Ambrus Tamás

# Non-Asymptotic Confidence Regions for the Least Absolute Deviations Estimate

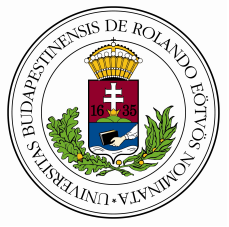
Thesis

Mathematics BSc
Applied Mathematics major

Supervisors:

Balázs Csanád Csáji
Senior Reasearch Fellow
Hungarian Academy of Sciences
Institute for Computer Science and Control (SZTAKI)

Ágnes Mariann Backhausz
Assistant Professor
ELTE, Department of Probability Theory and Statistics




Eötvös Loránd University
Faculty of Science

2018

**Acknowledgement**

First of all I would like to say thanks to Balázs Csanád Csáji, my supervisor who was very helpful when I searched for the thesis theme and introduced this specific topic to me. I am very grateful for him for challenging me and encouraging me to write this thesis. I also would like to express my thankfulness for spending so much of his precious time patiently consulting with me about my questions. I could learn a lot from him during these times.

I would also like to say thanks to Ágnes Mariann Backhausz who was my Probability Theory teacher and during my years in the university brought this specific field close to me. I am grateful for her for inspiring me to ponder about problems relating to Probability Theory. Many thanks for helping me to find the proper supervisor who is proficient in the topic I was interested in. I really appreciate her advices on this thesis as well.

# Contents

# Chapter 1

# Introduction

Linear regression has become an extremely important topic of modern statistical science applied for example in system identification, machine learning and financial mathematics. Finding suitable models to observed unknown systems is a classic statistical problem. The linear model developed by Gauss is applied all around the world in a variety of fields. Standard methods pick a single "point estimate" out of the possible ones, but sometimes we need a guarantee of finding the true model. For that reason we often build confidence regions around the point estimators. These set estimators are usually based on the asymptotic property (e.g. limiting distribution) of a given point estimator, therefore they can only be applied when the number of observations is large enough.

In this thesis we are going to investigate two well-known point estimators and get to know an algorithm, namely the Sign-Perturbed Sums (SPS) method, which builds distribution-free exact confidence regions around these estimators even for finitely many observations.

First of all the least squares (LS) estimator is going to be introduced. Its most important properties are going to be overviewed and the corresponding fundamental results are going to be presented.

Second, the Sign-Perturbed Sums (SPS) method is going to be introduced. Examining this method is the main subject of this thesis. This algorithm has been developed by my supervisor Csáji Balázs, Erik Weyer and Marco Campi in 2012 [10]. Its main advantage is that it builds an exact confidence region with user-chosen probability for finitely many observations under very mild assumptions on the noise. In the original version, this method builds the region around the LS estimator. This is why LS is thoroughly examined before this section. The SPS' nice properties are summerized and an extension is made

for autoregressive systems. These results were published by the mentioned authors.

In the second part of the thesis the least absolute deviatons (LAD) estimator is going to be investigated. Similarly to the LS its most important both asymptotic and non-asymptotic properties are going to be presented. Even a short comparison will be made to the LS.

Finally a modified version of the SPS is going to be shown which builds exact confidence region around the LAD estimator. There are open questions about the confidence region that are constructed by this algorithm. The pointwise consistency for this method was proven and the extention of this algortihm to the ARX system was made by me and my supervisor. It is in the last section of the thesis.

In the appendix some simulation examples and the MATLAB codes are presented for demonstrating the algorithm.

# Chapter 2

# Least Squares Criterion

## 2.1 Least squares data fitting

In this chapter I am going to introduce the least squares (LS) estimate which is a standard method to approximate solutions in linear systems. It is extremely important because of its many advantageous properties and extensive applicability. It was first introduced by Legendre [17] and Gauss [15] in the beginning of the $19^{th}$ century. Since then an enourmous developement has been made on the method in terms of probability theory by Gauss himself, Laplace and many others. Nowadays its leading role in linear regression is not questionable. The most importantant features are going to be presented in this chapter. We are also going to see some of the fundamental results related to this particular estimate such as the Gauss-Markov theorem and confidence ellipsoids.

Point estimation is a method which, trying to find the true parameter of an unknown random variable, picks a single point out of the parameter space based on finitely many observations. There are several point estimators (such as maximum likelihood, method of moments etc.) based on different ideas, each with a variety of great properties. One well-known estimator is the Least Sqaures estimator (LS). When we have a finite sample $\{(Y_1, \varphi_1), (Y_2, \varphi_2), ..., (Y_n, \varphi_n)\}$ it is reasonable to search for the parameter $\theta$ which gives us the solution with the least residual w.r.t. $Y_t - \varphi_t^T\theta$ in any given data point. "Least squares" means that we want to minimize the sum of the squares of the residuals made in each observed point. There are other methods minimizing different moments of the residual function (see Chapter 3), but this chapter deals with the LS.

### 2.1.1 Determenistic case

#### Normal equation

Assume that for a certain linear system with known structure but unknown true parameter $\theta^*$ we have recorded inputs and outputs over a time interval $1 \leq t \leq n$:

$$\{(Y_1, \varphi_1), (Y_2, \varphi_2), ..., (Y_n, \varphi_n)\}.$$

In this notation $Y \in \mathbb{R}^n$ is the output also called "dependent variable" vector where $Y_t \in \mathbb{R}$ is the output for any given $t$ and $\{\varphi_t\}_{t=1}^n \subseteq \mathbb{R}^d$ are the inputs also called regressors or "explanatory variables". We assume that we have a linear relation between the variables of the system and the output. Assuming a linear system behind simplifies the real problem, but also makes it easier to find applicable estimators. In ideal circumstances the following equation is true in linear systems:

$$Y_t = \varphi_t^T \theta^* + N_t \tag{2.1}$$

where $N_t$ is real valued variable for all $t = 1, \ldots, n$ indicating the noise in each measurement. The constant unknown true parameter is $\theta^*$. Our goal is to find this parameter. We assume that this equation describes the observed system correctly.

Based on this equation that we know for linear systems we can predict a certain $Y_t$. For any $\theta$ parameter we can calculate the prediction for $\widehat{Y}_t$ this way:

$$\widehat{Y}_t(\theta) = \varphi_t^T \theta \tag{2.2}$$

where $\widehat{Y}_t(\theta)$ is called the predictor function and the quantity of $Y_t - \widehat{Y}_t(\theta)$ is called the *residual* or *prediction error* for a given $t$ and $\theta$. For all $t = 1, \ldots, n$, $\varphi_t \in \mathbb{R}^d$ are assumed to be constant regressors for simplicity. Let $\Phi$ denote the matrix we get if we put the row vectors $\varphi_t^T$ for each $t = 1, \ldots, n$ into a matrix this way

$$\Phi \triangleq \begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_n^T \end{bmatrix}.$$

Using this matrix we can rewrite the known equation this way

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_n^T \end{bmatrix} \begin{bmatrix} \theta_1^* \\ \theta_2^* \\ \vdots \\ \theta_d^* \end{bmatrix} + \begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_n \end{bmatrix} \tag{2.3}$$

where $\theta_i^*$ is the corresponding coordinate of the true parameter $\theta^*$. Similiarly in a more simple notation

$$Y = \Phi\theta^* + N. \tag{2.4}$$

**A 1.** *Assume that $\Phi$ is "skinny" ($n \geq d$) and has a full column rank (rank$(\Phi) = d$).*

Having more observations than variables is an obvious criterion for estimation. Fewer variables would lead us into different problems such as system design or controll. We are going to see later that full column rank is also important, though it is still a very mild condition.

