[17]. *By Lemma 3 from Tibshirani (2013) the Lasso solution will thus be unique. When looking at the "active"*[18] *shrinkage areas with respect to* $X'\varepsilon$, *which are displayed in Figure 3.6, we see that each of the three regressors could end up in the model with positive probability*[19]. *To conclude this example, observe that in this case, even though every regressor will appear in in some model with positive probability, not all models are chosen with positive probability. Indeed, note that by Remark 34 the resulting Lasso estimator will never have three non-zero components, as no corner of the* $\lambda$*-box is intersected by the linear span of* $X'$. *Also note that even though each facet of the* $\lambda$*-box is intersected by the linear span of* $X'$, *implying that each component of the estimator could be positive, or negative, not all sign-combinations of the estimator's components are possible.*

We will now consider an example where $p = 3$ and $n = 2$ in which the structural active set contains exactly two regressors.

**Example 39.** *Take* $X = \left(\begin{smallmatrix} 1 & 0 & 1 \\ 4 & 1 & 0 \end{smallmatrix}\right)$ *and* $\lambda = (1, 1, 1)'$. *Moreover, assume that* $\beta = (0, 0, 0)'$ *and that* $\varepsilon \sim N((0, 0)', \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right))$ *and note that, as in the previous example, the columns of* $X$ *are in general (affine) position and the Lasso solution will thus be unique for each realization of* $\varepsilon \in \mathbb{R}^2$.

*Looking at the intersections of the shrinkage areas for* $X'\varepsilon$ *with the span of* $X'$, *which are displayed in Figure 3.7, reveals that, in this case, one of the regressors will never enter the model. Also note that in this example, the each sign-combination of the parameters that are contained in the structural active set is possible, as the estimator will behave like a low-dimensional Lasso.*

[17] C.f. Definition 54 in Appendix C.1.

[18] I.e., the ones that have an non-empty intersection with the span of $X'$.

[19] Note that since $X'\varepsilon$ follows a normal distribution that is supported on the span of $X'$, each point in the span of $X'$ carries positive probability density.

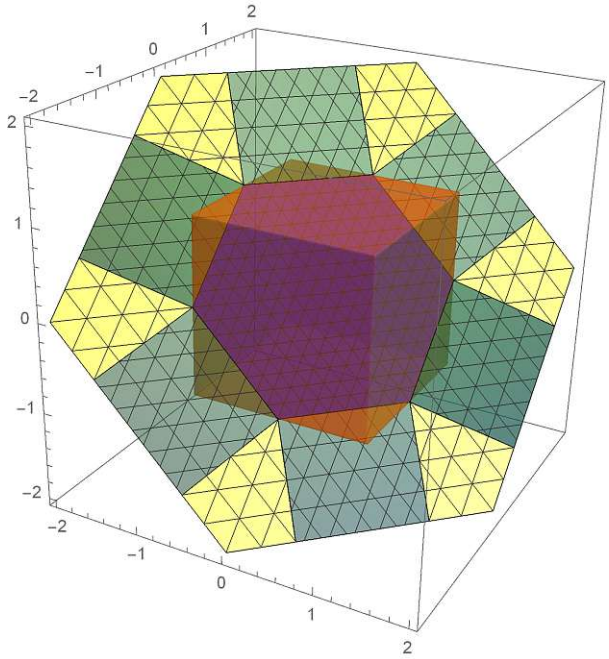



Figure 3.6: The intersection of the shrinkage areas with respect to $X'\varepsilon$ and the span of $X'$ along with the $\lambda$-cube from Example 38. The shrinkage areas corresponding to single-regressor models are displayed in grey, while the shrinkage areas that correspond to two-regressor models are displayed in yellow. The intersection of the $\lambda$-cube with the span of $X'$, which corresponds to the zero-estimator, is displayed in blue. The $\lambda$-cube itself is displayed in orange.

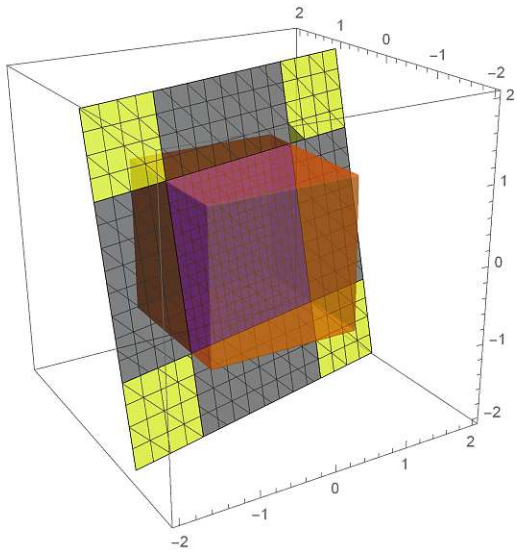



Figure 3.7: The intersection of the shrinkage areas with respect to $X'\varepsilon$ and the span of $X'$ along with the $\lambda$-cube from Example 39. The shrinkage areas corresponding to single-regressor models are displayed in grey, while the shrinkage areas that correspond to two-regressor models are displayed in yellow. The intersection of the $\lambda$-cube with the span of $X'$, which corresponds to the zero-estimator shrinkage set, is displayed in blue. The $\lambda$-cube itself is displayed in orange.

The examples in this section give a brief overview of the Lasso estimator's (structural) model selection properties. In these simple examples it is illustrated how the interaction of the regressor matrix $X$ and the penalization vector $\lambda$ influence the estimator's structural properties, i.e., uniqueness and the estimator's structural active set. The geometric condition that characterizes the structural active set that is given in Remark 34 is quite easy to understand on an intuitive level and has been further formalized in Ewald & Schneider (2020), who give necessary and sufficient conditions on the Lasso's uniqueness that are based on the above properties. Interestingly, these geometric arguments can be generalized to penalized Least-squares estimators whose penalty term takes the form of any norm with a polytope shape, as is done in Schneider & Tardivel (2020).

Another reference that deals with the topic of *accessible* models, i.e., models that are chosen by the Lasso, is Sepehri & Harris (2017) who give a condition for when this is the case. In contrast to the condition provided here, which is given on $\mathbb{R}^p$, that reference uses geometric considerations in $\mathbb{R}^n$ under a uniqueness assumption.

The theme of variables (not) being selected also appears in El Ghaoui et al. (2012); Ndiaye et al. (2017) and Tibshirani et al. (2012) who provide the so-called *SAFE* and *STRONG* rules, respectively. The difference is that these rules specify sets of variables that do not enter any Lasso solution, for no value of the tuning parameter and *for a given response* $y$, whereas the structural active set only depends on $X$ and $\lambda$.

We now turn to an interesting implication of the concept of structural active sets.

### 3.4.1 A note on high-dimensional confidence sets

Dealing with high-dimensional models again raises the question of which parameter, or parameters, should be the targeted by our analysis, since the true parameter is not identified in the high-dimensional model.

Interestingly, the PoSI target (see Section 2.2) can be used in the high-dimensional regression setting to develop a method of inference that is compatible with classical theory. In Example 35 from Section 3.4, there exists a structural active set $\mathcal{A}^S$ containing exactly $n$ regressors. Under conditions in which this property holds more generally, the parameter is identified in the model containing the regressors that are contained in the structural active set, if we assume that $\operatorname{rank}(X_{\mathcal{A}^S}) = n$, i.e., if all columns of $X$ corresponding to the structural active set, are linearly independent.

Moreover, considering this parameter solves the identifiability issues of the high-dimensional setup, since, using the notation from Section 2.2, the parameter $\tilde{\beta}_{\mathcal{A}^S}$ is identified in the model

$$y = X_{\mathcal{A}^S}\tilde{\beta}_{\mathcal{A}^S} + \epsilon. \tag{3.2}$$

Additionally, in this case the parameter $\tilde{\beta}_{\mathcal{A}^S}$ is a true parameter in the full model

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

in the sense that $X_{\mathcal{A}^S}\tilde{\beta}_{\mathcal{A}^S} = X\beta$ and thus, the error vector from model (3.2) is equal to the error vector from the full model: $\epsilon = \varepsilon$.

To conclude this section, we note that in cases where the combination of $X$ and $\lambda$ yields a setting in which the Lasso estimator possesses a structural active set with exactly $n$ regressors, the use of the regression parameter given the corresponding model, as proposed by Berk et al. (2013), yields a well-interpretable model that is also compatible with classical statistical theory. Moreover, given such a structural active set, the Lasso would be equivalent to a minimization of the Lasso objective function restricted to certain values being equal to zero, which in turn is equivalent to using a corresponding[20] Lasso in the low-dimensional model containing the sub-matrix of $X$ that corresponds to the structural active set, as is to be seen in Example 35. Furthermore, this would enable the construction of confidence sets for the parameter in the model corresponding to the structural active set, i.e., $\beta_{\mathcal{A}^S} = (X'_{\mathcal{A}^S}X_{\mathcal{A}^S})^{-1}X'_{\mathcal{A}^S}\beta$ (for any $\beta \in \mathcal{B}_0$), using the methods presented in Chapter 2. One has to keep in mind, however, that this kind of procedure would be feasible in a setup yielding an implicit[21] non-random selection of a low-dimensional sub-model and that any other choice of submodel with size $n$ would also yield a valid model, irrespective of the way the sub-model is chosen, at least as long as the selection procedure is non-random.

## 3.5 Discussion and conclusion

In this chapter we have analyzed the distribution of the Lasso estimator in finite samples, while the corresponding asymptotic distribution can be obtained in a similar fashion when additionally using arguments from Chapter 2. In a low-dimensional setting it is shown that the Lasso estimator creates so-called shrinkage areas inside of which some the Least-squares estimator's components are shrunk to zero, while the other components are shifted towards zero. This implies that the Least-squares estimator's distribution's probability mass is "compressed" into lower-dimensional densities that can be specified conditional on the Lasso's active set. As a result the distribution looks like a pieced-together combination of Gaussian-like densities, whereby each active set has its own distribution-piece whose dimension depends on the number of non-zero components. The dimension of each of these densities is given by the number of non-zero components, resulting in point-mass at the origin. Moreover, the concept of shrinkage areas establishes a unique relationship between the least-squares estimator and the Lasso.

The form of the distribution is even more intricate in the high-dimensional case in which the

[20] In the sense that the remaining components' penalization terms are identical to the ones in the full model.
[21] By the choice of $\lambda$ for a given $X$, for example.

estimator may not be unique anymore. In this case, one can again specify shrinkage areas, this time for the term $X\beta$. However, this quantity's probability mass is concentrated on an $n$-dimensional linear subspace of $\mathbb{R}^p$. This fact leads to a valuable insight into the behavior of the estimator, since it is shown that some models will never be selected by the estimator. In fact, it is shown that what models are excluded merely depend on the regressor matrix and tuning-vector and are thus not influenced by the observed response data. In some cases the Lasso estimator just selects a sub-model of a predetermined model, with this structural pre-selection being determined by the regressor matrix and the tuning-vector. In case this pre-determined *structural active set* is low-dimensional, the estimator simply acts like a low-dimensional Lasso in the model containing the structurally active components. This opens up possibilities of obtaining well-interpretable confidence regions for the high-dimensional case using the results from the previous chapter, at least for certain regressor matrices and tuning-vectors. However, it is also shown that the structural active set is not necessarily low-dimensional and it remains to be seen whether a well-interpretable and easy-to-verify condition on $X$ and $\lambda$ can be derived under which this property holds.

# Chapter 4

# Extensions

## 4.1 Introduction

The final chapter provides some extensions to the preceding ones, mostly on the topic of confidence sets in a low-dimensional setup. Most considerations are conducted in simple settings and should give a taste of how the previously developed methods can or cannot be applied to, or refined for, other settings.

In this chapter's first section we extend the finite-sample results on the confidence sets' coverage probabilities from Chapter 2 to the more realistic case of unknown error-variance. The analysis will show that the results from the previous chapter carry over to this case with only a slight adjustment of the confidence set's size parameter which is necessary to account for the estimation of the error-variance.

In Section 4.3 we discuss the question of how to construct confidence intervals for single components of the true parameter that are based on the Lasso estimator using the results from Chapter 2. To deal with the case where one is primarily interested in just a sub-vector of the true parameter and hence may not want to penalize these particular components, thus effectively excluding them from model selection, the *partial Lasso* estimator is considered in addition to the fully penalized one as starting point for the development of such confidence intervals. The main challenge in this section is to find the optimal shape for such a confidence set, i.e., a set that complies with Condition A from Chapter 2 while giving the smallest possible projection onto the subspace associated with the component of interest and maintaining the desired level of coverage.

Another question being investigated is the one of whether it is valid to choose the sub-parameter to be covered based on the Lasso estimator, i.e., to construct a confidence set for the non-zero estimates only. Using the results of Section 4.3, a simulation study is carried out in a $p = 2$ setting to determine whether this proposed procedure, which will certainly appear quite attractive to many users, can yield valid confidence sets.

Moreover, if some parameters appear to be very large, one might be tempted to try to use this information to conclude at least the sign of the true parameter of interest and subsequently take this knowledge into account when choosing the shape of the confidence set. However, it is shown in a simple setting that such an adaptive procedure will not yield uniformly valid methods of inference in a moving-parameter framework, even in large samples, as $n \to \infty$.

Since the sections of this final chapter do not have a common sub-setting of the one introduced in Section 1.2, the assumptions corresponding to the setups in the sub-sections are presented where appropriate.

## 4.2 Confidence sets for unknown error-variance

Throughout Chapter 2 we have assumed the error-variance $\sigma^2$ to be known. The question now arises, whether the results obtained in that chapter can also be used if the error-variance is estimated. To that end recall the finite-sample setting from Section 2.3. As a variance-estimator we consider the unbiased estimator based on the Least-squares residuals:

$$\hat{\sigma}^2 = \frac{1}{n-p} \hat{\varepsilon}'_{\text{LS}} \hat{\varepsilon}_{\text{LS}}$$

where $\hat{\varepsilon}_{\text{LS}} = y - X\hat{\beta}_{\text{LS}}$. To apply the previous results to this setting, we need to make the penalty-vector, $\lambda$, depend on the variance-estimate in the following way. For this subsection, set

$$\lambda = \hat{\sigma} \cdot l,$$

where $l \in \mathbb{R}^p_+$ some fixed penalization vector[1]. As already mentioned in Remark 5, the characterization of the minimal coverage probability in Theorem 4 does not depend on the stochastic properties of $\varepsilon$, but on the minimization problem alone. Hence, the parts of the theorem's proof which give the lower bound for the coverage probability also apply in the case we are considering now. What is left to be done is to determine the distributions of the $\hat{u}^d$'s that appear in the lower bound's formula. Not too surprisingly, these $\hat{u}^d$'s turn out to follow (multivariate) t-distributions instead of normal distributions, such that the confidence set has to be scaled slightly differently. This is formalized in the following proposition.