As it was mentioned earlier, we would like to minimize the squares of the residuals to approximate the true parameter $\theta^*$. We choose the parameter where this minimum is pursued for our estimate.

$$\widehat{\theta}_n := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{t=1}^{n} (Y_t - \widehat{Y}_t(\theta))^2 \tag{2.5}$$

If we apply the linear prediction for $\widehat{Y}_t(\theta)$ the estimator can be expressed as

$$\widehat{\theta}_n := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{t=1}^{n} (Y_t - \varphi_t^T\theta)^2. \tag{2.6}$$

In addition, if we use the matrix $\Phi$, it simplifies the notation to

$$\widehat{\theta}_n := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|Y - \Phi\theta\|^2. \tag{2.7}$$

Notice that the sum we want to minimize is a convex quadratic function of $\theta$ and it is greater than (or equal to) zero. Therefore, we can seek a minimum point by taking the derivative with respect to $\theta$ and set it equal to zero. As we take the gradient we get this:

$$\frac{\partial}{\partial\theta}(\|Y-\Phi^T\theta\|^2) = -2\Phi^T(Y-\Phi\theta) \tag{2.8}$$

Consequently

$$-2\Phi^T(Y-\Phi\theta) = 0 \tag{2.9}$$

$$\Phi^T Y = \Phi^T\Phi\widehat{\theta}_n. \tag{2.10}$$

This equation is called the normal equation. We see that $(Y-\Phi\theta)$ is orthogonal to the column vectors of $\Phi^T$. The $\Phi$ matrix is "skinny" and has a full column rank therefore $\Phi^T\Phi$ is invertible so there is only one unique solution to the problem. It can be calculated this way:

$$\widehat{\theta}_n = (\Phi^T\Phi)^{-1}\Phi^T Y \tag{2.11}$$

As it was shown, this estimate has a unique analitical solution under mild assumptions which makes it possible to use in many cases.

### Linear solution

The least squares estimate is linear. As we saw in (2.11) this estimate has a unique analitical solution. The matrix that we have on the right hand side of the equation $(\Phi^T\Phi)^{-1}\Phi^T \in \mathbb{R}^{d\times n}$ is a linear operator.

### Orthogonal projection

**Theorem 2.1.1.** *The matrix $P = \Phi(\Phi^T\Phi)^{-1}\Phi^T$ is an orthogonal projection on the range of $\Phi$ assuming A1.*

It means that $P$ is idempotent and self-adjoint.

*Proof.* First, I am going to show that P is idempotent meaning that $P^2 = P$.

$$P^2 = (\Phi(\Phi^T\Phi)^{-1}\Phi^T)^2 = \Phi(\Phi^T\Phi)^{-1}\Phi^T\Phi(\Phi^T\Phi)^{-1}\Phi^T \tag{2.12}$$

$$= \Phi[(\Phi^T\Phi)^{-1}(\Phi^T\Phi)](\Phi^T\Phi)^{-1}\Phi^T \tag{2.13}$$

$$= \Phi(\Phi^T\Phi)^{-1}\Phi^T = P \tag{2.14}$$

Second, I prove that P is self-adjoint.

$$P^T = (\Phi(\Phi^T\Phi)^{-1}\Phi^T)^T = \Phi[(\Phi^T\Phi)^{-1}]^T\Phi^T = \Phi[(\Phi^T\Phi)^T]^{-1}\Phi^T = \Phi(\Phi^T\Phi)^{-1}\Phi^T = P \tag{2.15}$$

We get that P is an orthogonal projection. □

### Moore-Penrose pseudoinverse

In linear algebra an $A \in \mathbb{R}^{m\times n}$ does not have an inverse. For this reason a generalization of the inverse concept has been introduced. The widely known Moore-Penrose pseudoinverse was developed independently by Moore [18] (1920), Bjerhammar [4](1952) and Penrose [19] (1955).

**Definition 2.1.2.** *For an $A \in \mathbb{R}^{m\times n}$ a pseudoinverse of $A$ is $A^+ \in \mathbb{R}^{n\times m}$ that satisfies these criterions:*

$$AA^+A = A \tag{2.16}$$

$$A^+AA^+ = A^+ \tag{2.17}$$

$$(AA^+)^{\mathrm{T}} = AA^+ \tag{2.18}$$

$$(A^+A)^{\mathrm{T}} = A^+A. \tag{2.19}$$

Note that the concept of the Moore-Penrose pseudoinverse can be extended to complex matrices. In that case reasonably in the third and fourth criterions we need to adjungate $AA^+$ and $A^+A$ instead of just transposing. Though when $A$ is real valued its pseudoinverse is real valued as well so now we do not need to deal with the complex case. We know that $A^+$ always exists and is unique [22]. If the first criterion is met for a matrix it is called generalized inverse. If the second criterion is satisfied then the matrix is called generalized reflexive inverse. Uniqueness is not always held for generalized inverse. The last two criterions makes the Moore-Penrose pseudoinverse unique. Notice that when $A$ has a full column rank we can find a nice algebraic formula for the pseudoinverse.

$$A^+ = (A^{\mathrm{T}}A)^{-1}A \tag{2.20}$$

This formula is the same that we got by solving the normal equation for the LS estimator,

therefore LS is a special case of the Moore-Penrose pseudoinverse. This way

$$\widehat{\theta} = A^{+}Y \tag{2.21}$$

estimator always exists and when $A$ is skinny and has a full column rank $\widehat{\theta}$ equals to the LS estimator, but when the equation is underdetermined it gives us the "Least-norm" solution.

### 2.1.2 Stochastic case

#### Unbiasedness

Assume we have a finite sample $X$ with $n$ observations. In fundamental Statistics we defined ubiasedness this way.

**Definition 2.1.3.** *The statistic T(X) is an unbiased estimator of the $g(\theta)$ function if and only $\mathbb{E}_\theta[T(X)] = g(\theta)$ for all $\theta \in \Theta$.*

Here, $X$ is the sample and $T$ is a function on the sample space. The LS estimate can be seen as a function on the sample space which estimates the parameter $\theta^*$. For LS:

$$X = \{Y_t, \varphi_t\},$$
$$T(X) = (\Phi^T\Phi)^{-1}\Phi^T Y \text{ and } g = id.$$

**A 2.** *: The noise has a zero mean, $\mathbb{E}(N_t) = 0$ for all $t = 1, \ldots, n$.*

**Theorem 2.1.4.** *Assuming A1 and A2 the least squares estimator is unbiased.*

*Proof.*

$$\mathbb{E}[\widehat{\theta}_n] = \mathbb{E}[(\Phi^T\Phi)^{-1}\Phi^T Y] = \mathbb{E}[(\Phi^T\Phi)^{-1}\Phi^T(\Phi\theta^* + N)] = \theta^* + (\Phi^T\Phi)^{-1}\Phi^T E[N] \tag{2.22}$$

We know that the noise has a zero mean. It gives us the following:

$$E[\widehat{\theta}_{LS}] = \theta^*$$

which proves that the LS estimate is unbiased.

□

### Gauss-Markov theorem

The least squares estimator is the best linear unbiased estimator (BLUE). Best means that the least squares has the "least" variance out of the linear unbiased estimators. Neither linearity nor unbiasedness can be left out. In general, variance is a positive semi-definite matrix also called covariance matrix for each random vector. In this scenario least variance means that its covariance matrix is the least in the Loewner partial order which is a partial order defined on the Hermitien matrices.

**Definition 2.1.5.** *Let $A$ and $B$ be two symmetric matrices of order $n$. $B \preccurlyeq A$ if and only if $A - B$ is a positive semi-definite matrix.*

Using this ordering we can state the Gauss-Markov theorem. Assume that for a linear system the followings are true:

**A 3.** *The noises are uncorrelated.* $\mathbb{E}(NN^T) = \sigma^2 I$.

**A 4.** *The noises are homoscedastics and they have the same finite variance,* $var(N_i) = \sigma^2 < \infty \; \forall i = 1, \ldots, n$.