**Proposition 40.** *For $\lambda = \hat{\sigma} \cdot l$ and any non-random $M \subseteq \mathbb{R}^p$ that complies with Condition A we have that*

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

*where $\hat{t}^d \sim T_{n-p}(C^{-1}, C^{-1}\,\text{diag}(l)d)$ is a multivariate t-distribution[2] with $n-p$ degrees of freedom, correlation matrix $C^{-1}$ and non-centrality parameter $C^{-1}\,\text{diag}(l)d$.*

[1] Note, however, that the vector $l$ may very well depend on $n$, but is fixed for each $n \in \mathbb{N}$.

[2] As in Definition 52 in Appendix C.1.1.

*Proof.* The first equality follows from the same arguments as in the proof of Theorem 4.

Since $\{\hat{u}^d \in \hat{\sigma} M\} = \{\hat{\sigma}^{-1}\hat{u}^d \in M\}$, we simply need to determine the distribution of $\hat{t}^d = \hat{\sigma}^{-1}\hat{u}^d$. We have

$$\hat{t}^d = \hat{\sigma}^{-1} C^{-1} W + C^{-1} \operatorname{diag}(l) d.$$

Noting that $W \sim N(0, \sigma C)$ as well as $\sigma^{-2}\hat{\varepsilon}'_{\mathrm{LS}}\hat{\varepsilon}_{\mathrm{LS}} \sim \chi^2_{n-p}$, one sees that $\hat{t}^d$ indeed follows a multivariate t-distribution with $n-p$ degrees of freedom, correlation matrix $C^{-1}$ and non-centrality parameter $C^{-1} \operatorname{diag}(l) d$. □

Using Theorem 4 and Proposition 40 one can now construct valid confidence set in cases where $\sigma^2$ is estimated by $\hat{\sigma}^2$ assuming i.i.d. Gaussian errors. Noting that the contour sets of a normal distribution with variance-covariance matrix $\Sigma$ and a multivariate t-distribution with the same variance-covariance matrix have the same shape, i.e., they are both of the form $\{z \in \mathbb{R}^p : z'\Sigma z \leq k\}$, we see that all considerations with respect to the shape of the confidence sets from Section 2.5 also apply in this setting. The only difference in the resulting confidence sets will be the size parameter $k$, since a contour-set of the above-mentioned distributions will correspond to different[3] contour-levels.

## 4.3 Inference on single components

As users are sometimes interested in only a few (or even just a single) regressors' effects we will now consider the case where only one of the true parameter's components is supposed to be covered by a confidence set. Hereby we will examine two regimes. One in which each of the estimator's components is penalized, hereafter referred to as the *full penalization case*, and one where the component of interest is not penalized, which we refer to as *partial penalization*[4]. We will start with the former one.

### 4.3.1 Full penalization case

Recall model (1.1). We consider the case in which all components are penalized, i.e., $\lambda_j > 0$ for all $j \in \{1, \ldots, p\}$. Suppose $\operatorname{rank}(X) = p$ (implying that $p \leq n$).

In this case Theorem 4 can be used to construct a confidence set for one single component of the true parameter vector, say $\beta_1$, in the following way: Given a desired level of coverage $1-\alpha$, we want to find a set $M \subseteq \mathbb{R}^p$ which complies with Condition A and satisfies that its projection onto the subspace associated with the component of interest is as short as possible, while the parameter of

[3] For more details on this refer to Appendix C.1.1.

[4] A Lasso with a tuning parameter that only penalizes some components is also referred to as a *partial Lasso*.

interest is uniformly[5] covered with the desired probability $1-\alpha$. For simplicity's sake, we restrict[6] ourselves to symmetric intervals. More formally, the set $M$ should be chosen such that

- $M \subseteq \mathbb{R}^p$ satisfies Condition A.
- $\max_{m \in M} |m_1| \leq a$ for some $a > 0$.
- $\inf_{\beta \in \mathbb{R}^p} P_\beta(\hat{u} \in M) \geq 1-\alpha$ for some $0 < \alpha < 1$.

To find such a set, one may first determine the smallest set that complies with Condition A for each fixed $a > 0$. Next, $a$ is chosen such that the desired coverage probability is attained. Note that formally we still construct a confidence set for the entire parameter vector, even though this set may be unbounded with respect to the components which are not of primary interest. The analysis will reveal, however, that this will only be the case when the regressors are orthogonal.

**Constructing the optimal shape in case p=2:**

Throughout the following suppose that $p = 2$ such that $C = \left(\begin{smallmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{smallmatrix}\right)$ and suppose that $C$ is positive definite. Assume, without loss of generality[7], that $c_{12} \geq 0$ . In case $c_{12} = 0$, it is easily seen that the set

$$\overline{M} = \{z \in \mathbb{R}^2 : |z_1| \leq a\}$$

complies with Condition A and cannot be enlarged while maintaining a fixed projection onto the subspace associated with the first component. Also note that in this case the confidence interval constructed in the above-described way will be equivalent to the procedure proposed by Pötscher & Schneider (2010).

We thus turn to the more interesting case where $c_{12} > 0$. Defining our (prospective) confidence set on each quadrant $\mathcal{O}^d$ separately, let the part of the set that lies in the orthant $\mathcal{O}^d$ be denoted by $M^d$, such that

$$M = \bigcup_{d \in \{-1,1\}^2} M^d.$$

We define these four parts in the following way: Take

$$M^\iota = M \cap \mathcal{O}^\iota = \{z \in \mathcal{O}^\iota : |z_1| \leq a\} \cap \{z \in \mathcal{O}^\iota : (Cz)_1 \leq (C\underline{a})_1\}$$

with $\underline{a} = (a, 0)'$, $a > 0$. Next, define

$$M^{(-1,1)} = M \cap \mathcal{O}^{(-1,1)} = \{z \in \mathcal{O}^{(-1,1)} : |z_1| \leq a\} \cap \{z \in \mathcal{O}^{(-1,1)} : (Cz)_2 \leq (C\underline{b})_2\}$$

[5] For all possible values of $\beta$.

[6] Note that the generalization to asymmetric intervals would be straight-forward given the results that follow in this section.

[7] Otherwise construct a confidence interval for $\beta_1$ from the model $y_i = \beta_1 x_{i1} + \tilde{\beta}_2 \tilde{x}_{i2} + \varepsilon_i$ where $\tilde{\beta}_2 = -\beta_2$ and $\tilde{x}_{i2} = -x_{i2}$.

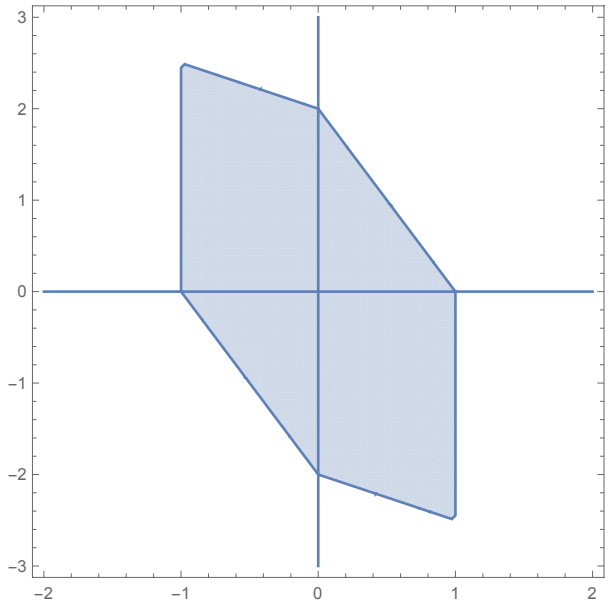



Figure 4.1: The set $M$ with $a = 1$ and $C = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$.

where $\underline{b} = (0, b)'$ is chosen such that $(C\underline{a})_1 = (C\underline{b})_1$. The rest of the set is defined by mirroring the parts of the above defined sets around both axes[8]:

$$M^{-\iota} = -M^{\iota}$$

and

$$M^{(1,-1)} = -M^{(-1,1)}.$$

Figure 4.1 shows the shape of the set M for $a = 1$ and $C = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$. Note that, even though we are only interested in a confidence set that provides a bound for one of the components, the need to comply with Condition A forces us to bound the set in the other component as well whenever $c_{12} \neq 0$. The interpretation of this fact is the following: As the Lasso can be viewed as a shifted Least-squares estimator whereby the size and direction of the shift depend (among others) on both components of the Least-squares estimator, it needs to be ensured that the influence of the parameter vector's second component on the shift is also taken into account by the procedure.

We will now proceed by showing that the set $M$ indeed complies with Condition A. Towards that end we start with a technical lemma that gives some inequalities related to the ones which appear in the condition.

**Lemma 41.** *Consider $v = (v_1, v_2)' \in \mathbb{R}^2$ and $w = (w_1, w_2)' \in \mathbb{R}^2$, suppose that $c_{12} \geq 0$ and that $C$ is symmetric and positive definite. Then the following holds.*

[8]Note that an asymmetric interval $[-\tilde{a}, a]$ could easily defined at this point by constructing $M^{(1,1)}(\tilde{a})$ as well as $M^{(1,-1)}(\tilde{a})$ and mirroring these sets around both axes to obtain the lower endpoint of the interval.

*(a) If $v_1 \leq w_1$ or $v_2 \geq w_2$, then*

$$(Cv)_2 \leq (Cw)_2 \implies (Cv)_1 \leq (Cw)_1.$$

*(b) If $(Cv)_1 \geq (Cw)_1$ and $(Cv)_2 \leq (Cw)_2$, then*

$$v_1 \geq w_1 \text{ and } v_2 \leq w_2.$$

*Proof.* To see part (a), assume that the implication does not hold:

$$(Cv)_2 = c_{12}v_1 + c_{22}v_2 \leq c_{12}w_1 + c_{22}w_2 = (Cw)_2$$

and

$$(Cv)_1 = c_{11}v_1 + c_{12}w_2 > c_{11}w_1 + c_{12}v_2 = (Cw)_1.$$

Taken together the above inequations yield

$$\begin{cases} c_{12}(v_1 - w_1) \leq c_{22}(w_2 - v_2) \\ c_{11}(v_1 - w_1) > c_{12}(w_2 - v_2). \end{cases} \tag{4.1}$$

Now note that (4.1) is equivalent to

$$\begin{cases} (v_1 - w_1) \leq \frac{c_{22}}{c_{12}}(w_2 - v_2) \\ (v_1 - w_1) > \frac{c_{12}}{c_{11}}(w_2 - v_2) \end{cases}$$

which implies that

$$\frac{c_{11}c_{22}}{c_{12}^2}(w_2 - v_2) > (w_2 - v_2), \tag{4.2}$$

a contradiction in case $v_2 \geq w_2$, since $C$ is positive definite and hence, $\frac{c_{11}c_{22}}{c_{12}^2} > 1$. Similarly, (4.1) implies that

$$\frac{c_{11}c_{22}}{c_{12}^2}(v_1 - w_1) > (v_1 - w_1), \tag{4.3}$$

a contradiction in case $v_1 \leq w_1$.

For part (b) note that the two in-equations in the assumption also yields inequations in (4.1). Hence, also (4.2) applies which yields a contradiction in case $v_2 > w_2$. Similarly, (4.3) needs to hold as well and yields a contradiction for the case $w_1 > v_1$. □

**Proposition 42.** *The set $M$, as defined in this subsection, satisfies Condition A.*

*Proof.* We need to show that for each $d \in \{-1, 1\}^d$ and each $m \in M$,

$$A_C^d(m) \subseteq M.$$

We first check this condition for $A_C^{-\iota}$:

- For $m \in M^\iota$, we have that

$$\begin{aligned} A_C^{-\iota}(m) &= \bigcap_{j\in\{1,2\}} \{z \in \mathcal{O}^\iota : (Cz)_j \leq (Cm)_j\} \\ &\subseteq \{z \in \mathcal{O}^\iota : (Cz)_1 \leq (Cm)_1\} \\ &\subseteq \{z \in \mathcal{O}^\iota : (Cz)_1 \leq (C\underline{a})_1\} \subseteq M, \end{aligned}$$

  because $(Cm)_1 \leq (C\underline{a})_1$, which follows from $m \in M^\iota$.

- For $m_2 \leq 0$, which implies that $m \in M^{(1,-1)} \cup M^{-\iota}$, we have

$$\begin{cases} (Cm)_1 = c_{11}m_1 + c_{12}m_2 \leq c_{11}a = (C\underline{a})_1 \\ (Cm)_2 = c_{12}m_1 + c_{22}m_2 \leq c_{12}a = (C\underline{a})_2, \end{cases}$$

  since $c_{12} \geq 0$ and $m_1 \leq a$. This implies that

$$A_C^{-\iota}(m) \subseteq A_C^{-\iota}(\underline{a}) \subseteq M^\iota \subseteq M.$$

- For $m \in M^{(-1,1)}$ first note that by Lemma 41, part (a) (with exchanged indices) we have that

$$(Cz)_1 \leq (Cm)_1 \implies (Cz)_2 \leq (Cm)_2$$

  for each $z \in A^{-\iota}(m)$, since $m \in M^{(-1,1)}$ implies that $z_1 \geq 0 \geq m_1$. Now define $\tilde{b}$ such that $(Cm)_1 = (C\underline{\tilde{b}})_1$ where $\underline{\tilde{b}} = (0, \tilde{b})'$. Note that by the above inequality and because $m \in M^{(-1,1)}$ we now have that $(C\underline{\tilde{b}})_2 \leq (Cm)_2 \leq (C\underline{b})_2$ which entails that $\tilde{b} \leq b$. Since $c_{12} \geq 0$ it follows that $(Cm)_1 = (C\underline{\tilde{b}})_1 \leq (C\underline{b})_1 = (Ca)_1$ so that

$$A^{-\iota}(m) \subseteq \{z \in \mathcal{O}^\iota : (Cz)_1 \leq (Cm)_1\} \subseteq \{z \in \mathcal{O}^\iota : (Cz)_1 \leq (C\underline{a})_1\} = M^\iota.$$

Next, we verify the condition for $A_C^{(1,-1)}$:

- For $m \in M^{(-1,1)}$ we have that

$$\begin{aligned} A_C^{(1,-1)}(m) &= \{z \in \mathcal{O}^{(-1,1)} : -(Cz)_1 \leq -(Cm)_1, (Cz)_2 \leq (Cm)_2\} \\ &= \{z \in \mathcal{O}^{(-1,1)} : z_1 \geq m_1, -(Cz)_1 \leq -(Cm)_1, (Cz)_2 \leq (Cm)_2\} \\ &\subseteq \{z \in \mathcal{O}^{(-1,1)} : z_1 \geq -a, (Cz)_2 \leq (C\underline{b})_2\} \\ &= M^{(-1,1)} \subseteq M, \end{aligned}$$

where the second equality holds by an argument similar to the proof of Lemma 41.

- For $m \in M \setminus M^{(-1,1)}$, implying that $m_1 > 0$ or $m_2 < 0$, we use the second part of Lemma 41:

$$(Cz)_1 \geq (Cm)_1 \text{ and } (Cz)_2 \leq (Cm)_2 \implies z_1 \geq m_1 \text{ and } z_2 \leq m_2,$$

which yields that $A_C^{(1,-1)}(m) = \varnothing$ whenever $m_1 > 0$ or $m_2 < 0$, since $A_C^{(1,-1)}(m) \subseteq \mathcal{O}^{(-1,1)}$.

For $A_C^{\iota}(m)$ and $A_C^{(-1,1)}(m)$ note that

$$\begin{aligned} -A_C^d(m) &= -\{z \in \mathbb{R}^p : d_j(Cz)_j \leq d_j(Cm)_j, d_j z_j \geq 0 \ \ \forall j\} \\ &= \{z \in \mathbb{R}^p : -d_j(Cz)_j \leq -d_j(C(-m))_j, -d_j z_j \geq 0 \ \ \forall j\} \\ &= A_C^{-d}(-m). \end{aligned}$$

and that we thus have by the previous results

$$A_C^{\iota}(m) = -A_C^{-\iota}(-m) \subseteq -M = M$$

for all $m \in M$ and similarly

$$A_C^{(-1,1)}(m) = -A_C^{(1,-1)}(-m) \subseteq -M = M$$

for all $m \in M$. □

Having established that the set $M$ complies with Condition A, we will now show that this is indeed the largest set that does so while its projection on the subspace associated with the first component does not exceed $a$ in absolute value.

**Proposition 43.** *If $\tilde{M} \subseteq \mathbb{R}^2$ satisfies Condition A and $\sup_{m \in \tilde{M}} |m_1| \leq a$, then*

$$\tilde{M} \subseteq M.$$

*Proof.* We assume that $\tilde{M} \ni z \notin M$ and show that $\max_{m \in \tilde{M}} |m_1| > a$, if $\tilde{M}$ satisfies Condition A:

- $z \in \mathcal{O}^{\iota}$: Suppose that $z \notin M^{\iota}$. Lemma 41 then entails[9], that

$$(Cz)_j > (C\underline{a})_j \ \ \forall j \in \{1, 2\}.$$

Now choose[10] $\tilde{\underline{a}} = (\tilde{a}, 0)$ with $\tilde{a} \geq 0$ such that

[9] Indeed, in case $(Cz)_2 \leq (C\underline{a})_2$ Lemma 41 yields that also $(Cz)_1 \leq (C\underline{a})_1$, since $z_2 \geq 0$. But then $z \in M$.

[10] Note that such an $\tilde{a} \geq 0$ exists, since $(C(z_1, 0)')_j \leq (Cz)_j$ for each $j \in \{1, 2\}$ and hence the set $\{z_1 \in \mathbb{R}_+ : (C(z_1, 0)')_j \leq (Cz)_j \ \ \forall j \in \{1, 2\}\}$ is non-empty.

$$(C\underline{\tilde{a}})_j \leq (Cz)_j \quad \forall j \in \{1,2\}$$

and such that

$$(C\underline{\tilde{a}})_{j_0} = (Cx)_{j_0}$$

for some $j_0 \in \{1,2\}$.

Clearly, we have $\underline{\tilde{a}} \in A_C^{-\iota}(z)$ and $(C\underline{\tilde{a}})_{j_0} = (Cz)_{j_0} > (C\underline{a})_{j_0}$ implying that

$$\tilde{a} > a,$$

so that $\sup_{m \in \tilde{M}} m_1 > a$.

- $z \in \mathcal{O}^{(-1,1)}$: Suppose $z \notin M$, i.e., $(Cz)_2 > (C\underline{b})_2$.

  Now choose $\tilde{b} \geq 0$ such that $(C\underline{\tilde{b}})_2 = (Cz)_2$, where $\underline{\tilde{b}} = (0, \tilde{b})'$. We have that $\underline{\tilde{b}} \in A_C^{(1,-1)}(z)$ :

  Suppose not, i.e., $(C\underline{\tilde{b}})_1 < (Cz)_1$, then, since $\tilde{b} = \frac{c_{12}}{c_{22}} z_1 + z_2$,

$$c_{12}\tilde{b} < c_{11}z_1 + c_{12}z_2 \iff c_{12}(\tfrac{c_{12}}{c_{22}}z_1 + z_2) < c_{11}z_1 + c_{12}z_2 \iff \tfrac{c_{12}^2}{c_{11}c_{22}} z_1 < z_1,$$

  a contradiction whenever $z_1 > 0$, since $\frac{c_{12}^2}{c_{11}c_{22}} < 1$.

  Finally, we have $\tilde{b} > b$, as $c_{22}\tilde{b} = (C\underline{\tilde{b}})_2 = (Cz)_2 > (C\underline{b})_2 = c_{22}b$ and thus also $(C\underline{\tilde{b}})_1 > (C\underline{b})_1$. Noting that $\underline{\tilde{b}} \in \mathcal{O}^{\iota}$ leads us back to the previous case.

- For $z \in \mathcal{O}^{-\iota}$, we have that $z \notin M \iff -z \notin -M$. Noting that, in this case, $-z \in \mathcal{O}^{\iota}$ and that $-M = M$ reduces this case to the first case. Finally, using the same argument, the case $z \in \mathcal{O}^{(1,-1)}$, can be reduced to the case $z \in \mathcal{O}^{(-1,1)}$.

□

**Remark 44.** *It is again easily seen that the interval's half-length, a, must be greater than the half-length of the (canonical) confidence interval that is based on the Least-squares estimator, i.e., the* $(1 - \frac{\alpha}{2})$*-quantile of the standard normal distribution, since, as described in Chapter 2,* $2^p = 4$ *Gaussian random variables*[11]*,* $\hat{u}^d$*, need to be covered by the (shifted) confidence set.*

Given the preceding remark, one might now be interested in the size difference between the confidence intervals that are constructed based on the Lasso and Least-squares estimates, respectively. Pötscher & Schneider (2010) have already shown that in the orthogonal regressor case, the length

[11] Having the same variance-covariance structure as the Least-squares estimator, but different means.

| $\lvert c_{12}\rvert$ | 0.25 | 0.5 | 0.75 | 0.9 |
|---|---|---|---|---|
| $\lambda_1 = 0$ | 1.96 | 1.96 | 1.96 | 1.96 |
| $\lambda_1 = 0.1$ | 2.1 | 2.4 | 3.1 | 4.9 |
| $\lambda_1 = 0.5$ | 2.4 | 2.9 | 4.5 | 8.8 |
| $\lambda_1 = 1$ | 3.0 | 3.9 | 6.5 | 13.8 |
| $\lambda_1 = 2$ | 4.4 | 5.9 | 10.5 | 23.8 |
| $\lambda_1 = 3$ | 5.7 | 7.9 | 14.5 | 33.8 |

Table 4.1: Half-lengths of the 95% confidence intervals based on the fully penalized Lasso estimator for $c_{11} = c_{22} = 1$, $\sigma^2 = 1$ and different penalization parameters $\lambda_1$. (Values rouned to one digit.)

of confidence intervals which are based on the Lasso is greater than the length of the canonical Least-squares interval as well as that the length of the intervals is increasing in the penalization parameter[12], $\lambda = (\lambda_1, \lambda_1)'$. Table 4.1 contains the required values of $a$, i.e., the half-length of the interval, of the Lasso confidence interval for $c_{11} = c_{22} = 1$ and $\sigma^2 = 1$ and various combinations of $\lambda_1$ and $c_{12}$. Note that in this case the half-length of the Least-squares interval is 1.96 and that the Least-squares estimator is equal to the Lasso when $\lambda_1 = 0$.

The picture that is drawn here is the following: For small values of $\lambda_1$ and $c_{12}$ the resulting confidence interval is only slightly longer than the one based on the Least-squares estimator. For increasing $\lambda_1$ and $|c_{12}|$ the required length of the interval increases significantly, in particular in the latter case, with the length more than doubling as $c_{12}$ increases from 0.25 to 0.9 for each of the presented values of $\lambda_1 > 0$. This effect is even more extreme for larger values of $\lambda_1$. Two effects are at work here: On the one hand, the volume of $M$ decreases for fixed $a > 0$ as $c_{12}$ increases. On the other hand, the corners of the distorted $\lambda$-box, $C^{-1}\Lambda d$ ($d \in \{-1, 1\}^p$) that are the means of the normal distributions whose probability mass must be covered shift further apart as $c_{12}$ increases in absolute value. Obviously, increasing the components of the tuning vector $\lambda$ shifts the distribution's means further away from the origin, which results in larger confidence sets.

In terms of practical implications, this shows that the "cost" of, in this case Lasso-based, variable selection in terms of estimation accuracy can be quite severe. Given the construction of the confidence set in this case, it is not hard to imagine that a version of such a confidence set will carry an even higher penalty in a setting where more than two components are present and multiple correlations have to be taken into account. Also note that this property is not unique to the Lasso, as can be seen in Taylor & Tibshirani (2015), for example.

### 4.3.2 Partial penalization case

We will now consider the scenario where only some parameters are penalized in the estimation, i.e., $\lambda_j = 0$ for some $j$'s. We refer to this type of Lasso estimator as a *partial Lasso*. Using such

[12] In this case we choose the weights to be equal and thus the penalization vector $\lambda$ is specified by a single parameter.

a choice of penalization vector seems quite attractive in some cases. Indeed, if one is primarily interested in only a few of the parameter vector's components, this choice of weights will result in an estimator that can still yield relatively small models, since all other parameter estimates are still being shrunk towards, or exactly to, zero while promising smaller confidence intervals for the sub-parameter in question. It appears reasonable to assume that the penalization's adverse effects on the size of the confidence regions that have been outlined so far will be somewhat mitigated if the parameter of primary interest is not penalized. To verify that this is actually the case and to quantify the effect's magnitude we do the following. Let

$$\mathcal{R} = \{j \in \{1, \dots, p\} : \lambda_j > 0\}$$

denote the index set containing all penalized components and let

$$\mathcal{N} = \{j \in \{1, \dots, p\} : \lambda_j = 0\}$$

denote the remaining components, which are not subject to any penalization.

Note that the theory developed in the previous parts[13] also comprises the partial Lasso case and valid confidence sets could be constructed using Theorem 4. It is, however, possible to produce an improved version of the theorem, as Condition A can be slightly relaxed when considering the partial Lasso. This alternative condition is given in the following.

**Condition B.** *Let $\bar{C} \in \mathbb{R}^{p \times p}$ be positive definite. A set $M \subseteq \mathbb{R}^p$ satisfies Condition B with matrix $\bar{C}$ if*

$$A^d_{\mathcal{R},\bar{C}}(m) = \bigcap_{j=1}^{p} \Big( \{z \in \mathbb{R}^p : (\bar{C}z)_j = (\bar{C}m)_j, d_j z_j \leq 0 \quad \forall j \in \mathcal{N}\} \\ \cap \{z \in \mathbb{R}^p : d_j(\bar{C}z)_j \geq d_j(\bar{C}m)_j, d_j z_j \leq 0 \quad \forall j \in \mathcal{R}\} \Big) \subseteq M$$

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

**Remark 45.** *Condition B is less restrictive than Condition A. Indeed, we have that $A^d_{\mathcal{R},\bar{C}}(m) \subseteq A^d_{\bar{C}}(m)$ for each $m \in \mathbb{R}^p$ and each matrix $\bar{C}$ and hence, a set which satisfies Condition A also satisfies Condition B.*

**Proposition 46.** *If $M \subseteq \mathbb{R}^p$ satisfies Condition B with $\bar{C} = C$, then*

$$\inf_{\beta \in \mathbb{R}^p} P_\beta \left(\hat{u} \in M\right) = \min_{d \in \{-1,1\}^p} P\left(\hat{u}^d \in M\right),$$

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

[13] In particular in Chapter 2 and Section 4.3.1.

**Remark 47.** *Note that, since $\lambda_j = 0$ for each $j \in \mathcal{N}$, the random variables $\hat{u}^d$ in Proposition 46 only have $2^{|\mathcal{R}|}$ distinct distributions, as their means, which are given by $C^{-1}\Lambda d$, are identical when flipping the sign of $d_j$ for any $j \in \mathcal{N}$.*

*Proof of Proposition 46.* The proof is essentially the same as the proof of Theorem 4 including the discussion leading up to that result. The only difference is in the proof of the equivalent of Proposition 2:

First note that in the partial Lasso case the function $V(u)$ can be written as

$$V(u) = u'Cu + 2u'W + 2\sum_{j\in\mathcal{R}} \lambda_j \left(|u_j + \beta_j| - |\beta_j|\right)$$

and also the functions $V^d$ now reduce to

$$V^d(u) = u'Cu - 2u'W + 2\sum_{j\in\mathcal{R}} \lambda_j u_j d_j.$$

Instead of Fact (a) in the proof of Proposition 2, the same arguments now yield that

$$(C\hat{u})_j = (C\hat{u}^\iota)_j \quad \forall j \in \mathcal{N}$$

and as before

$$(C\hat{u})_j \leq (C\hat{u}^\iota)_j \quad \forall j \in \mathcal{R}.$$

From this we see that indeed Condition A may be replaced by Condition B in this setting, while the rest of the proof is completely analogous. □

We will now determine the optimal shape of the confidence set based on the partial Lasso in case of a model that contains two regressors.

**Constructing the optimal shape in case p=2:**

As in the previous subsection we will again consider the case of $p = 2$ where only the second component is penalized and we are primarily interested in the first component, which is not penalized in the estimation. We can now use Condition B to construct a confidence interval for that component which is slightly shorter than the one constructed in the previous subsection. Similarly to the previous case, we now use Proposition 46 to find a set $M_\mathcal{R} = M_\mathcal{R}(a)$ which satisfies

- $M_\mathcal{R} \subseteq \mathbb{R}^p$ complies with Condition B.
- $\max_{m\in M_\mathcal{R}} |m_1| \leq a$ for some $a > 0$.
- $\inf_{\beta\in\mathbb{R}^p} P_\beta(\hat{u} \in M_\mathcal{R}) \geq 1 - \alpha$ for some $0 < \alpha < 1$.