**Theorem 2.1.6** (Gauss-Markov). *[16] If A1, A2, A3 and A4 are valid, then the least squares estimate has the least variance in the Loewner partial order out of the linear unbiased estimators.*

*Proof.* Suppose we have an unbiased estimate $\widetilde{\theta}$ other than LS which is linear in $Y$.

$$\widetilde{\theta} = CY \tag{2.23}$$

It is easy to see that $C$ can be rewritten as

$$C = (\Phi^T \Phi)^{-1} \Phi^T + D \tag{2.24}$$

Knowing that $C$ is unbiased gives us the following.

$$\mathbb{E}(\widetilde{\theta}) = \mathbb{E}[((\Phi^T \Phi)^{-1} \Phi^T + D)(\Phi\theta^* + N)] = \theta^* + D\Phi\theta^* \Rightarrow D\Phi = 0 \tag{2.25}$$

The covariance matrix of $\theta^*$ is going to be:

$$\mathbb{E}[(\theta^* - \widetilde{\theta})(\theta^* - \widetilde{\theta})^T] = \mathbb{E}\Big[[(\Phi^T\Phi)^{-1}\Phi^T + D]NN^T[\Phi(\Phi^T\Phi)^{-1} + D^T]\Big]. \tag{2.26}$$

Using that $\sigma^2 I = \mathbb{E}(NN^T) \in \mathbb{R}^{n\times n}$ and the linearity of expectation

$$\mathbb{E}[(\theta^* - \widetilde{\theta})(\theta^* - \widetilde{\theta})^T] = [(\Phi^T\Phi)^{-1}\Phi^T + D]\mathbb{E}(NN^T)[\Phi(\Phi^T\Phi)^{-1} + D^T] \tag{2.27}$$

$$= [(\Phi^T\Phi)^{-1}\Phi^T + D]\sigma^2 I[\Phi(\Phi^T\Phi)^{-1} + D^T] = \sigma^2[(\Phi^T\Phi)^{-1}\Phi^T + D]I[\Phi(\Phi^T\Phi)^{-1} + D^T] \tag{2.28}$$

$$= \sigma^2(\Phi^T\Phi)^{-1} + \sigma^2(\Phi^T\Phi)^{-1}\Phi^T D^T + \sigma^2 D\Phi(\Phi^T\Phi)^{-1} + \sigma^2 DD^T. \tag{2.29}$$

Since $D\Phi = 0$

$$E[(\theta^* - \widetilde{\theta})(\theta^* - \widetilde{\theta})^T] = \sigma^2(\Phi^T\Phi)^{-1} + \sigma^2 DD^T. \tag{2.30}$$

It is clear that $DD^T$ is a positive semi-definite matrix. Consequently the covariance matrix of $\widetilde{\theta}$ equals to the covariance matrix of $\widehat{\theta}_n$ plus a positive semidefinite matrix. Hence $\widehat{\theta}_n$ has a smaller variance using the Loewner partial order relative to any other linear unbiased estimator. □

**Efficiency in case of Gaussian noise**

The Gauss-Markov theorem showed us that the least squares estimator is the best linear unbiased operator. With more assumptions additional theorems can be proven.

Efficiency can be another good property of an estimator. It measures the quality of an estimator. A more efficient estimator needs less observation or data to produce a given performance. We use the Fisher information matrix to measure the performance. It was named after Sir Ronald Fisher British statistician who played a leading role in the foundation of modern statistical science. The Fisher information can be defined this way.

**Definition 2.1.7.** *Assume that $\Theta \subseteq \mathbb{R}^p$ is an open (usually convex) set of parameters. Let $l(\theta) \triangleq \log(f(X;\theta))$ be the natural logarithm of the likelihood function of $X$. Suppose that the function $\theta \mapsto \log f(X;\theta)$ is differentiable, then $\frac{\partial l(\theta)}{\partial \theta}$ is a $p$ dimensional column vector. Simplifying the notation let $\partial l(\theta)$ be $\frac{\partial l(\theta)}{\partial \theta}$. The Fisher information is defined as the*

*following symmetric, positive semidefinite matrix:*

$$I(\theta) = \mathbb{E}[\partial l(\theta) \partial l(\theta)^T] \tag{2.31}$$

Suppose we have an unknown parameter $\theta^*$ which is to be estimated with an unbiased estimator $\widehat{\theta}$. Then, the variance of $\widehat{\theta}$ is bounded by the inverse of the Fisher information matrix of the true parameter.

$$\operatorname{cov}(\widehat{\theta}) \succeq \mathrm{I}^{-1}(\theta^*) \tag{2.32}$$

This bound is called the Cramér-Rao lower bound. It was discovered by Harald Cramér [9] and Calyampudi Radhakrishna Rao [23] independently in the 1940*s*. Now let $e(\theta)$ denote $\operatorname{cov}(\widehat{\theta})^{-1}\mathrm{I}^{-1}(\theta^*)$. Consequently the following is true

$$e(\theta) \preceq I.$$

Using this notation we can define efficient estimators.

**Definition 2.1.8.** *An unbiased estimator is efficient if* $e(\theta) = 1$ *for all* $\theta \in \Theta$.

Equivalently an estimator is efficient if $\operatorname{cov}(\widehat{\theta})$ is equal to the Cramér-Rao lower bound.

Suppose that the noise has a normal distribution. In this case it can be proven that the least squares estimator is efficient. It means that the LS estimator covariance matrix reaches the Cramér-Rao lower bound.

Note that this assumption is not always met. Usually we do not have this information about the noise in our hand. Still this result is important. Proof can be found in Chingnun Lee's notes about linear regression models [16]. It is important to see that in case of Gaussian noise the LS estimator is the same as the maximum likelihood estimator, which is defined as

$$\widehat{\theta}_{ML} \triangleq \operatorname*{argmax}_{\theta} f_\theta(X) = \operatorname*{argmax}_{\theta} L(\theta) \tag{2.33}$$

where $f$ is the density function of the sample and $L$ is the log-likelihood function. Note that the maximum likelihood estimator does not always exist or is not always unique, but in this special case it can be showed that it is equal to the LS.

Now I am going to list some of the most important asymptotical properties of the least squares estimator. These are extremely useful and give us a reasonable approach to the construction of confidence regions.

### Consistency

We are going to investigate the properties of the least squares as $n$, the "sample size", tends to infinity.

**Definition 2.1.9.** *The $\{\widehat{\theta}_n\}$ estimate sequence is consistent if $\widehat{\theta}_n \to \theta^*$ in probability as $n \to \infty$.*

**Definition 2.1.10.** *The $\{\widehat{\theta}_n\}$ estimate sequence is strongly consistent if $\widehat{\theta}_n \to \theta^*$ almost surely as $n \to \infty$.*

**A 5.** $\lim_{n\to\infty}(\frac{\Phi_n^T\Phi_n}{n}) = Q$ *where $Q$ is finite and nonsingular.*

**Theorem 2.1.11.** *Assume that A1 A2, A3, A4 and A5 hold then the least squares estimator is consistent. [16]*

$$\widehat{\theta}_n \xrightarrow{p} \theta^* \tag{2.34}$$

*Proof.*

$$\widehat{\theta}_n = (\Phi_n^T\Phi_n)^{-1}\Phi_n^T Y = \theta^* + (\Phi_n^T\Phi_n)^{-1}\Phi_n^T N = \theta^* + \left(\frac{\Phi_n^T\Phi_n}{n}\right)^{-1}\left(\frac{\Phi_n^T N}{n}\right) \tag{2.35}$$

We know that:

$$\mathbb{E}\left(\frac{(\Phi_n^T N)}{n}\right) = 0$$

and

$$\mathbb{E}\left[\frac{(\Phi_n^T N)}{n}\frac{(\Phi_n^T N)^T}{n}\right] = \frac{\sigma^2}{n}Q$$

therefore

$$\lim_{n\to\infty}\mathbb{E}\left(\frac{(\Phi_n^T N)}{n}\right) = \lim_{n\to\infty}\frac{\sigma^2}{n}Q = 0. \tag{2.36}$$

Consequently

$$\tfrac{(\Phi_n^T N)}{n} \xrightarrow{p} 0 \Rightarrow \widehat{\theta}_n \xrightarrow{p} \theta^*.$$

□

Note that it can be shown that $\lim_{n\to\infty}\widehat{\theta}_n = \theta^*$ almost surely as well. It means that the strong consistency is also true as $n$ tends to infinity under certain circumstances. See [8] for more information on this topic.