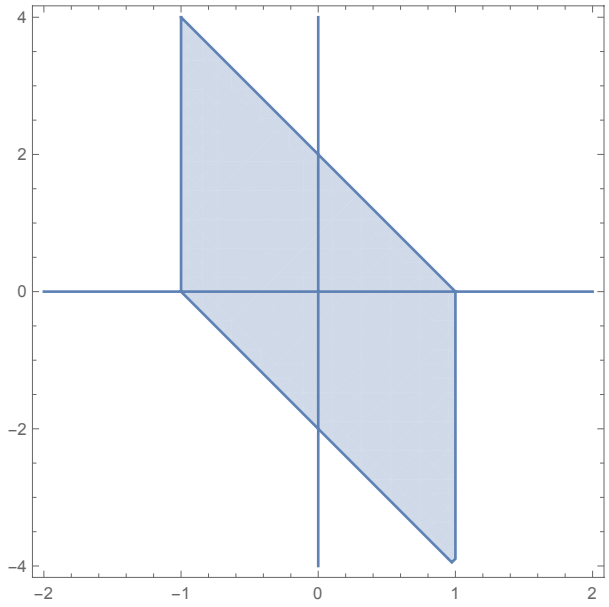



Figure 4.2: The set $M_{\mathcal{R}}$ with $a = 1$, $\mathcal{R} = \{2\}$ and $C = \left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$.

We will again assume that $C$ is positive definite and that, without loss of generality[14], $c_{12} \geq 0$. As in the fully penalized Lasso case it is easily seen that for a diagonal matrix $C$, the set

$$\overline{M} = \{z \in \mathbb{R}^2 : |z_1| \leq a\}$$

is the largest set that complies with Condition B and cannot be enlarged while maintaining a fixed projection onto the subspace associated with the first component. Also, in this case, it is easily seen that the confidence interval constructed in this way is again equivalent to the procedure proposed by Pötscher & Schneider (2010), when considering a confidence set for both components. However, note that in this case, by orthogonality of the regressors, the marginal distribution of the component of primary interest is simply equal to the Least-squares estimator's marginal distribution, thus yielding the same (marginal) confidence interval, since the component of interest is not being penalized.

Turning to the case of $c_{12} > 0$ we define the set

$$M_{\mathcal{R}}^{+} = \{z \in \mathbb{R}^2 : |z_1| \leq a, z_2 \geq 0, (Cz)_1 \leq (C\underline{a})_1\}$$

where $\underline{a} = (a, 0)'$ and $a \geq 0$. Now determine the shape of the confidence set by mirroring the set $M_{\mathcal{R}}^{+}$ around both axes:

$$M_{\mathcal{R}} = -M_{\mathcal{R}}^{+} \cup M_{\mathcal{R}}^{+}$$

[14] As before, simply consider the model $y = \tilde{x}_{\cdot 1}\tilde{\beta}_1 + x_{\cdot 2}\beta_2 + \varepsilon$ with $\tilde{x}_{\cdot 1} = -x_{\cdot 1}$, in case $c_{12} < 0$.

An example of the set $M_{\mathcal{R}}$ is displayed in Figure 4.2. Note that $M_{\mathcal{R}} = -M_{\mathcal{R}}$ and that $M(a) \subseteq M_{\mathcal{R}}(a)$ for fixed $a \geq 0$, as the set now contains additional points in $\mathcal{O}^{(-1,1)}$ and $\mathcal{O}^{(1,-1)}$. (For a visualization of this compare Figure 4.2 to Figure 4.1 which displays the set for the full penalization case with otherwise identical parameters.)

**Proposition 48.** *The set $M_{\mathcal{R}}$ satisfies Condition B.*

*Proof.* By symmetry it is again sufficient to verify that $A^{d}_{\mathcal{R},C}(m) \subseteq M_{\mathcal{R}}$ for $d \in \{-\iota, (1,-1)'\}$, where $m \in M_{\mathcal{R}}$ is arbitrary. By Lemma 41 we have that for $m \in \mathbb{R}^p$ and $z \in \mathbb{R}^p$

$$(Cz)_1 \geq (Cm)_1 \text{ and } (Cz)_2 \leq (Cm)_2 \implies z_1 \geq m_1 \text{ and } z_2 \leq m_2.$$

Using this fact we see that for any $m \in M_{\mathcal{R}}$

$$\begin{aligned} A^{-\iota}_{\mathcal{R},C}(m) \cup A^{(1,-1)}_{\mathcal{R},C}(m) &= \{z \in \mathbb{R}^p : z_2 \geq 0, (Cz)_1 = (Cm)_1, (Cz)_2 \leq (Cm)_2\} \\ &= \{z \in \mathbb{R}^p : z_2 \geq 0, z_1 \geq m_1, (Cz)_1 = (Cm)_1, (Cz)_2 \leq (Cm)_2\} \\ &\subseteq \{z \in \mathbb{R}^p : z_2 \geq 0, z_1 \geq -a, (Cz)_1 = (Cm)_1\} \subseteq M_{\mathcal{R}}, \end{aligned}$$

since $m \in M_{\mathcal{R}}$ implies that $(Cm)_1 \leq (C\underline{a})_1$. □

We now show that the set $M_{\mathcal{R}}$ is the largest set with a fixed projection onto the subspace associated with the first component that satisfies Condition B.

**Proposition 49.** *If $\tilde{M}_{\mathcal{R}} \subseteq \mathbb{R}^2$ satisfies Condition B and that $\sup_{m \in \tilde{M}_{\mathcal{R}}} |m_1| \leq a$, then*

$$\tilde{M}_{\mathcal{R}} \subseteq M_{\mathcal{R}}.$$

*Proof.* The proof is completely analogous to the first part of the proof of Proposition 43. □

We are again interested in the behavior of the length of the confidence interval in dependence of both the penalization parameter's second component, $\lambda_2$, and the regressors' correlation. We will again consider the case where $c_{11} = c_{22} = 1$ and $\sigma^2 = 1$. Note that in case of an orthogonal design, i.e., for $c_{12} = 0$, the interval is the same as the Least-squares interval ($a \approx 1.96$). It is again easy to see that also in the partial Lasso case the size of a confidence set that is based on that estimator will be larger than a Least-squares confidence set whenever $c_{12} \neq 0$. Hereby, one can again observe the same effects as in the previously considered full penalization case. First, note that as in the case of the fully penalized Lasso there are multiple normal distributions with the same variance-covariance structure, but different means which have to be covered. This effect will be a bit less severe in the partial Lasso case, as there will be fewer distinct means for the random variables that have to be covered, an effect that occurs if not all components are penalized (c.f. Remark 47). Also note that for fixed $a > 0$, $M(a) \subseteq M_{\mathcal{R}}(a)$ and that both sets decrease[15] in volume as $|c_{12}|$ increases.

[15] For fixed $a > 0$.

| $\lvert c_{12}\rvert$ | 0.25 | 0.5 | 0.75 | 0.9 |
|---|---|---|---|---|
| $\lambda_2 = 0$ | 1.96 | 1.96 | 1.96 | 1.96 |
| $\lambda_2 = 0.1$ | 2.1 | 2.3 | 3 | 4.6 |
| $\lambda_2 = 0.5$ | 2.1 | 2.4 | 3.4 | 6.2 |
| $\lambda_2 = 1$ | 2.2 | 2.6 | 4.2 | 8.5 |
| $\lambda_2 = 2$ | 2.4 | 3.3 | 5.9 | 13.3 |
| $\lambda_2 = 3$ | 2.6 | 3.9 | 7.6 | 18.0 |

Table 4.2: Half-lengths of the 95% confidence intervals based on the partial Lasso estimator for $c_{11} = c_{22} = 1$ and $\sigma^2 = 1$, rounded to one digit.

Table 4.2 displays the half-lengths of the resulting confidence sets for various values of $\lambda_2$ and $c_{12}$. While qualitatively similar to the full penalization case, the results differ from that case in the following way: For small absolute values of $c_{12}$ the increase in $\lambda_2$ is far less pronounced than before, whereas for large values of $c_{12}$ the confidence sets still have to be substantially larger than for small values of this parameter. Note that the resulting confidence sets are considerably smaller than those that are based on the fully penalized Lasso estimator (Table 4.1). This size difference is due to the fact that the bias induced by penalizing the component of interest is removed in the partial Lasso case. The bias induced by penalizing the second component, however, remains due to the regressors' correlation. This "carry-over" bias is made visible by the increased size of the corresponding confidence sets for higher-correlated regressors.

## 4.4 Adaptive choice of the sub-parameter being covered

In this section we discuss the validity of a procedure that produces confidence sets only for the non-zero parameter estimates. Since the components of the finally selected model are random, the choice of shape of the confidence set is consequently random[16] as well. It is thus not clear, a priori, whether such a procedure will result in valid confidence sets in general, as a certain model's selection event and the parameter's coverage event are dependent in an intricate manner. However, a procedure of this type will be quite attractive to users, since they are, often, only interested in the components of the active set. Note, however, that we still consider confidence sets designed to cover the entire parameter vector[17], albeit with different priorities[18] on certain components, depending on which model is chosen by the procedure. This is in contrast to the post-selection inference procedures proposed by Berk et al. (2013) and Lee et al. (2016) as well as similar works

[16]Note, however, that the shape and size are fixed for any sub-model that may be selected. This means that the randomness of the confidence set's shape merely stems from the fact that different parts of the parameter should be covered in an "optimal" way.

[17]This, as we have seen in the previous sections, also yields a bound for all components of the parameter vector, whenever the design matrix' columns are not orthogonal.

[18]These priorities will be reflected in the choice of shape for the confidence sets, depending on the selected model.

(also c.f. Section 2.2). The main goal of this section is to determine whether such a procedure, if implemented in a rather simple but quite intuitive fashion, will produce valid confidence sets in the sense that the unconditional probability of the randomly chosen parameter being covered is at least equal to the nominal coverage probability.

To formalize the procedure, we do the following. For an active set $\mathcal{A} = \mathcal{A}(\hat{\beta}_{\mathrm{L}}) = \{j \in \{1, \dots, p\} : \hat{\beta}_{\mathrm{L},j} \neq 0\}$, we want to construct a confidence set $M_{\mathcal{A}} \subseteq \mathbb{R}^{|\mathcal{A}|}$ for the sub-parameter $\beta_{\mathcal{A}}$. Given this definition we will have to consider the special case in which the Lasso estimator yields the empty model, i.e., $\hat{\beta}_{\mathrm{L}} = 0$. Note that this model is assumed with positive probability whenever $\lambda_j > 0$ for all $j = 1, \dots, p$, as seen in Corollary 22. By the above definition we would thus want to construct a confidence set for an "empty" parameter, $\beta_{\varnothing}$, which would simply mean not to conduct any inference at all. Despite this being a valid procedure in principle, it seems reasonable that in such a case the user would be more interested in the question of how close to the origin the entire parameter will actually be. We thus define this "empty" parameter to be equal to the whole parameter vector: $\beta_{\varnothing} = \beta$.

To describe the approach, we denote the powerset by $\mathscr{P}(\cdot)$. For each possible sub-model $\mathcal{I} \in \mathscr{P}(\{1, \dots, p\})$ consider a set $M_{\mathcal{I}}$ that is designed to uniformly cover the estimation error and is fixed with respect to its size and shape. We now estimate the unknown parameter using the Lasso and from the above family of sets, $\{M_{\mathcal{I}}\}_{\mathcal{I} \in \mathscr{P}(\{1,\dots,p\})}$, pick the one that corresponds to the active set, $M_{\mathcal{A}}$, and center it at the Lasso estimate. In order to define the overall coverage probability of this procedure, we first consider the conditional[19] coverage probability and multiply it by the probability of selecting the corresponding model. This is equivalent to the joint probability of selecting a certain model and covering the corresponding sub-parameter. To cover every possible model, we now take the sum over all these cases:

$$P_\beta(\beta_{\mathcal{A}} \in \hat{\beta}_{\mathrm{L}} - M_{\mathcal{A}}) = \sum_{\mathcal{I} \in \mathscr{P}(\{1,\dots,p\})} P_\beta(\beta_{\mathcal{I}} \in \hat{\beta}_{\mathrm{L}} - M_{\mathcal{I}} \text{ and } \mathcal{A}(\hat{\beta}_{\mathrm{L}}) = \mathcal{I}). \tag{4.4}$$

In lack of a general procedure on how to construct confidence sets for sub-parameters that are based on the Lasso estimator, we limit ourselves to the case where $p = 2$ and use the confidence sets constructed in Chapter 2 in case $\beta_{\mathcal{A}} = \beta$ and Section 4.3.1 in case $\mathcal{A} \in \{\{1\}, \{2\}\}$.

In this setting a simulation study was carried out to determine the respective overall coverage probability. The simulations were set up in the following way: For a given regressor matrix $X$ which satisfies that $\frac{1}{n}X'X = \frac{1}{n}C = \begin{pmatrix} 1 & c_{12}^* \\ c_{12}^* & 1 \end{pmatrix}$, the response $y$ is generated from the model in (1.1). Next, the Lasso estimator is calculated for $\lambda = (\sqrt{n}, \sqrt{n})'$ and the selected model is recorded, as well as whether the corresponding parameter is covered by its 95% confidence set. Repeating this procedure 100,000 times and calculating the means of both the model selection indicators and the coverage-indicators, we obtain estimates for model the selection probabilities (c.f. Figure 4.4) and,

[19] For a given active set.

more importantly, the coverage probabilities as defined in (4.4). Note that the simulation's margin of error is upper-bounded by 0.0032[20]. This was repeated for different values of $\beta$ and $c_{12}^*$. The simulation was carried out using the statistical software package R with the (main) simulation code being provided in Appendix B.

The results for the coverage probability, which are displayed in Figure 4.3, strongly suggest that the procedure under consideration is valid: For $c_{12}^* = 0.5$, the lowest coverages have been observed for large absolute values of $\beta_1$ and $\beta_2$ and where $\operatorname{sgn}(\beta_1) = \operatorname{sgn}(\beta_2)$. Note that in these cases[21] the Lasso almost always selects a model without any zero-components, thus effectively not performing model selection, as can be seen in Figure 4.4. Hence, this case essentially corresponds to the case of having to cover $\hat{u}^{\iota}$ and $\hat{u}^{-\iota}$, respectively as described in Chapter 2. By construction, the coverage probability is almost exactly $0.95 = 1-\alpha$ in case $\operatorname{sgn}(\beta_1) = \operatorname{sgn}(\beta_2)$ and approximately 0.992 in the case where $\operatorname{sgn}(\beta_1) \neq \operatorname{sgn}(\beta_2)$ which corresponds to $P(\hat{u}^{(-1,1)} \in M)$. This is due to the fact that the minimal coverage probability is attained for large absolute values of the true parameter, as pointed out in Chapter 2. Indeed, the confidence sets produced in Chapter 2 and Section 4.3.1 are over-covering the true parameter in case it is small while attaining their nominal coverage probabilities for large parameters (that have certain signs). This is due to the fact that the Lasso confidence sets have to correct for the bias that is induced by shrinking the parameter estimates. If the true parameters actually are almost zero, this bias is over-corrected leading to a higher-than-nominal coverage probability for the corresponding values of $\beta$. As a result of this, we see that in cases where $\beta$ is small[22] and the empty model is frequently selected by the estimator, the true parameter is covered by the confidence set in over 99% of cases. A similar effect appears to hold in the setting where a model with only one non-zero component is selected often: If the non-zero component is small in absolute value, then the coverage probability is very high, i.e., almost exactly one. Moving away from that area, the true coverage slightly declines, but still ranges between 0.98 and 0.99 depending on the value of the component of the true parameter that is estimated to be exactly zero.