### 2.1.3 Asymptotical Gaussianity

**Theorem 2.1.12.** *[16] Under the conditions of A1, A2, A3, A4 and A5 the least squares estimate is asymptotically Gaussian.*

If we assume that $\Phi^T\Phi$ is $O(n)$ then $lim_{n\to\infty}(\Phi^T\Phi)^{-1} = 0$. We know that $\sqrt{n}(\widehat{\theta}_n - \theta) \sim N(0, \sigma^2(\Phi^T\Phi/n)^{-1})$ for each $n$. Hence $\sqrt{n}(\widehat{\theta}_n - \theta) \to N(0, \sigma^2 Q^{-1})$ in distribution.

### 2.1.4 Asymptotical efficiency

LS is an efficient estimator for Gaussian noise. Its information matrix reaches the Cramér-Rao lower bound for all $n$. Consequently the Fisher information matrix reaches the Cramér-Rao bound asymptotically too. Therefore LS is asymptotically efficient. It also means that LS is a minimume-variance unbiased estimate (MVUI) asymptotically.

### 2.1.5 Confidence ellipsoids

Using a point estimator gives us one specific $\widehat{\theta}$ from the parameter space as an estimator. This parameter can have great properties such as unbiasedness, consistency or efficiency just like the least squares estimate had, but typically it is going to be equal to the true parameter $\theta^*$ with zero probability. It is often useful to find a set of parameters which contains the true parameter with at least a user-chosen probability $q$. In many cases, we need such guarantee to prove that our calculation is correct for the purpose of safety, quality or punctuality. It is critical for robust method and risk management.

**Definition 2.1.13.** *A $\Theta$ set of parameters is a confidence region if $\mathbb{P}(\theta^* \in \Theta) \geq q$.*

Note that here the $q$ probability is just a lower bound therefore these regions can contain more parameters than necessary. For example the entire parameter space is a confidence region for every $q \in [0, 1]$ because it contains the true parameter with exactly 1 probability which is definitely more than $q$. Nevertheless this confidence region is useless. It does not give us any new information about the parameters.

Sometimes it is possible to reach the lower bound and find exact confidence regions which leads us to the following definition.

**Definition 2.1.14.** *A $\Theta$ set of parameters is an exact confidence region if $\mathbb{P}(\theta^* \in \Theta) = q$.*

Finding confidence regions is a fundamental problem in Statistics. Many different algorithms have been developed for such aims. There are many different methods with a variety of assumptions and approximations.

Now, I am going to show a standard example for a non-exact asymptotic confidence region. This approach comes from the central limit theorem which explains why the regions are asymptotic. Even though we cannot construct exact regions with this method and we need many observations there is a variety of applications of this method in different fields of Science such as Economy, Social Sciences and Biology.

Knowing that under some moment conditions the LS estimate is asymptotically Gaussian gives us a great approach to the construction of a confidence region around the LS estimate.

Let $\widehat{\theta}_n$ denote the LS estimate based on $n$ data points. Then, according to asymptotic Gaussianity $\sqrt{n}(\widehat{\theta}_n - \theta^*) \longrightarrow N(0, \sigma^2 Q^{-1})$ in distribution, where $N(0, \sigma^2 Q^{-1})$ is the normal distribution with zero mean and covariance matrix $\sigma^2 Q^{-1}$, $\sigma^2$ is the variance of the noise and $Q = \lim_{n\to\infty}(\frac{1}{n}\Phi_n^T \Phi_n)$ which is finite and nonsingular as stated in A5. Reasonably we are going to approximate the quantity of $\sqrt{n}(\widehat{\theta}_n - \theta^*)$.

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \sim N(0, \sigma^2 Q^{-1}) \tag{2.37}$$

Note that usually we do not know $\sigma^2$ exactly since it depends on the noise which is unknown. Most of the time we need to estimate $\sigma^2$. For this aim we use $\widehat{\sigma}_n^2$ which is an unbiased estimator of $\sigma^2$.

$$\widehat{\sigma}_n^2 = \frac{1}{n-d}\sum_{t=1}^{n}(Y_t - \varphi^T \widehat{\theta}_n)^2 \tag{2.38}$$

Now we are able to estimate the previous quantity in (2.37). Obtaining (approximately)

$$\frac{1}{\widehat{\sigma}_n}(\frac{1}{n}\Phi_n^T \Phi_n)^{\frac{1}{2}}\sqrt{n}(\widehat{\theta}_n - \theta^*) \sim N(0, \mathbb{I}). \tag{2.39}$$

Consequently (approximately)

$$\frac{n}{\widehat{\sigma}_n^2}(\widehat{\theta}_n - \theta^*)(\Phi_n^T \Phi_n)(\widehat{\theta}_n - \theta^*) \sim \chi^2(d) \tag{2.40}$$

where $\chi^2(d)$ denotes the $\chi^2$ distribution with $d$ degrees of freedom. Then we can build an

approximate confidence region by

$$\Theta_n^q = \{\theta \in \mathbb{R}^d : (\widehat{\theta}_n - \theta)(\Phi_n^T \Phi_n)(\widehat{\theta}_n - \theta) \leq \mu \frac{\widehat{\sigma}_n^2}{n}\} \tag{2.41}$$

where $q$ is user-chosen probability level. We can find $\mu$ by solving $q = F_{\chi^2(d)}(\mu)$ where $F$ is the cumulative distribution function of the $d$ dimensional $\chi^2$ distribution. Then

$$\mathbb{P}(\theta^* \in \Theta_n^q) \approx q.$$

Note that this is a heuristic method which often gives us inaccurate regions for small samples.

Sometimes we cannot repeat the experiment or measurement as many times as it would be necessary to build a great confidence region therefore it is reasonable to find a non-asymptotic method which constructs regions with good properties even for small number of data points. In the next section I am going to introduce an algorithm which gives us exact, finite sample confidence regions.

## 2.2 Non-asymptotic confidence region for the LS

Finding the true paramter can be very hard because of the lack of informations. Our knowlegde about the noise is usually limited and the number of observations is also finite. I am going to introduce you a method which proposes a solution with these restrictions. It was originally developed by Balázs Csanád Csáji, Marco C. Campi and Erik Weyer in 2012 [10]. This method is called the Sign-Perturbed Sums (SPS). It constructs exact confidence regions under mild statistical assumptions around the least squares estimate. As we are going to see compared to the asymptotic confidence regions these regions are going to be exact. The main assumption that we have to make is that the noises are symmetrically distributed around zero. The method exploits this symmetry as much as possible. This method can be used with small changes for many different systems. It is also possible to build confidence regions around estimators other than the LS. The SPS for the least absolute deviations (LAD) estimate is going to be presented in the next chapter.

### 2.2.1 The Sign-Perturbed Sums method

Consider a linear system. Assume we have $n$ observations like we had in the previous section:

$$\{(Y_1, \varphi_1), (Y_2, \varphi_2), ..., (Y_n, \varphi_n)\}. \tag{2.42}$$

Suppose that the explanatory variables are deterministic here. There is neither autoregression nor correlation between the inputs. This simple system can be for example a finite impulse system with order $d$ ($FIR(d)$) where the dimension of the regressors is the order of the FIR system. Nevertheless we assume that each output was processed this way

$$Y_t = \varphi_t^T \theta^* + N_t \tag{2.43}$$

The notation is the same as it was in the previous sections.
As I mentioned there are only a few mild assumptions.

**B 1.** $(N_t)_t$ *is an independent random noise sequence (not necessary identically distributed) and each* $N_t$ *is symmetrically distributed around zero.*

**B 2.** $\det(R_n) \neq 0$ *where* $R_n = \frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^T$

Note that these assumptions are really mild. The noise does not need to have special moment conditions. We do not assume stationarity for the noise terms nor a specific distribution. Though independence is the strongest assumption with small changes in the SPS this condition can be relaxed. Practically many of the standard distributions satisfy the assumption about the symmetricity (Gauss, Lagrange, Cauchy etc.).