This behavior may be attributed to the fact that, while the confidence sets are designed to cover the entire parameter vector (albeit with a shape that is chosen to be optimized for a specific sub-parameter), we only review whether the component of interest is covered in the simulations. This indicates that there may still be some room to improve the length of the component-wise confidence interval. However, this may only be achieved when the confidence intervals for single components are not based on Proposition 46, since the shape of the confidence set based on that result is already optimal in the sense of Proposition 49.

[20] Note that the simulation's outcome follows a binomial distribution. Bounding its variance by $\frac{0.5^2}{N}$, where $N$ denotes the number of repetitions, and using an approximation by the normal distribution, one may add and subtract two standard deviations to obtain the end-points of a 95.4% confidence interval.

[21] I.e., in cases all components of $\beta$ are large in absolute value.

[22] In the Euclidean norm, for example.

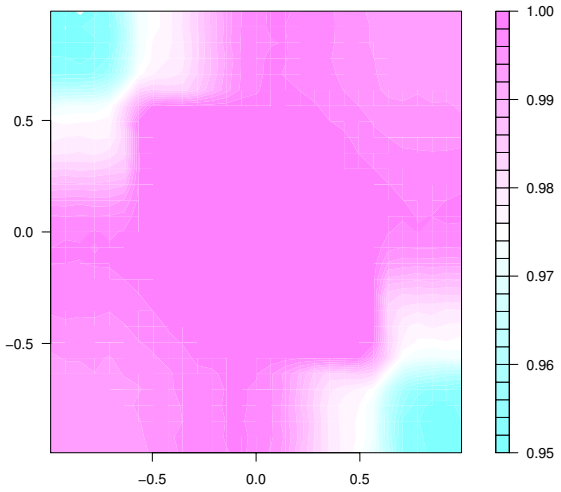



Figure 4.3: The coverage probabilities according to simulations in dependence of $\frac{1}{2\sqrt{n}}\beta_1$ (x-axis) and $\frac{1}{2\sqrt{n}}\beta_2$ (y-axis) for $c_{12}^* = 0.5$ and $\lambda = (\sqrt{n}, \sqrt{n})'$.

Turning to a different choice of parameter, $c_{12}^* = 0.25$, we see a qualitatively similar picture to the previous choice of parameter: The coverage is very high for small $\beta$'s and decreases as the components of the true parameter get larger. Despite being lower than for the previous value of $c_{12}^*$, the true coverage probability always lies above 0.95. This can be explained by a now less severe over-correction compared to the previous case, as there is now less correlation between the two regressors. The coverage and model selection probabilities for this case are visualized in Figure A.1 which can be found in Appendix A.

For $c_{12}^* = 0.75$, the picture is again similar. However, the area in which the parameter of interest is covered with probability almost equal to 1 now stretches over all parameters that do not have components with differing signs and a large norm. This is again to be attributed to the larger correlation in this case, which in turn leads to stronger over-correction for non-coverage-minimizing parameters. The results for this correlation parameter are visualized in Figure A.3 which can again be found in Appendix A.

Finally, to ensure that the results are not the result of the specific choice of design matrix $X$, another set of simulations was run, this time generating a new regressor-matrix $X$ for each repetition. The results confirm the above findings, but are not reported, as they hardly give any new insights.

Summing things up, the results of the simulation study indicate that the proposed procedure which includes a different choice of sub-parameter to be covered for each selected model is valid in the sense that the resulting overall coverage probability for the quantity of interest never, i.e., for

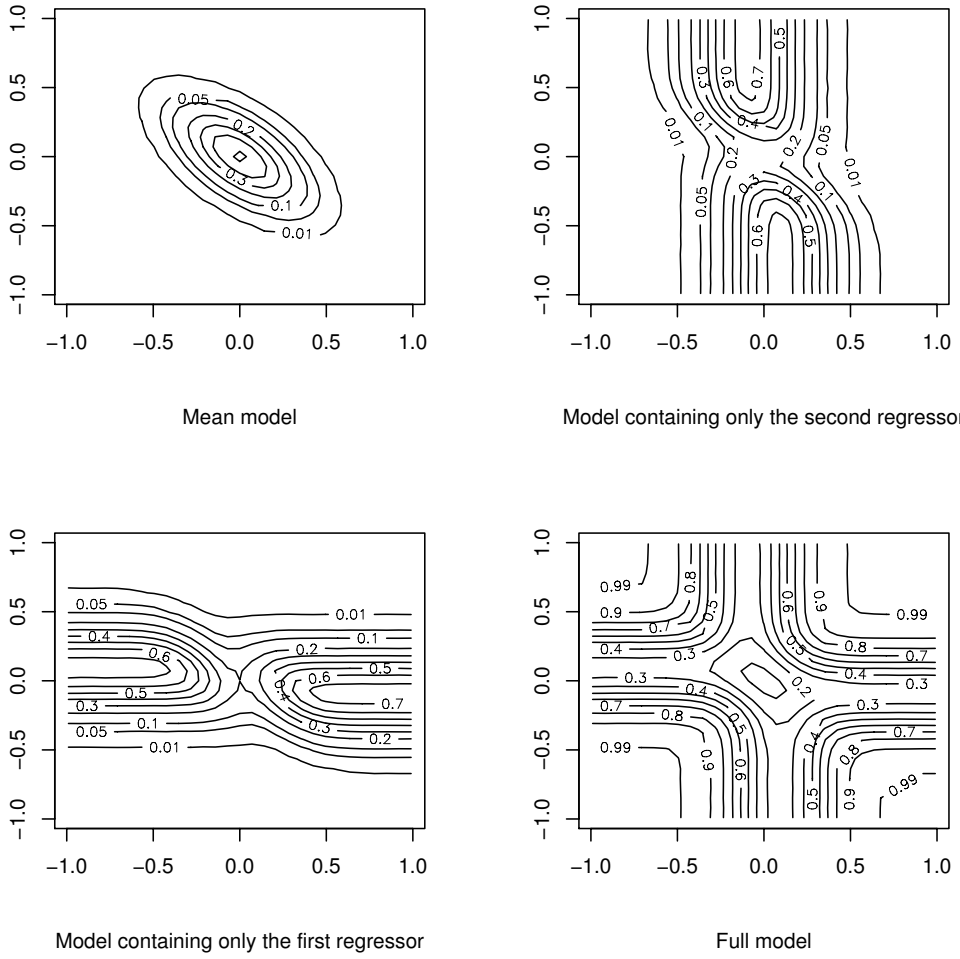



Figure 4.4: The model selection probabilities according to the simulations in dependence of $\frac{1}{2\sqrt{n}}\beta_1$ (x-axis) and $\frac{1}{2\sqrt{n}}\beta_2$ (y-axis) for $c_{12}^* = 0.5$ and $\lambda = (\sqrt{n}, \sqrt{n})'$.

no value of $\beta$, falls below the confidence set's nominal level $1-\alpha$. At first this may seem to be in contrast to findings of Kabaila & Leeb (2006) who show that, in a classical model selection setting, the true coverage probability of "naive" confidence sets for the true (sub)parameter often falls well below the nominal coverage probability. This, however, is not a contradiction at all, since the confidence sets that are constructed in this simulation are not naive by any means, but designed to account for the Lasso's model selection properties and cover the true parameter uniformly. Indeed, it seems to rather over-cover the desired parameter in cases of a single-regressor model being chosen by the procedure. Naturally, there is room for further research, both in terms of giving an analytical confirmation of these results in higher dimensions (i.e., $p > 2$), as well as a more optimal choice of confidence set for certain sub-parameters of interest.

## 4.5 Adaptive confidence regions

As previously shown, the minimal coverage probability for the proposed confidence sets is assumed when the components of the true parameter vector are, or become, large in absolute value, i.e., as $|\beta_j| \to \infty$ for each $j$. Hereby the coverage-minimizing signs of $\beta$ depend on the (sequence of the) regressor matrix $X$. Knowing the signs of the true parameter vector, at least partially, would allow us to construct confidence sets that are smaller in volume than the sets proposed in Chapter 2 using Proposition 2. In this case it is sufficient to ensure that the minimal coverage probability over all $\beta$'s having certain signs is sufficiently large, thus reducing the number of random variables, $\hat{u}^d$, that have to be covered, if a confidence set is constructed in the way described in the previous parts of this thesis. Being aware of the fact that Lasso-based confidence sets can be reduced in size by either penalizing only some of the parameters, or by considering only some components, one may ask the question whether information about the parameter's true sign can also be used to obtain smaller confidence sets.

The true signs of $\beta$ are, however, rarely[23] known. It may thus may seem compelling to try and estimate, or test for, the signs of the parameter vector prior to choosing the shape for the confidence set. The reasoning behind this idea is that a "sufficiently conservative" test for the signs may, at least asymptotically, as $n \longrightarrow \infty$, detect the true sign, thus enabling the user to construct the confidence set as if the signs of the true parameter vector were known.

To explore this idea, we will thus consider an asymptotic framework and make the dependence of the parameters on the sample size explicit again, so that $X = X_n$ and $\lambda = \lambda_n$. And as before, we will suppress the dependence of the estimators $\hat{\beta}_{\mathrm{L}}$ and $\hat{\beta}_{\mathrm{LS}}$ on $n$ in the notation. We will consider the very simple case of the mean-model where $X_n = \iota$ for each $n \in \mathbb{N}$ (hence, $p = 1$). The model thus reads

$$y_i = \beta + \varepsilon_i \tag{4.5}$$

[23] Note that the signs may be known due to some external information about the model. One may argue that this should rather be reflected in one's choice of estimator, which would lead to a different procedure.

where $1 \leq i \leq n$, $\varepsilon_i \sim_{iid} N(0,1)$ and $\beta \in \mathbb{R}$ is the parameter of interest. Note that the Lasso can be stated explicitly in this simple model:

$$\hat{\beta}_{\mathrm{L}} = (\hat{\beta}_{\mathrm{LS}} - \eta_n)\mathbb{1}_{\{\hat{\beta}_{\mathrm{LS}} > \eta_n\}} + (\hat{\beta}_{\mathrm{LS}} + \eta_n)\mathbb{1}_{\{\hat{\beta}_{\mathrm{LS}} < -\eta_n\}}, \tag{4.6}$$

where $\eta_n = \frac{1}{n}\lambda_n$. Suppose that $\sqrt{n}\eta_n \longrightarrow \eta < \infty$ as $n \longrightarrow \infty$, thus putting us into a conservative model selection framework. We now want to define the sign estimator that will be used to determine the sign of the parameter of interest in a conservative manner. It appears reasonable to base the sign estimator on the Least-squares estimator, as this estimator is unbiased and consistent. We thus define the sign estimator as

$$\hat{s}_n = \mathbb{1}_{\{\hat{\beta}_{\mathrm{LS}} > n^{-\gamma}\}} - \mathbb{1}_{\{\hat{\beta}_{\mathrm{LS}} < -n^{-\gamma}\}},$$

where $0 < \gamma < \frac{1}{2}$ is a tuning parameter that determines the procedure's "conservativeness". The procedure is to be interpreted in the following way: if $\hat{s}_n = 0$ we cannot determine whether the true sign is positive, negative, or actually zero based on the given sample size. The decision whether $\hat{s} = 0$ can also be interpreted as a statistical test of the question where the outcome $\hat{s} = 0$ does not reject the null-hypothesis $\mathrm{sgn}(\beta) = 0$. If, however, the null-hypothesis is rejected in the pretest, then the observed sign is taken as an estimate for the true sign. Note that since $\gamma < \frac{1}{2}$ this procedure is more conservative than the Lasso in the sense that its "critical values" go to zero at a rate that is slower than $\frac{1}{\sqrt{n}}$. Indeed, we have that

$$\lim_{n\to\infty} \sup_{\beta\in\mathbb{R}} P\left(\hat{s}_n \cdot \mathrm{sgn}(\beta) = -1\right) = \lim_{n\to\infty} \Phi\left(-n^{-\gamma+\frac{1}{2}}\right) = 0.$$

Since the case $\hat{s}_n = 0$ is to be interpreted as "undecided", the estimator $\hat{s}_n$ is uniformly consistent in the sense that the supremum probability of obtaining a wrong sign-estimate tends to zero as $n$ goes to infinity. Hereby the rate of convergence depends on the tuning parameter $\gamma$.