Our goal is to construct an exact confidence region with user-chosen probability. The idea is going to lean on the symmetricity. Now let

$$S_0(\theta) \triangleq \frac{1}{n}(R_n)^{-1/2} \sum_{t=1}^{n} \varphi_t (Y_t - \varphi_t^T \theta) \tag{2.44}$$

be the reference sum and the sign-perturbed sums which are indicated in the name of the method are

$$S_i(\theta) \triangleq \frac{1}{n}(R_n)^{-1/2} \sum_{t=1}^{d} \alpha_{i,t} \varphi_t (Y_t - \varphi_t^T \theta) \tag{2.45}$$

for $i = 1, \ldots, m-1$, where $\{\alpha_{i,t}\}$ are i.i.d. random signs meaning $\alpha_{i,t} = \pm 1$ with $1/2$ probability for each $i$ and $t$. Note that $R_n^{-1/2}$ is not necessary but it helps to shape the region and $1/n$ is introduced only for numerical stability.

The intuitive idea is that when we get far enough from the true parameter $\|S_0(\theta)\|^2$ increases faster than $\{\|S_i(\theta)\|^2\}$ so eventually it will dominate the ordering. Formally when $\|\tilde{\theta}\|$ is large enough, where $\tilde{\theta} \triangleq \theta^* - \theta$

$$\| \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \sum_{t=1}^{n} \varphi_t N_t \|^2_{R_n^{-1}} > \| \sum_{t=1}^{n} \alpha_{i,t} \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \sum_{t=1}^{n} \alpha_{i,t} \varphi_t N_t \|^2_{R_n^{-1}} \tag{2.46}$$

with high probability. In fact $\sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta}$ increases faster than $\sum_{t=1}^{n} \alpha_{i,t} \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta}$. The formal proof of this claim relies on the Schur complement argument.

If we calculate these equations for the true parameter ($\theta^*$) we get the following

$$S_0(\theta^*) = \frac{1}{n}(R_n)^{\frac{1}{2}} \sum_{t=1}^{d} \varphi_t N_t$$

$$S_i(\theta^*) = \frac{1}{n}(R_n)^{\frac{1}{2}} \sum_{t=1}^{d} \alpha_{i,t} \varphi_t N_t \tag{2.47}$$

where $i = 1, \ldots, m-1$. The main idea is that since $N_t$ has a symmetric distribution around zero, $N_t$ and $\alpha_{i,t}N_t$ have the same distribution so $S_0(\theta^*)$ and $S_i(\theta^*)$ also have the same distribution for all $i \in \{0, \ldots, m-1\}$. Therefore if we somehow sort the $S_i$ vectors for $i = 0, \ldots, m-1$ there is no reason why $S_0$ should be greater or smaller than any other $S_i$. Proof can be found in [11]. To sort the vectors we are going to use the 2-norm. Note that we can use any norm to construct regions though some norms result smaller regions than others, there is no significant difference in the size.

Consequently $S_0$ is going to be in each place in the order with $\frac{1}{m}$ probability so $S_0$ is going to be among the first $q$ with $\frac{q}{m}$ probability. We are going to overview the property of exactness more precisely later.

### 2.2.2 The algorithm

The SPS algorithm is going to test each parameter whether it is in the confidence region or not. It has two parts. The first part is the initialization. Here we have to set the hyperparameters. The confidence region is going to contain the true parameter with a user-chosen probability so we need to set $p$ first. For any chosen rational $p$ we can find $q \in \mathbb{N}$ and $m \in \mathbb{N}$ such that $p = 1 - \frac{q}{m}$. As we can see there are many different usable (q,m) pairs. Though each pair may construct different region the difference in shape and size is not significant. During the initialization process we calculate the shaping matrix, $R^{1/2} = \frac{1}{n}(R_n)^{1/2}$. It exists because of A2 and used only for shaping the region and numerical stability. There are many methods to find $R^{1/2}$. One is the Cholesky factorization, which I used during the simulation process (see Chapter 4). We also have to generate $n(m-1)$ random signs $\alpha_{i,t}$ such that $\mathbb{P}(\alpha_{i,t} = \pm 1) = \frac{1}{2}$. These random signs are fixed for each parameter. We also need to provide a strict total order on the sign perturbed sums. Vector norms can be equal. For this reason we use a random permutation $\pi$ which chooses between the sums in case of tie. So the strict total order on the sign perturbed sums is defined as

$$S_k \succ_\pi S_j \qquad \text{if and only if}$$

$$\left( \|S_k\|^2 > \|S_j\|^2 \right) \text{ or } \left( \|S_k\|^2 = \|S_j\|^2 \text{ and } \pi(k) > \pi(j) \right).$$

The pseudocode for the initialization is given in Table 2.1.

| Pseudocode: SPS-Initialization |
|---|
| 1. Given a (rational) confidence probability $p \in (0,1)$, set integers $m > q > 0$ such that $p = 1 - q/m$; |
| 2. Calculate the outer product $$R_n \triangleq \frac{1}{n}\sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}},$$ and find a factor $R_n^{1/2}$ such that $$R_n^{1/2} R_n^{(1/2)\mathrm{T}} = R_n;$$ |
| 3. Generate $n\,(m-1)$ i.i.d. random signs $\{\alpha_{i,t}\}$ with $$\mathbb{P}(\alpha_{i,t} = 1) = \mathbb{P}(\alpha_{i,t} = -1) = \tfrac{1}{2},$$ for $i \in \{1, \ldots, m-1\}$ and $t \in \{1, \ldots, n\}$; |
| 4. Generate a random permutation $\pi$ of the set $\{0, \ldots, m-1\}$, where each of the $m!$ possible permutations has the same probability $1/(m!)$ to be selected. |

Table 2.1

Note that $\pi$ is a bijection from $\{0, \ldots, m-1\}$ to itself, thus for $k \neq j$, $\pi(k) \neq \pi(j)$.

After the initialization process the indicator function can test whether a parameter is included in the confidence region or not. This function is given in the Table 2.2. For any given parameter first we calculate the predicted residuals, then we evaluate the $m-1$ sign-perturbed sums and the reference sum. In order to sort the vectors we use the the total order that was defined earlier. Finally we compute the rank of the value corresponding to the reference sum in the order. If the rank is $q$ then the parameter is in the confidence region with confidence level of $1 - \frac{q}{m}$. In other words the indicator function is going to return with the value of 1 if $\|S_0(\theta)\|^2$ is not among the $q$ largest in the strict total order or else it will exclude the parameter and return with 0 value.

Using this function now we can express the confidence region we found as

$$\widehat{\Theta}_n = \{\theta \in \Theta : SPS - indicator(\theta) = 1\} \tag{2.48}$$

In the next section we are going to see that this confidence region is exact. Moreover, since $S_0(\widehat{\theta}_n) = 0$ if the region is non-empty the LS is always included and it is in the center of

| PSEUDOCODE: SPS-INDICATOR ( $\theta$ ) |
|---|
| 1. For the given $\theta$, compute the prediction errors for $t \in \{1, \dots, n\}$ $$N_t(\theta) \triangleq Y_t - \varphi_t^{\mathrm{T}}\theta;$$ |
| 2. Evaluate $$S_0(\theta) \triangleq R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t N_t(\theta),$$ $$S_i(\theta) \triangleq R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \alpha_{i,t}\, \varphi_t N_t(\theta),$$ for $i \in \{1, \dots, m-1\}$; |
| 3. Order scalars $\{\|S_i(\theta)\|^2\}$ according to $\succ_\pi$; |
| 4. Compute the rank $\mathcal{R}(\theta)$ of $\|S_0(\theta)\|^2$ in the ordering, where $\mathcal{R}(\theta) = 1$ if $\|S_0(\theta)\|^2$ is the smallest in the ordering, $\mathcal{R}(\theta) = 2$ if $\|S_0(\theta)\|^2$ is the second smallest, and so on. |
| 6. Return 1 if $\mathcal{R}(\theta) \leq m - q$, otherwise return 0. |

Table 2.2

the region. It can be shown that this estimator is star convex with the center of LS and also can be proven that it is strongly consistent.

However we can only evaluate the indicator function for finitely many data points we will see how we are able to represent these regions nicely using its good properties.

### 2.2.3 Exact confidence regions

The most important property of SPS is that it constructs a confidence region with exact user-chosen probability for finitely many data points meaning that $\widehat{\Theta}_n$ is an exact confidence region.

**Theorem 2.2.1.** *[11] Assuming B1 and B2 the confidence probability of the constructed confidence region is exactly p,*

$$\mathbb{P}(\theta^* \in \widehat{\Theta}_n) = 1 - \frac{q}{m}. \tag{2.49}$$

A formal proof was presented in [11]. It is based on the perception that $\{||S_i(\theta^*)||^2\}_{i=1}^{m-1}$ are uniformly ordered.

**Definition 2.2.2.** *Let $Z_1, Z_2, \ldots, Z_k$ be finitely many random variables and $\succ$ is a strict order on them. Suppose for all permutations $i_1, i_2, \ldots, i_k$ of indices $1, \ldots, k$ we have*

$$\mathbb{P}(Z_{i_k} \succ Z_{i_{k-1}} \succ \cdots \succ Z_{i_1}) = \frac{1}{k!}. \tag{2.50}$$

*Then we call $\{Z_i\}$ uniformly ordered with respect to order $\succ$.*

If $\{\|S_i(\theta^*)\|^2\}_{i=0}^{m-1}$ is uniformly ordered, then $\|S_0(\theta^*)\|$ is included in the first $q$ elements in the order with exactly $\frac{q}{m}$ probability, which justifies the theorem.

The proof of the claim that $\{\|S_i(\theta^*)\|^2\}_{i=0}^{m-1}$ are uniformly ordered relies on three lemmas. The proof of these lemmas are rather technical than complicated therefore not all details are presented here.

**Lemma 1.** *Let $\alpha, \beta_1, \ldots, \beta_k$ are i.i.d. random signs, then the random variables $\alpha, \alpha\beta_1, \ldots, \alpha\beta_k$ are also i.i.d. random signs.*

The formal proof uses that the original variables are independent and that $\alpha\beta_i$ and $\beta_i$ are identically distributed.

**Lemma 2.** *Let $X$ and $Y$ be two independent, $\mathbb{R}^d$-valued and $\mathbb{R}^k$-valued vector variable, respectively. Consider a measurable function $g : \mathbb{R}^d \times \mathbb{R}^k \mapsto \mathbb{R}$ and a Borel-set $A \subseteq \mathbb{R}$. If we have $\mathbb{P}(g(x, Y) \in A) = p$ for all (fixed) $x \in \mathbb{R}^d$, then $\mathbb{P}(g(X, Y) \in A) = p$ is also true.*

**Lemma 3.** *Let $Z_1, \ldots, Z_k$ be real-valued, i.i.d. random variables. Then, they are uniformly ordered w.r.t. $\succ_\pi$.*

A detailed argument about these can be found in [11] Appendix A. Using these lemmas we are able to prove the theorem (2.49).

*Proof.* As it was said we would like to verify that $\{\|S_i(\theta^*)\|^2\}_{i=0}^{m-1}$ are uniformly ordered.

Notice, that for $\theta = \theta^*$ we can express all $S_i(.)$ function as

$$S_i(\theta^*) = R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \alpha_{i,t} \varphi_t N_t \tag{2.51}$$

for all $i = 0, \ldots, m-1$, where $\alpha_{0,t} = 1$ for all $t \in \{1, \ldots, n\}$. We can see that all $S_i(.)$ functions depend on the perturbed noise sequence $\{\alpha_{i,t} N_t\}_{t=1}^n$ via the same measurable function. We denote this function by $S(\alpha_{1,1}N_1, \ldots, \alpha_{i,n}N_n) \triangleq S_i(\theta^*)$.

Since each $N_t$ is symmetric $\text{sign}(N_t)$ and $|N_t|$ are independent. Now we introduce

$$\gamma_{i,t} \triangleq \alpha_{i,t} \,\text{sign}(N_t). \tag{2.52}$$

We can use lemma 1 because $\alpha_{i,t}$ are i.i.d. random signs independent of $\text{sign}(N_t)$ so $\{\gamma_{i,t}\}_{t=1}^n$ are not only independent of $\{|N_t|\}_{t=1}^n$ but also i.i.d. random signs.

Now look at one constant realization of $\{|N_t|\}_{t=1}^n$, called $\{v_t\}_{t=1}^n$. We define the real valued variables $Z_i \triangleq \|S(\gamma_{i,1}v_1, \ldots, \gamma_{i,n}v_n)\|^2$. We apply the same (measurable) function to each element of an i.i.d. sample therefore the result we get are going to be i.i.d. as well, so $\{Z_i\}$ are i.i.d. random variables. Hence, lemma 3 can be applied. Consequently $\{Z_i\}$ are uniformly ordered.

We proved that for a fixed realization of $|N_t|$ the uniform ordering property is achieved. Let $\{|N_t|\}$ be $X$, $\{\gamma_{i,t}\}$ be $Y$ and $\|S(.)\|^2$ is $g$. Applying lemma 2 we see that the probabilities are independent of the particular realization of $\{|N_t|\}$ so we obtain the unconditional uniform ordering property for $\{\|S_i(\theta^*)\|^2\}$, from which the theorem follows. □

### 2.2.4 Star convexity

We said before that the SPS constructs the confidence region around the LS estimator. Now we are going to interpret this claim more precisely. It has been seen that the LS is in the confidence region unless it is empty coming from the fact that $\|S_0(\widehat{\theta}_n)\| = 0$. To punctuate our claim that LS is in the center of the confidence region let's recall the definition of star convexity.

**Definition 2.2.3.** *$S \subseteq \mathbb{R}^n$ is star convex if and only there exists an $x \in S$ called the star center that for all $\alpha \in [0,1]$ and for all $y \in S$ it is true that $x\alpha + (1-\alpha)y$ is an element of $S$.*

It is easy to see that all convex sets are star convex even though the controverse is not true. Now the following theorem holds.

**Theorem 2.2.4.** *If B1 and B2 are true then the confidence region built by SPS is star convex with the LS estimate as a star center or empty.*

Convexity is not necessarely held. For example for $q = 1$ the constructed region is the union of ellipsoids, which are not usually convex. A formal proof of this theorem can be found in [11]. It is not so simple. the proof relies on the Schur complement argument. The

main idea is to express $\widehat{\Theta}_n$ as unions and intersections of star convex sets having $\widehat{\theta}_n$ as a common star center. Also notice that the region can be empty with small probability. It can happen that the generated signs are all one vectors so accidentally the $\widehat{\theta}_n$ is excluded during the process, but the probability of this event is really small and decreases with an exponantial rate. It can be showed that if we do not allow two identical random sign vectors the method still works, though in practise this is not a problem. The property of star convexity gives us a good approach to find the boundary of the confidence region, which helps us representing these regions in lower dimensions.

### 2.2.5 Ellipsoidal outer approximation

Now we would like to represent the confidence region that is constructed by the SPS. As we can see in the method, for any given parameter it is easy to find out if it is in the region or not. We have to calculate $\{\|S_i(\theta)\|^2\}$ for all $i = 0, \dots, m-1$ for that parameter and compare them. It is possible to check each parameter in a grid and test if that is included. However, it is computationally demanding and for higher dimensions representation remains a big issue as always. Another problem is that we do not know how big is our region estimator. There has not been anything so far that guarantees that the region is not going to be too big or too small. Finding a great grid and interval is also an arising question. Although the LS estimator gives us a great starting point because it should be the center of the grid (see Section 2.2.4) the size and shape depend on a lot of factors such as the variance of the noise, the number of observations, which norm we use and even on the chosen $m$ and $q$ in the method. Some of these factors are unknown like the variance of the noise.

Although, using the property of star convexity gives us a better approach to representation, it is still going to be computationally demanding. We know that LS is the star center of the region (if it is non-empty) therefore it is possible to find the boundaries starting to test parameters from the LS in each direction. Using binary search algorithm is a relatively fast way to find the boundary starting from the LS estimator in any direction with a user-chosen accuracy. The problem is that in the parameter space usually there are infinitely many directions. Hence compact representation remains an issue.