Having obtained the sign-estimates, we denote the corresponding length-minimizing confidence intervals $K_n(\hat{s})$ in the following way:

- $K_n(1) = [\hat{\beta}_{\mathrm{L}} - a_n, \hat{\beta}_{\mathrm{L}} + b_n]$
- $K_n(-1) = [\hat{\beta}_{\mathrm{L}} - b_n, \hat{\beta}_{\mathrm{L}} + a_n]$
- $K_n(0) = [\hat{\beta}_{\mathrm{L}} - c_n, \hat{\beta}_{\mathrm{L}} + c_n]$

where $c_n$ is chosen such that $\Phi(\sqrt{n}c_n + \sqrt{n}\eta_n) - \Phi(-\sqrt{n}c_n + \sqrt{n}\eta_n) = 1 - \alpha$ for some prescribed level $\alpha \in (0,1)$. To determine the smallest confidence interval for the known-sign cases, we do the

following: For $a_n \geq 0$ and $b_n \geq 0$ and $\beta \geq 0$ Proposition 2 yields[24]

$$\begin{aligned}\inf_{\beta\geq 0} P_\beta(\beta \in [\hat{\beta}_{\mathrm{L}} - a_n, \hat{\beta}_{\mathrm{L}} + b_n]) &= \inf_{\beta\geq 0} P_\beta(-b_n \leq \hat{\beta}_{\mathrm{L}} - \beta \leq a_n) \\ &= P(-b_n \leq \hat{u}^+ \leq a_n) \\ &= P(-\sqrt{n}b_n + \sqrt{n}\eta_n \leq \sqrt{n}(\hat{u}^+ + \eta_n) \leq \sqrt{n}a_n + \sqrt{n}\eta_n) \\ &= \Phi(\sqrt{n}a_n + \sqrt{n}\eta_n) - \Phi(-\sqrt{n}b_n + \sqrt{n}\eta_n).\end{aligned}$$

where $\hat{u}^+ \sim N(-\eta_n, \frac{1}{n})$. From the previous display, we see that setting $\sqrt{n}a_n + \sqrt{n}\eta_n = \Phi^{-1}(1-\frac{\alpha}{2})$ and $-\sqrt{n}b_n + \sqrt{n}\eta_n = -\Phi^{-1}(1-\frac{\alpha}{2})$ will yield the shortest interval having coverage probability $1-\alpha$. This interval can, however, only be chosen as long as $\Phi^{-1}(1-\frac{\alpha}{2}) \geq \sqrt{n}\eta_n$, since otherwise, either $a_n < 0$ or $b_n < 0$, thus violating Condition A. In case $\Phi^{-1}(1-\frac{\alpha}{2}) < \sqrt{n}\eta_n$ it is easy to see that choosing $a_n = 0$ and $b_n$ such that $\Phi(\sqrt{n}\eta_n) - \Phi(-\sqrt{n}b_n + \sqrt{n}\eta_n) = 1-\alpha$ will yield the shortest interval. Inspection of the above formulae shows that the shortest Lasso-based confidence interval for a known sign is simply given by the "standard" confidence set based on the Least-squares estimator whenever the Lasso is contained in that interval. When, by heavy penalization, the Lasso lies outside the Least-squares interval, the estimator is taken as the end-point that is closer to the origin and the other endpoint is chosen to achieve the desired coverage probability. Note that by symmetry, the length-minimizing $a_n$ and $b_n$ are identical in case $\beta < 0$ and hence for $K_n(-1)$.

Note that, as mentioned earlier, we have $\inf_{\beta_n\in\mathbb{R}} P(\beta_n \in K_n(0)) = 1-\alpha$ and $\inf_{\beta:\ \mathrm{sgn}(\beta)=s} P(\beta \in K_n(s)) = 1-\alpha$ for $s \in \{-1,1\}$. The the vital question now is whether the proposed procedure, which switches between these intervals, yields asymptotically valid confidence sets, that is, whether

$$\liminf_{n\to\infty} \inf_{\beta\in\mathbb{R}} P(\beta \in K_n(\hat{s}_n)) \geq 1-\alpha.$$

It turns out that this does not hold, as is shown in the following result.

**Proposition 50.** *Let $K_n(s)$ be the above-defined minimum-length confidence intervals for known signs $s \in \{-1,1\}$ and unknown sign $s=0$ and assume that $\sqrt{n}\eta_n \longrightarrow \eta < \infty$. For the adaptive confidence interval based on the sign estimator $\hat{s}_n = \mathbb{1}_{\{\hat{\beta}_{LS} > n^{-\gamma}\}} - \mathbb{1}_{\{\hat{\beta}_{LS} < -n^{-\gamma}\}}$ and $\alpha < \frac{1}{2}$ we have that*

$$\liminf_{n\to\infty} \inf_{\beta\in\mathbb{R}} P_\beta(\beta \in K_n(\hat{s}_n)) < 1-\alpha.$$

*Proof.* To show that the actual coverage probability may indeed fall below the nominal level of $1-\alpha$ in the limit, as $n\to\infty$, it is sufficient to find a sequence $(\beta_n)_{n\geq 1}$ such that $P(\beta_n \in K_n(\hat{s}_n)) \longrightarrow q < 1-\alpha$.

Note that since $\sqrt{n}\eta_n \longrightarrow \eta < \infty$, we have that $\sqrt{n}\ c_n \longrightarrow c > \Phi^{-1}(1-\frac{\alpha}{2})$ as $n \to \infty$ (at least along some subsequence). For later use let $\hat{u}^\infty_{\mathrm{LS}} = \lim_{n\to\infty} \sqrt{n}\hat{u}_{\mathrm{LS}} = \lim_{n\to\infty} \sqrt{n}(\hat{\beta}_{\mathrm{LS}} - \beta_n)$ and

[24] Note that Condition A merely requires the confidence set to be an interval containing $\hat{\beta}_{\mathrm{L}}$ in case $p=1$.

note that both $\sqrt{n}\hat{u}_{\mathrm{LS}}$ and $\hat{u}^{\infty}_{\mathrm{LS}}$ follow a standard normal distribution. Now take the sequence

$$\beta_n = n^{-\gamma} - n^{-\frac{1}{2}}\left(\frac{c+\eta+\Phi^{-1}(1-\frac{\alpha}{2})}{2}\right).$$

First note that for this sequence of parameters we have

$$P(\hat{\beta}_{\mathrm{LS}} \geq \eta_n) = P(\sqrt{n}(\hat{\beta}_{\mathrm{LS}} - \beta_n) \geq \sqrt{n}\eta_n - \sqrt{n}\beta_n) = \Phi(-\sqrt{n\eta_n} + \sqrt{n}\beta_n) \longrightarrow 1$$

as $n$ goes to infinity and hence, by (4.6),

$$\hat{\beta}_{\mathrm{L}} = \hat{\beta}_{\mathrm{LS}} - \eta_n + o_p(1).$$

The desired coverage probability is now given by

$$P(\beta_n \in K_n(\hat{s}_n)) = P(\beta_n \in K_n(1), \hat{s}_n = 1) + P(\beta_n \in K_n(-1), \hat{s}_n = -1) + P(\beta_n \in K_n(0), \hat{s}_n = 0).$$

We will now treat these three terms separately. As we have seen before, the form of $K_n(1)$ depends on $\eta_n$. We thus first consider the case[25] where $\sqrt{n}\eta_n < \Phi^{-1}(1-\frac{\alpha}{2})$ for each $n$:

$$\begin{aligned}
&P\left(\beta_n \in K_n(1) \text{ and } \hat{s}_n = 1\right)\\
&= P\left(\hat{\beta}_{\mathrm{LS}} - \tfrac{1}{\sqrt{n}}\Phi^{-1}(1-\tfrac{\alpha}{2}) \leq \beta_n \leq \hat{\beta}_{\mathrm{LS}} + \tfrac{1}{\sqrt{n}}\Phi^{-1}(1-\tfrac{\alpha}{2}) \text{ and } \hat{\beta}_{\mathrm{LS}} > n^{-\gamma}\right) + o(1)\\
&= P\left(-\Phi^{-1}(1-\tfrac{\alpha}{2}) \leq \sqrt{n}(\hat{\beta}_{\mathrm{LS}} - \beta_n) \leq \Phi^{-1}(1-\tfrac{\alpha}{2})\sqrt{n} \text{ and } \sqrt{n}(\hat{\beta}_{\mathrm{LS}} - \beta_n) > \sqrt{n}(n^{-\gamma} - \beta_n)\right) + o(1)\\
&= P\left(\sqrt{n}\hat{u}_{\mathrm{LS}} > \tfrac{1}{2}(c+\eta+\Phi^{-1}(1-\tfrac{\alpha}{2})) \text{ and } -\Phi^{-1}(1-\tfrac{\alpha}{2}) \leq \sqrt{n}\hat{u}_{\mathrm{LS}} \leq \Phi^{-1}(1-\tfrac{\alpha}{2})\right) + o(1) = o(1),
\end{aligned}$$

since $c+\eta+\Phi^{-1}(1-\frac{\alpha}{2}) > 2\Phi^{-1}(1-\frac{\alpha}{2})$ and the event in the last line of the above display is thus empty. When $\sqrt{n}\eta_n \geq \Phi^{-1}(1-\frac{\alpha}{2})$ for each $n$, we have

$$\begin{aligned}
P\left(\beta_n \in K_n(1) \text{ and } \hat{s}_n = 1\right) &= P\left(\hat{\beta}_{\mathrm{L}} \leq \beta_n \leq \hat{\beta}_{\mathrm{L}} + b_n \text{ and } \hat{\beta}_{\mathrm{LS}} > n^{-\gamma}\right)\\
&= P\left(-b_n \leq \hat{\beta}_{\mathrm{LS}} - \eta_n - \beta \leq \text{ and } \hat{\beta}_{\mathrm{LS}} - \beta > n^{-\gamma} - \beta\right) + o(1)\\
&\leq P\left(\sqrt{n}(n^{-\gamma} - \beta) < \sqrt{n}(\hat{\beta}_{\mathrm{LS}} - \beta) \leq \sqrt{n}\eta_n\right) + o(1)\\
&= P\left(\tfrac{c+\eta+\Phi^{-1}(1-\frac{\alpha}{2})}{2} < \sqrt{n}\hat{u}_{\mathrm{LS}} \leq \eta + o(1)\right) = o(1),
\end{aligned}$$

since the probability in the last line of the above display is eventually empty, as $\Phi^{-1}(1-\frac{\alpha}{2}) > 0$ and $c > \eta$. (To see the latter inequality note that in the limit we have $1-\alpha = \Phi(c+\eta) - \Phi(-c+\eta) \leq 1 - \Phi(-c+\eta) \implies \Phi(-c+\eta) < \frac{1}{2} \implies c \geq \eta$, since $\alpha < \frac{1}{2}$.)

[25] Assume that either $\sqrt{n}\eta_n < \Phi^{-1}(1-\frac{\alpha}{2})$ for each $n$, or $\sqrt{n}\eta_n \geq \Phi^{-1}(1-\frac{\alpha}{2})$ for each $n$, otherwise pass to subsequences.

Next, we see that

$$P(\beta_n \in K_n(-1), \hat{s}_n = -1) \leq P(\hat{s}_n = -1) = P(\hat{\beta}_{\text{LS}} < -n^{-\gamma}) \longrightarrow 0$$

as $n \longrightarrow \infty$. Finally, we have

$$\begin{aligned}
&P\left(\beta_n \in K_n(0) \text{ and } \hat{s}_n = 0\right) \\
&= P\left(\hat{\beta}_{\text{LS}} - \eta_n - c_n \leq \beta_n \leq \hat{\beta}_{\text{LS}} - \eta_n + c_n \text{ and } -n^{-\gamma} \leq \hat{\beta}_{\text{LS}} \leq n^{-\gamma}\right) + o(1) \\
&= P\left(\sqrt{n}(-c_n + \eta_n) \leq \sqrt{n}\hat{u}_{\text{LS}} \leq \sqrt{n}(c_n + \eta_n) \text{ and } -2n^{\frac{1}{2}-\gamma} + O(1) \leq \sqrt{n}\hat{u}_{\text{LS}} \leq \tfrac{c+\eta+\Phi^{-1}(1-\frac{\alpha}{2})}{2}\right) + o(1) \\
&= P\left(c + \eta \leq \hat{u}^{\infty}_{\text{LS}} \leq c + \eta \text{ and } \hat{u}^{\infty}_{\text{LS}} \leq \tfrac{c+\eta+\Phi^{-1}(1-\frac{\alpha}{2})}{2}\right) + o(1) \\
&= \Phi\left(\tfrac{c+\eta+\Phi^{-1}(1-\frac{\alpha}{2})}{2}\right) - \Phi(-c + \eta) + o(1) \\
&= 1 - \alpha - \left(\Phi(c + \eta) - \Phi\left(\tfrac{c+\eta+\Phi^{-1}(1-\frac{\alpha}{2})}{2}\right)\right) + o(1).
\end{aligned}$$

Since $c + \eta > \Phi^{-1}(1 - \frac{\alpha}{2})$, we eventually have $P\left(\beta_n \in K_n(0) \text{ and } \hat{s}_n = 0\right) < 1 - \alpha$, completing the proof. □

**Remark 51.** *Note that the assumption that $\alpha < \frac{1}{2}$ is not too restrictive, since for almost all practical purposes $\alpha$ is typically chosen to be "close" to zero. Indeed, $\alpha \geq \frac{1}{2}$ would mean that the resulting "confidence set" is more likely not to contain the true parameter than it is to contain it, a choice that seems strange in almost any imaginable application.*

Proposition 50 shows that the proposed adaptive procedure does not yield valid confidence sets. Even though the procedure has no problems with parameter-sequences of order $\frac{1}{\sqrt{n}}$ by using a pre-test, or sign-estimator of higher order, $n^{-\gamma}$, the "problematic" rate of parameter sequences is merely shifted to the pre-test's rate. Also note that this problem arises, since the pre-test and the corresponding estimators are (necessarily) dependent. Considering the simplicity of the model and the fact that the assumptions in the setting under consideration are quite standard, this highly suggests that adaptive procedures for producing confidence sets are not uniformly valid in a larger class of settings.

## 4.6 Conclusion

In this chapter we have seen that the approach to construct confidence sets that are based on the Lasso estimator that is presented in Chapter 2 easily extends to the unknown variance case, where $\sigma^2$ has to be estimated, with only minor adaptations.

We also explored what Lasso-based confidence sets that are constructed based on the theory that has been developed in Chapter 2 look like when optimized to cover just one of the parameter

vector's components in a two-dimensional setup. It is illustrated that the length of these intervals increases with both the absolute correlation between the regressors, and the size of the penalization parameter. Similarly, a confidence set is also constructed for a partial Lasso where the non-penalized component of the parameter vector is to be covered. This approach somewhat mitigates the adverse effects of size of the intervals that arise from the regressor's correlation.

To study whether the shape and sub-parameter to be covered may be chosen in dependence of the Lasso's initial outcome, a simulation study was carried out. In the considered two-dimensional setting the results indicate that such a procedure may indeed be valid.

Finally, it is shown that an adaptive procedure for producing confidence sets in which the sign of the true parameter is first tested with a higher order pre-test does not yield valid inference, as the "problematic" sequences of parameters are merely shifted to the pre-test's order.

# Closing remarks

In this thesis, we have made advances in the knowledge about and understanding of the Lasso estimator.

We have found a stochastic bound for the Lasso's estimation error which enables the construction of uniformly valid confidence sets for the entire parameter vector in the, in practice quite relevant, low-dimensional setting, which had not been fully covered in the academic literature so far. To that end, we have derived a formula for the minimal coverage probability of a large class of sets that satisfy a rather mild condition which depends on the model's design matrix.

We have further explored various aspects of such confidence sets, such as the choice of shape both for the full parameter vector of interest, as well as a version that is optimized for a single component. We have also compared the Lasso and its confidence sets with the Least-squares estimator, showing that the resulting confidence sets will be larger than those based on the Least-squares estimator.