As we could not find a truely compact representation for the confidence region, trying to find a good way of approximating the region is reasonable. Even though this way our confidence region is not going to be exact anymore a guaranteed probability is still

achievable meaning that the true parameter, $\theta^*$ is included in the confidence region with at least a user-chosen $p$ probability. An outer approximation method has been developed by Balázs Csáji, Erik Weyer and Marco Campi in [11]. This approximation is efficiently computed (in polynomial time) and can be represented very efficiently in a compact way.

Expanding $\|S_0(\theta)\|^2$ we can rewrite the reference sum as

$$\|S_0(\theta)\|^2 = \left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t(Y_t - \varphi_t^{\mathrm{T}}\theta)\right]^{\mathrm{T}} R_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t(Y_t - \varphi_t^{\mathrm{T}}\theta)\right] \tag{2.53}$$

$$= \left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}}(\theta - \hat{\theta}_n)\right]^{\mathrm{T}} R_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t\varphi_t^{\mathrm{T}}(\theta - \hat{\theta}_n)\right] \tag{2.54}$$

$$= (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n). \tag{2.55}$$

where $\hat{\theta}_n$ is the LS estimator. First, consider those parameters where there are at least $q$ $\|S_i(\theta)\|^2$ greater or equal to $\|S_0(\theta)\|^2$. Now we do not worry about the random ordering used in the SPS method. Our estimator is going to be an outer approximation so we may allow more parameters to be included in the region than we did before. First we look at this set:

$$\widehat{\Theta}_n \subseteq \left\{\, \theta \in \mathbb{R}^d \,:\, (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq r(\theta) \,\right\},$$

where $r(\theta)$ denotes the $q^{th}$ greatest value of $\|S_i(\theta)\|^2$.

Our goal is to find an upper bound instead of $r(\theta)$ independent of $\theta$. Let $r$ denote this fix boundary. Using this $r$ we can see that the region we get is going to be a similar ellipsoid as in the asymptotic theorem. At least the shape and the orientation of the confidence ellipsoid are going to be the same as it was there because those depend just on the matrix $R_n$. Even though this ellipsoid has a different size depending on $r$ it is a guaranteed confidence region for finitely many data points. In addition to that the values of $\hat{\theta}_n$, $R_n$ and $r$ give us a compact representation to this region.

### 2.2.6 Convex programming formulation

Looking for a fix $r$ as upper bound instead of $r(\theta)$ leads us to a convex optimization problem.

Comparing the reference sum to just a single sign-perturbed sum with index $i$ for each

parameter we can get this set of parameters:

$$\begin{aligned}&\{\,\theta : \|S_0(\theta)\|^2 \le \|S_i(\theta)\|^2\,\}\\ &\quad\subseteq \{\,\theta : \|S_0(\theta)\|^2 \le \max_{\theta:\|S_0(\theta)\|^2\le\|S_i(\theta)\|^2} \|S_i(\theta)\|^2\,\}.\end{aligned}$$

We can rewrite this relation as:

$$(\theta-\hat\theta_n)^{\mathrm{T}} R_n(\theta-\hat\theta_n) \le$$

$$\begin{aligned}&\left[\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t(Y_t-\varphi_t^{\mathrm{T}}\theta)\right]^{\mathrm{T}} R_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t(Y_t-\varphi_t^{\mathrm{T}}\theta)\right]\\ &\quad= \theta^{\mathrm{T}}Q_iR_n^{-1}Q_i\theta - 2\,\theta^{\mathrm{T}}Q_iR_n^{-1}\psi_i + \psi_i^{\mathrm{T}}R_n^{-1}\psi_i,\end{aligned}$$

where $Q_i \in \mathbb{R}^{d\times d}$ and $\psi_i \in \mathbb{R}^d$ are defined this way:

$$\begin{aligned}Q_i &\triangleq \frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t\varphi_t^{\mathrm{T}},\\ \psi_i &\triangleq \frac{1}{n}\sum_{t=1}^{n}\alpha_{i,t}\varphi_t Y_t.\end{aligned}$$

We know that

$$\max_{\theta:\|S_0(\theta)\|^2\le\|S_i(\theta)\|^2} \|S_i(\theta)\|^2 = \max_{\theta:\|S_0(\theta)\|^2\le\|S_i(\theta)\|^2} \|S_0(\theta)\|^2.$$

Let's use this notation: $z \triangleq R_n^{\frac{1}{2}\mathrm{T}}(\theta-\hat\theta_n)$. Now the quantity we seek

$$\max_{\theta:\|S_0(\theta)\|^2\le\|S_i(\theta)\|^2} \|S_0(\theta)\|^2$$

can be found as the solution of the following optimization problem:

$$\begin{aligned}\text{maximize}\quad & \|z\|^2\\ \text{subject to}\quad & z^{\mathrm{T}}A_iz + 2z^{\mathrm{T}}b_i + c_i \le 0,\end{aligned} \tag{2.56}$$

where $A_i$, $b_i$ and $c_i$ are defined as

$$\begin{aligned} A_i &\triangleq I - R_n^{-\frac{1}{2}} Q_i R_n^{-1} Q_i R_n^{-\frac{1}{2}\mathrm{T}}, \\ b_i &\triangleq R_n^{-\frac{1}{2}} Q_i R_n^{-1} (\psi_i - Q_i \hat{\theta}_n), \\ c_i &\triangleq -\psi_i^{\mathrm{T}} R_n^{-1} \psi_i + 2\hat{\theta}_n^{\mathrm{T}} Q_i R_n^{-1} \psi_i - \hat{\theta}_n^{\mathrm{T}} Q_i R_n^{-1} Q_i \hat{\theta}_n. \end{aligned}$$

This problem is not convex in general, but it can be shown that strong duality holds. Formal proof can be found in [11]. It means that the maximum we look for is equal to the optimum value of the dual problem, which can be expressed as

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & \lambda \geq 0 \\ & \begin{bmatrix} -I + \lambda A_i & \lambda b_i \\ \lambda b_i^{\mathrm{T}} & \lambda c_i + \gamma \end{bmatrix} \succeq 0, \end{aligned} \tag{2.57}$$

where "$\succeq 0$" denotes that a matrix is positive semidefinite. This problem is convex indeed, therefore it can be solved easily using Gurobi or other tools.
Let $\gamma_i^*$ be the value we get by solving the convex optimization problem. Now we know that

$$\{\theta : \|S_0(\theta)\|^2 \leq \|S_i(\theta)\|^2\} \subseteq \{\theta : \|S_0(\theta)\|^2 \leq \gamma_i^*\}.$$

Obtaining

$$\widehat{\Theta}_n \subseteq \widehat{\widehat{\Theta}}_n \triangleq \left\{ \theta \in \mathbb{R}^d \,:\, (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq r \right\},$$

where $r$ is the $q^{th}$ largest value of $\gamma_i^*, i = 1, \ldots, m-1$.
The outer approximation is $\widehat{\widehat{\Theta}}_n$. It contains more parameters than $\widehat{\Theta}_n$ so it is clear that it is a guaranteed confidence region, meaning

$$\mathbb{P}\big(\, \theta^* \in \widehat{\widehat{\Theta}}_n \,\big) \;\geq\; 1 - \frac{q}{m} \;=\; p,$$

for any finite $n$.
The pseudocode for computing $\widehat{\widehat{\Theta}}_n$ is given in Table 2.3.

| Pseudocode: SPS-Outer-Approximation |
|---|
| 1. Compute the least-squares estimate, $\hat{\theta}_n = R_n^{-1}\left[\frac{1}{n}\sum_{t=1}^{n}\varphi_t Y_t\right]$; |
| 2. For $i \in \{1,\dots,m-1\}$, solve the optimization problem (2.56), and let $\gamma_i^*$ be the optimal value; |
| 3. Let $r$ be the $q$th largest $\gamma_i^*$ value; |
| 4. The outer approximation of the SPS confidence region is given by the ellipsoid $\widehat{\widehat{\Theta}}_n = \left\{ \theta \in \mathbb{R}^d \,:\, (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \le r \right\}$. |

Table 2.3

### 2.2.7 Asymptotic properties

One of the most important properties of the SPS is that it is non-asymptotic. It gives us an estimate for finitely many data points. Nevertheless it would be great if SPS would construct a more accurate confidence region when we have more data points. In other words when we have more information about a system then we would like to get closer to the true parameter. In the next paragraphs we will see that SPS has nice asymptotical properties as well. In a way it is strongly consistent and its size and shape is similar to the asymptotic elipsoid as both $n$ and $m$ (the number of sums) goes to infinity.