Last, but not least, we have thoroughly analyzed the Lasso estimator's distribution in both low- and high-dimensional settings and given its cumulative distribution function. While other authors had already provided this distribution in different ways before, the approach taken in this thesis is much more intuitive and easier to grasp. Furthermore, we have described the connection between the Lasso and the Least-squares estimator by giving a one-to-one relation between the two. In the high-dimensional setting we have also gained better insights in the estimator's behavior and in particular its model selection properties, showing the importance of the choice of the regressors' scaling to the procedure's users.

The author would like to thank his readers for their interest and hopes that they found the thesis to be stimulating. And while it has most likely not answered all questions regarding the topic, the author hopes that this thesis has given the reader a few illuminating insights, while sparking a few more questions and ideas that will lead to further research.

# Bibliography


Ali A., Tibshirani R. J. (2019). 'The generalized Lasso problem and uniqueness'. *Electronic Journal of Statistics* **13**:2307–2347.

Alliney S., Ruzinsky A. (1994). 'An algorithm for the minimization of mixed $l_1$ and $l_2$ norms with applications to Bayesian estimation' **42**:618–627.

Amann N., Schneider U. (2018). 'Asymptotic confidence regions based on the adaptive Lasso with partial consistent tuning'. *arXiv:1810.02665* .

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

Cai T., Guo Z. (2017). 'Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity'. *The Annals of Statistics* **45**:615–646.

Caner M., Kock A. B. (2018). 'Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative Lasso'. *Journal of Econometrics* **203**:143–168.

Efron B., et al. (2004). 'Least angle regression'. *The Annals of Statistics* **32**:407–499.

El Ghaoui L., et al. (2012). 'Safe feature elimination for the Lasso and sparse supervised learning problems'. *Pacific Journal of Optimization 8, 667-698* **abs/1009.3515**.

Ewald K., Schneider U. (2018). 'Uniformly valid confidence sets based on the Lasso'. *Electronic Journal of Statistics* **12**:1358–1387.

Ewald K., Schneider U. (2020). 'On the distribution, model selection properties and uniqueness of the Lasso estimator in low and high dimensions'. *Electronic Journal of Statistics* **14**:944–969.

Friedman A. (1982). *Foundations of modern analysis*. Dover Books on Mathematics Series. Dover.

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

Hyun S., et al. (2016). 'Exact post-selection inference for changepoint detection and other generalized Lasso problems'. *arXiv:1812.03644* .

Jagannath R., Upadhye N. S. (2018). 'The Lasso estimator: Distributional properties'. *Kybernetika* **54**:778–797.

Javanmard A., Montanari A. (2014a). 'Confidence intervals and hypothesis testing for high-dimensional regression'. *Journal of Machine Learning Research* **15**:2869–2909.

Javanmard A., Montanari A. (2014b). 'Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic Theory'. *IEEE Trans. Information Theory* **60**:6522–6554.

Kabaila P., Leeb H. (2006). 'On the large-sample minimal coverage probability of confidence intervals after model selection'. *Journal of the American Statistical Association* **101**:619–629.

Kivaranovic D., Leeb H. (2018). 'On the length of post-model-selection confidence intervals conditional on polyhedral constraints'. *arXiv:1803.01665* .

Knight K., Fu W. (2000). 'Asymptotics of Lasso-type estimators'. *The Annals of Statistics* **28**:1356–1378.

Kotz S., Nadarajah S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Lee J. D., et al. (2016). 'Exact post-selection inference, with application to the Lasso'. *The Annals of Statistics* **44**:907–927.

Leeb H., et al. (2015). 'On various confidence intervals post-model-selection'. *Statistical Science* **30**:216–227.

Liu K., et al. (2018). 'More powerful post-selection inference, with application to the Lasso'. *arXiv:1801.09037* .

Lockhart R., et al. (2014). 'A significance test for the Lasso'. *The Annals of Statistics* **42**:413–468.

Meir A., Drton M. (2017). 'Tractable post-selection maximum likelihood inference for the Lasso'. *arXiv:1705.09417* .

Min S., Zhou Q. (2019). 'Constructing confidence sets after Lasso selection by randomized estimator augmentation'. *arXiv:1904.08018* .

Miolane L., Montanari A. (2018). 'The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning'. *arXiv:1811.01212* .

Ndiaye E., et al. (2017). 'GAP safe screening rules for sparse-group-Lasso'. *Journal of Machine Learning Research* **18**:1–33.

Pötscher B. (1991). 'Effects of model selection on inference' **7**:163–185.

Pötscher B. M., Leeb H. (2009). 'On the distribution of penalized maximum likelihood estimators: The Lasso, SCAD, and thresholding' **100**:2065–2082.

Pötscher B. M., Prucha I. R. (2001). *Basic elements of asymptotic theory*.

Pötscher B. M., Schneider U. (2009). 'On the distribution of the adaptive Lasso estimator' **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.

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'. *The Annals of Statistics* **35**:1012–1030.

Schneider U. (2016). 'Confidence sets based on thresholding estimators in high-dimensional Gaussian regression models'. *Econometric Reviews* **35**:1412–1455.

Schneider U., Tardivel P. (2020). 'The geometry of uniqueness, sparsity and clustering in penalized estimation'. *arXiv:2004.09106* .

Sepehri A., Harris N. (2017). 'The accessible lasso models'. *Statistics* **51**:711–721.

Taylor J., Tibshirani R. J. (2015). 'Statistical learning and selective inference'. *Proceedings of the National Academy of Sciences* **112**:7629–7634.

Tian X., Taylor J. (2017). 'Asymptotics of selective inference'. *Scandinavian Journal of Statistics* **44**:480–499.

Tibshirani R. (1996). 'Regression shrinkage and selection via the Lasso' **58**:267–288.

Tibshirani R., et al. (2012). 'Strong rules for discarding predictors in lasso-type problems'. *Journal of the Royal Statistical Society Series B* **74**:245–266.

Tibshirani R. J. (2013). 'The lasso problem and uniqueness'. *Electronic Journal of Statistics* **7**:1456–1490.

Tibshirani R. J., et al. (2018). 'Uniform asymptotic inference and the bootstrap after model selection'. *The Annals of Statistics* **46**:1255–1287.

Tibshirani R. J., et al. (2016). 'Exact post-selection inference for sequential regression procedures'. *Journal of the American Statistical Association* **111**:600–620.

Van de Geer S. (2017). 'On the efficiency of the de-biased Lasso'. *arXiv:1708.07986* .

Van de Geer S., et al. (2014). 'On asymptotically optimal confidence regions and tests for high-dimensional models'. *The Annals of Statistics* **42**:1166–1202.

Van de Geer S., Stucky B. (2016). '$\chi$2-confidence sets in high-dimensional regression'. In A. Frigessi, P. Bühlmann, I. K. Glad, M. Langaas, S. Richardson, & M. Vannucci (eds.), *Statistical Analysis for High-Dimensional Data*, pp. 279–306, Cham. Springer International Publishing.

Yuan M., Lin Y. (2007). 'On the non-negative garrotte estimator'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**:143–161.

Zhang C.-H., Zhang S. S. (2014). 'Confidence intervals for low dimensional parameters in high dimensional linear models'. *Journal of the Royal Statistical Society Series B* **76**:217–242.

Zhao P., Yu B. (2006). 'On model selection consistency of Lasso' **7**:2541–2563.

Zhao S., et al. (2017). 'In defense of the indefensible: A very naive approach to high-dimensional inference'. *arXiv:1705.05543* To appear in Statistical Science.

Zhou K., et al. (2019). 'Honest confidence sets for high-dimensional regression by projection and shrinkage'. *arXiv:1902.00535* .

Zhou Q. (2014). 'Monte Carlo simulation for Lasso-type problems by estimator augmentation'. *Journal of the American Statistical Association* **109**:1495–1516.

Zou H. (2006). 'The adaptive Lasso and its oracle properties' **101**:1418–1429.

# Appendices

# Appendix A

# Additional figures

This chapter presents additional figures that were left out of the main part as not to disturb the reading flow. The figures display the (unconditional) coverage[1] of the confidence sets constructed in Section 4.4 for two additional examples of covariance matrices of the regressors. Apart from indicating that the procedure indeed yields valid confidence sets, one can see that the produces more and conservative sets for the points that are not "close" to the minimal coverage probability as the absolute correlation between the regressors increases. This can be seen in Figure A.1 and Figure A.3. Additionally, the corresponding (simulated) model selection probabilities are displayed in Figure A.2 and Figure A.2.

[1] I..e, formula (4.4).

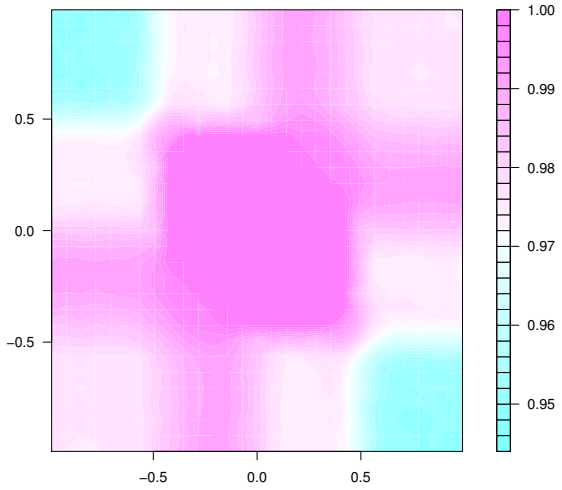



Figure A.1: The coverage probabilities according to simulations in dependence of $\frac{1}{2\sqrt{n}}\beta_1$ (x-axis) and $\frac{1}{2\sqrt{n}}\beta_2$ (y-axis) for $c_{12}^* = 0.25$ and $\lambda = (\sqrt{n}, \sqrt{n})'$.

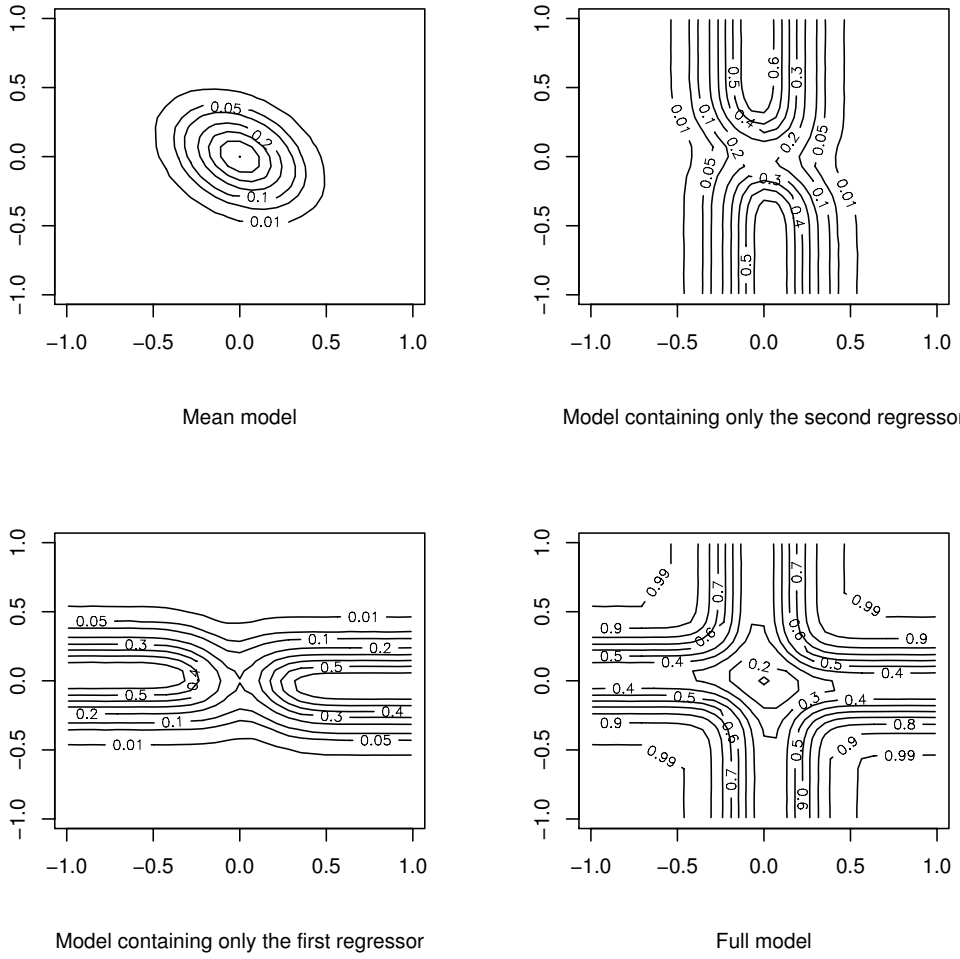



Figure A.2: The model selection probabilities according to simulations in dependence of $\frac{1}{2\sqrt{n}}\beta_1$ (x-axis) and $\frac{1}{2\sqrt{n}}\beta_2$ (y-axis) for $c_{12}^* = 0.25$ and $\lambda = (\sqrt{n}, \sqrt{n})'$.

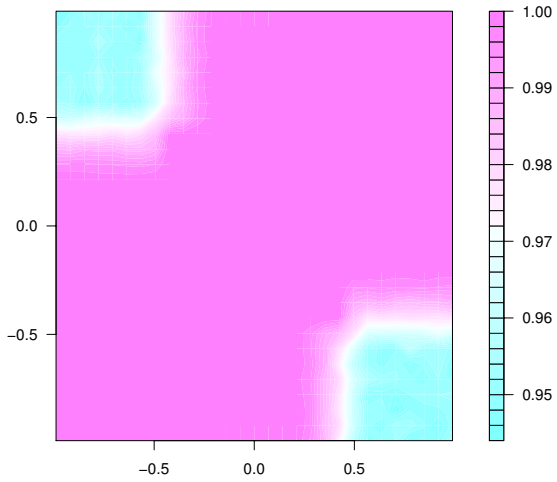



Figure A.3: The coverage probabilities according to simulations in dependence of $\frac{1}{2\sqrt{n}}\beta_1$ (x-axis) and $\frac{1}{2\sqrt{n}}\beta_2$ (y-axis) for $c_{12}^* = 0.75$ and $\lambda = (\sqrt{n}, \sqrt{n})'$.

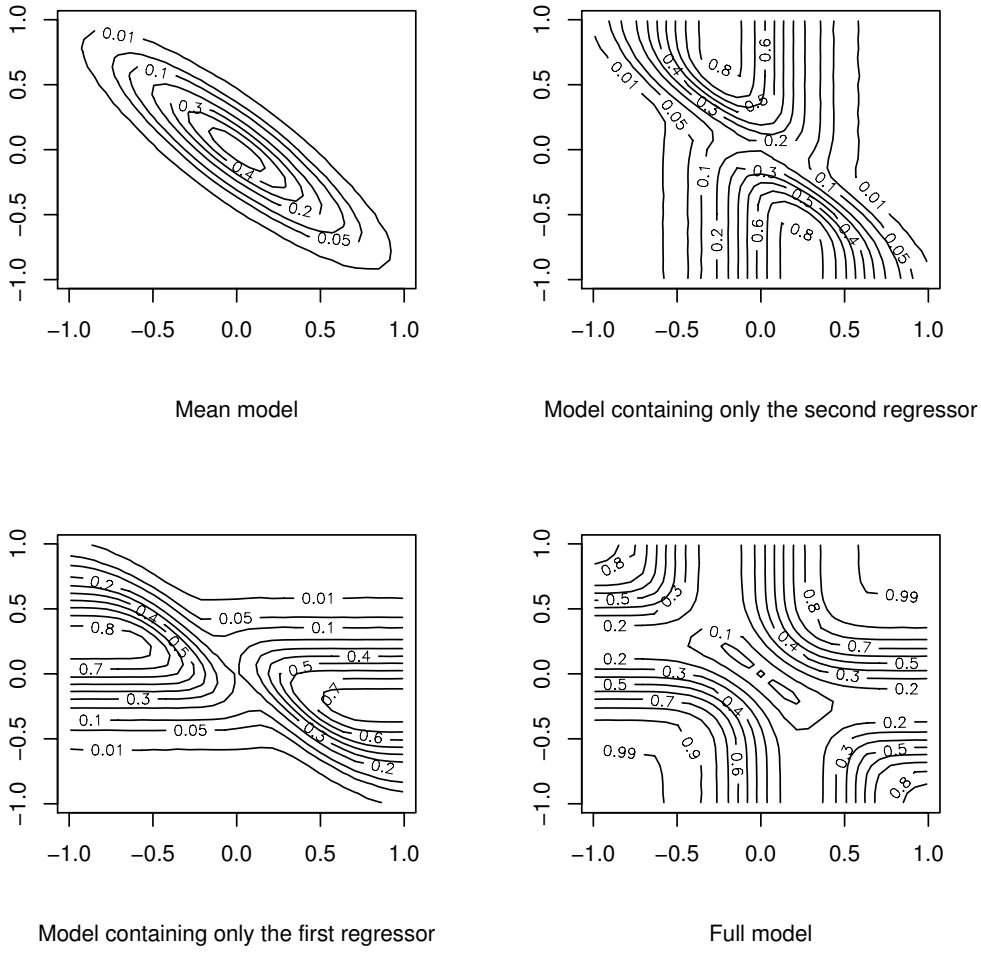



Figure A.4: The model selection probabilities according to simulations in dependence of $\frac{1}{2\sqrt{n}}\beta_1$ (x-axis) and $\frac{1}{2\sqrt{n}}\beta_2$ (y-axis) for $c_{12}^* = 0.75$ and $\lambda = (\sqrt{n}, \sqrt{n})'$.

# Appendix B

# Simulation code

This Chapter contains the essential code used for the Monte Carlo simulations that were performed to obtain the results in Section 4.5. The simulations were conducted using the statistical software package R.

The code in the succeeding display generates response data $y$ for given regressor matrix, tuning parameter $\lambda$ and regression parameter $\beta$. It then calculates the Lasso estimator and checks whether the true parameter is covered by the Confidence set that is constructed using an elliptic shape. Finally, the above-described steps are repeated for a given number of times and averages over the results taken to arrive at the (approximate) model selection probabilities of the estimator, as well as coverage probabilities both conditional and unconditional on the selected model. Note that the conditional coverage probabilities are less accurate than the unconditional ones, as less observations are available per selected model. The results should thus be viewed bearing the model selection probability in mind, especially if it is quite low.

```
lassofkt <- function(y_, X_, lambda_, par_){
t(y_-X_ %*% par_)%*%(y_-X_ %*% par_)+2*lambda_*sum(abs(par_))
}

lasso <- function(lambda_,y_,X_,digits=3){
round(optim(par=c(0,0), lassofkt, y_=y_, X_=X_, lambda_=lambda_)$par,
digits=digits)
}

sim.fkt1a <-
function(X, beta=c(0,0), sd.epsilon=1, lambda=NULL, a=3.9, k=14.7, n.rep=100000,
digits=3){
n <- length(X[,1])
XX <- t(X) %*%X
```

```
sqrt.n <- sqrt(n)
if(is.null(lambda)){lambda <- sqrt.n}
cvrd.vec <- rep(0, times=n.rep)
model.vec <- rep("00", times=n.rep)
C <- t(X) %*% X/n
for(i in 1:n.rep){
epsilon <- rnorm(n=n, mean=0, sd=sd.epsilon)
y <- X %*% beta+epsilon
beta.hat <- lasso(lambda_=lambda, y_=y, X_=X, digits = digits)
if(beta.hat[1]==0 && beta.hat[2]==0){
if(t(beta.hat-beta) %*% XX %*% (beta.hat-beta) < k){cvrd.vec[i] <- 1}
}
else if(beta.hat[1]!=0 && beta.hat[2]!=0){
model.vec[i] <- "11";
if(t(beta.hat-beta) %*% XX %*% (beta.hat-beta) < k){cvrd.vec[i] <- 1}
}
else if(beta.hat[2]==0){
model.vec[i] <- "10";
if(sqrt.n*abs(beta[1]-beta.hat[1])<a){cvrd.vec[i]<- 1}
}
else{
model.vec[i] <- "01"; if(sqrt.n*abs(beta[2]-beta.hat[2])<a){cvrd.vec[i]<- 1}
}
}

res <- NULL
res$coverage <- mean(cvrd.vec)
res$model <- c(sum(model.vec=="00"), sum(model.vec=="01"),
sum(model.vec=="10"),sum(model.vec=="11"))
res$cond_coverage00 <- mean(cvrd.vec[model.vec=="00"])
res$cond_coverage01 <- mean(cvrd.vec[model.vec=="01"])
res$cond_coverage10 <- mean(cvrd.vec[model.vec=="10"])
res$cond_coverage11 <- mean(cvrd.vec[model.vec=="11"])
return(res)
}
```

The function has been slightly altered to generate a new regressor matrix $X$ in each repetition to verify that the results do not depend on the particular regression matrix. The adapted function is not reported as the changes are quite minor.

Note that the constants defining the size of the confidence sets were obtained using numerical integration in the software package *Mathematica*. Since this is a straight-forward process, these codes will not be presented here.

# Appendix C

# Definitions and results used in the thesis

This chapter gives an overview of generally known results that are used or mentioned in the thesis. To avoid confusion, it also gives the definitions of some concepts that sometimes are defined in slightly different ways.

## C.1 Definitions

### C.1.1 Multivariate t-distributions

This Section is based on Kotz & Nadarajah (2004).

In general there are multiple forms of multivariate t-distributions, we will present the most common and most natural one, in the sense that it can be derived from the a multivariate normal distribution in the same way the univariate t-distribution can be derived from a univariate normal distribution.

**Definition 52.** *Let $Z$ be a random vector following a $p$-variate normal distribution with mean zero and variance-covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. Moreover, let $\chi_k$ denote an independent $\chi^2$ distributed variable and let $\mu \in \mathbb{R}^p$. Then, the variable*

$$T = \frac{Z}{\sqrt{\frac{\chi_k}{k}}} - \mu$$

*is said to follow a multivariate t-distribution with $k$ degrees of freedom, correlation matrix $\Sigma$ and non-centrality parameter $\mu$ and has the Lebesgue-density*

$$f(z) = \frac{\Gamma(\frac{k+p}{2})}{(\pi k)^{(p/2)}\Gamma(\frac{k}{2})\sqrt{\det(\Sigma)}}\left(1 + \frac{z'\Sigma^{-1}z}{k}\right)^{-\frac{k+p}{2}}, \tag{C.1}$$

*where $\Gamma(\cdot)$ denotes the Gamma function.*

**Remark 53.** *It is easily seen that if $Z \sim N(0, \Sigma)$ for some $p$-dimensional correlation matrix $\Sigma$ and $T$ follows a (central) t-distribution with correlation matrix $\Sigma$ and any number of degrees of freedom $k$, then the contour lines of the distributions of $T$ and $Z$ have the same shape in the sense that for each $\varrho \geq 0$ there exists an $\rho \geq 0$ such that*

$$\{z \in \mathbb{R}^p : f(z) \leq \varrho\} = \{z \in \mathbb{R}^p : \phi_{(0,\Sigma)}(z) \leq \rho\},$$

*since the contour sets of both densities will be of the form*

$$\{x \in \mathbb{R}^p : x'\Sigma^{-1}x \leq \rho\},$$

*for some constant $\rho \geq 0$, as is seen by inspecting the corresponding Lebesgue densities. Note, however, that any set of the form $\{z \in \mathbb{R}^p : z'\Sigma^{-1}z \leq \rho\}$ will correspond to different contour-levels for the two distributions, in general.*

### C.1.2 General position

The concept of the regressor matrix' columns being in *general position* can shown to be a sufficient condition for uniqueness of the Lasso, c.f. Lemma 3 in Tibshirani (2013)[1]. Since it is easily confused with another concept of the same name, the definition is given in the following.

**Definition 54** (General Position)**.** *We say that the matrix $X \in \mathbb{R}^{n\times p}$ has columns in general position if no $k$-dimensional affine subspace $L \subseteq \mathbb{R}^n$, for $k < \min\{n, p\}$, contains more than $k + 1$ elements of the set $\{\pm x_{\cdot 1}, \ldots, \pm x_{\cdot p}\}$, excluding antipodal pairs.*

*Another way of saying this: The affine span of any $k + 1$ points $s_1 x_{i,1}, \ldots, s_{k+1} x_{i,k+1}$ , for arbitrary signs $s_1, \ldots, s_{k+1} \in \{-1, 1\}$, does not contain any element of $\{\pm x_{\cdot i} : i \neq i_1, \ldots, i_{k+1}\}$.*

## C.2 On the asymptotics of convex stochastic optimization

In this section we take a look at the result that ensures that the minimizers of a sequence of convergent random functions indeed converges in distribution to the minimizer of the limiting function, thus enabling the proofs of Proposition 11 and Proposition 15. This result has been provided by Geyer (1996) and are much more general than needed in our application. To keep the notation as simple as possible we shall thus give a simplified version of the main result that is sufficient for our needs.

We start with a definition.

[1]The definition given here is also based on this reference, but clarifies that the subspaces used to define the condition are indeed affine subspaces.

**Definition 55.** *Let $C(\mathbb{R}^p)$ denote the space of all continuous functions from $\mathbb{R}^p$ to $\mathbb{R}$. Let*

$$\|F\|_K = \sup_{x \in K} |F(x)|.$$

*We say that a sequence $\{F_n\}_{n \geq 1} \subseteq C(\mathbb{R}^p)$ converges to a function $F \in C(\mathbb{R}^p)$ in the topology of uniform convergence on compact sets, if*

$$\|F_n - F\|_K \longrightarrow 0$$

*as $n \longrightarrow \infty$ for every compact subset $K$ of $\mathbb{R}^p$.*

Albeit not completely obvious it can be shown that this notion of convergence is induced by the metric

$$d(F; G) = \sum_{n=1}^{\infty} \frac{2^{-n} \|F - G\|_{B_n}}{1 + \|F - G\|_{B_n}}$$

where $B_n$ denotes a ball around zero having radius $n$ and the space $C(\mathbb{R}^p)$ equipped with the topology of uniform convergence is in fact a metric space.

Equipped with this concept we can now state the theorem that enables the asymptotic results in Chapter 2.

**Theorem 56.** *Suppose $F_n$ is a sequence of random elements of the Polish space $C(\mathbb{R}^p)$ of all continuous functions from $\mathbb{R}^p$ to $\mathbb{R}$ with the metric of uniform convergence on compact sets and $F$ is another random element of that Polish space having the property that $F$ has a unique global minimizer almost surely. Suppose $x_n$ is a sequence of random vectors. If $f_n \longrightarrow_d f$ and $x_n$ is bounded in probability and $F_n(x_n) - \inf_{y \in \mathbb{R}^p} F_n(y) \longrightarrow_d 0$, then*

$$x_n \longrightarrow_d x$$

*where $x$ is another random vector and*

$$F_n(x_n) \longrightarrow_d F(x)$$

*and $F(x) = \inf_{y \in \mathbb{R}^p} F(y)$ almost surely.*

## C.3 Further results

This Section contains two classical results from both Functional Analysis as well as Probability Theory that are used in the thesis. We start with The Banach-Steinhaus Theorem which is used in the proof of Remark 10.

**Theorem 57** (Banach-Steinhaus[2])**.** *Let $X$ be a Banach space and let $Y$ be a normed linear space. Let $\{T_\alpha\}$ be a family of bounded linear operators from $X$ to $Y$ and let $\|\cdot\|_{op}$ denote the operator norm. If for each $x \in X$ the set $\{T_\alpha x\}$ is bounded, then the set $\{\|T_\alpha\|_{op}\}$ is bounded.*

We now turn to the Central Limit Theorem which ensures the asymptotic normality of the Least-squares estimator in our setting and thus also enables the analysis of the asymptotic behavior of the Lasso estimator in Section 2.6.1. The following version of the CLT is Theorem 4.7 from Pötscher & Prucha (2001) with the notation being adapted to the one used in this dissertation.

**Theorem 58** (Central Limit Theorem)**.** *Let $\varepsilon_{n,i}, i \geq 1$ be a sequence of i.i.d. random variables with $E(\varepsilon_{n,i} = 0)$ and $E(\varepsilon_{n,i}^2) < \infty$. Let $X = X_n = (x_{ij})$ be a sequence of real non-stochastic $n \times p$ matrices with $\lim_{n\to\infty} \frac{1}{n} X_n' X_n = C_\infty$. Then*

$$\frac{1}{\sqrt{n}} X_n' \varepsilon_n \longrightarrow N(0, \sigma^2 C_\infty)$$

*as $n \longrightarrow \infty$.*

[2]As formulated in Friedman (1982), p.139.

# Appendix D

# Author's Curriculum Vitae

Karl Ewald was born on October 28th 1987 in Vienna, Austria. He attended primary school in Krottenbachstraße 108, 1190 Vienna. After graduating with distinction from high school, BRG 18 Schopenhauerstraße in 2006, he started studying Economics at University of Vienna. During that time he discovered his interest in Mathematics and Statistics, taking additional courses in those fields and eventually switching to a Statistics curriculum once having earned his Bachelor's degree in Economics in 2010. Karl Ewald concluded his Master's studies of Statistics in 2012 with distinction and subsequently started working as a pre-doctoral researcher at Vienna University of Technology. Following his short career in academia, he moved to the private sector in 2016.