#### Strong consistency

We are going to see that as the number of observations tends to infinity the size of the region gets smaller so the region shrinks around the true parameter and asymptotically all parameters are going to be excluded except $\theta^*$.
Strong consistency holds if we assume the following:

**B 3.** $\liminf_{n\to\infty} \lambda_{min}(R_n) = \widetilde{\lambda} > 0$ *where* $\lambda_{min}$ *is the minimum eigenvalue of a matrix,*

**B 4.** $\sum_{t=1}^{\infty} \frac{\|\varphi_t\|^2}{t^2} < \infty$ *(regressor growth rate restriction),*

**B 5.** $\sum_{t=1}^{\infty} \frac{\mathbb{E}(N_t^2)^2}{t^2} < \infty$ *(noise variance rate restriction).*

**Theorem 2.2.5.** *Assuming B1, B2, B3, B4 and B5* $\forall \varepsilon > 0$ *(almost surely)* $\exists \widetilde{N}$ *that* $\widehat{\Theta}_n \subseteq B_\varepsilon(\theta^*)$ $\forall n > \widetilde{N}$.

A formal proof of this theorem can be found here [25] in Appendix A. The proof relies on the strong Law of Kolmogorov, uses the assumptions and the Cauchy-Schwarz inequality. Note that in this theorem $\widetilde{N}$ is stochastic, it depends on the noise realization.

The noise can be non-stationary and the variance of the noise may itself tend to infinity, though the growth rate cannot be so big (see B6). Even the regressors themselves can tend to infinity though there is restriction for the rate as well (see B5).

### Asymptotic shape

The shape and size are comparable to the asymptotic standard ellipsoids as we assume that

**B 6.** $\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{\infty} \|\varphi_t\|^4 < \infty$ *(regressor growth rate restriction)*

**B 7.** $N_i$ *is i.i.d. with* $\mathbb{E}[N_t] = \sigma^2$ *and* $\mathbb{E}[N_t^4] < \infty$

To claim the theorem we need to define the relaxed asymptotic ellipsoids, which are:

$$\widetilde{\Theta}_n(\varepsilon) \triangleq \left\{ \theta : (\theta - \widehat{\theta}_n)^T R_n (\theta - \widehat{\theta}_n) \leqq \frac{\mu\sigma^2 + \varepsilon}{n} \right\}.$$

where $\varepsilon > 0$. We use $\widehat{\Theta}_{n,m}$ which refers to the confidence region for $n$ observations and $m-1$ sign-perturbed sums. As it was mentioned earlier both $n$ and $m$ goes to infinity. Let $q_m \triangleq \lfloor (1-p)m \rfloor$ so that the probability that $\widehat{\Theta}_{n,m}$ contains the true parameter is $p_m \triangleq 1 - \frac{q_m}{m}$. We know that $p_m \to p$ as $m \to \infty$, it comes from the construction.

**Theorem 2.2.6.** *Assume B1, B2, B3, B6 and B7. then, there exists a doubly indexed set of random variables* $\{\varepsilon_{n,m}\}$ *such that* $\lim_{n\to\infty} \lim_{m\to\infty} \varepsilon_{n,m} = 0$ *a.s., and*

$$\widehat{\Theta}_{n,m} \subseteq \widetilde{\Theta}_n(\varepsilon_{n,m})$$

The formal proof can be found in [25]. In the previous section we saw that LS is the Best Linear Unbiased Estimator (BLUE), more preciesely the Gauss-Markov theorem holds. This theorem shows us that in the long run $\widehat{\Theta}_{n,m}$ is almost surely contained in the standard asymptotic ellipsoid for the LS estimate, though noise variance can increase by an asymptotically vanishing margin.

## 2.3 SPS for ARX system

After dealing with a simple linear model we are going to consider an auto regressive system with exogenous input also called the ARX model. In this model the explanatory variables are previous outcomes and independent exogenous inputs, so the following holds

$$Y_t + \sum_{j=1}^{n_a} a_j^* Y_{t-j} \triangleq \sum_{j=1}^{n_b} b_j^* U_{t-j} + N_t. \tag{2.58}$$

It can be reformulated into linear regression form as

$$Y_t \triangleq -\sum_{j=1}^{n_a} a_j^* Y_{t-j} + \sum_{j=1}^{n_b} b_j^* U_{t-j} + N_t = \varphi\theta^* + N_t \tag{2.59}$$

where $Y_t$ is the $t^{th}$ outcome, $n_a$ and $n_b$ are known orders of the ARX system, and

$$\varphi_t \triangleq [-Y_{t-1}, \ldots, -Y_{t-n_a}, U_{t-1}, \ldots, U_{t-n_b}]^T \tag{2.60}$$

$$\theta^* \triangleq [-a_1^*, \ldots, -a_{n_a}^*, b_1, \ldots, b_{n_b}^*]. \tag{2.61}$$

The presence of past outputs makes it harder to apply SPS on this system. The standard SPS cannot be used directly, because the previous outcomes are not independent of the noise. More precisely the $\sum_{t=2}^{n} -Y_{t-1}N_t$ respective to $S_i(\theta^*)$ does not have the same distribution as $\sum_{t=2}^{n} -Y_{t-1}\alpha_{i,t}N_t$, since $Y_{t-1}$ depends on $N_{t-2}$,$N_{t-3}$ and so on. Therefore we would not get an exact region in this case. Our goal in this section is to extend the original SPS to the ARX model. This extended method was introduced in [10].

We are going to use alternative $\bar{Y}_t$ such that $\sum_{t=2}^{n} -\bar{Y}_{t-1}\alpha_{i,t}N_t$ has the same distribution as $\sum_{t=2}^{n} -Y_{t-1}N_t$ respective to $S_i(\theta^*)$. We can apply these trajectories in general ARX systems, but first let us consider a first order case where $n_a = n_b = 1$.

### 2.3.1 First order ARX system

In the first order case the following equation holds

$$Y_t = -a_1^* Y_{t-1} + b_1^* U_{t-1} + N_t. \tag{2.62}$$

with $|a_1^*| < 1$. Assume that B1 and B2 are true and we have $n$ observations $\{(Y_t, U_t)_{t=0}^n\}$, however notice that $U_n$ is not necessary. We also assume that the exogenous inputs are independent of each other and the noise. We would like to apply the SPS on this system. Our aim is to construct exact confidence region for $\theta^*$ around the LS estimator.

First, for any given parameter $\theta = [a_1, b_1]$ we can calculate the predicted residuals as in the previous section

$$\widehat{N}_t(\theta) \triangleq Y_t - \varphi^T \theta \tag{2.63}$$

for all $t = 1, \ldots, n$.

The level of the confidence region remains user-chosen so we need to choose integers $m$ and $q$ such that $\mathbb{P}(\theta^* \in \Theta_n) = 1 - \frac{q}{m}$. Again, in the initialization process we have to generate $m-1$ random sign vectors. As we recall these random signs were denoted this way: $\{\alpha_{i,t}\}_{t=1}^n$ for all $i = 0, \ldots, m-1$ where $\alpha_{0,t} = 1$ for all $t = 1, \ldots, n$. We know that each element of these signs are identically distributed and independent of any other variable of the system. With these random signs we can perturb the predicted errors and calculate $(\alpha_{i,t}\widehat{N}_t(\theta))_{t=1}^n$ for all $i = 0, \ldots, m-1$. Using these alternative perturbed errors we can predict alternative outcomes

$$\bar{Y}_t(\theta, \alpha_i) \triangleq -a_1 \bar{Y}_{t-1} + b_1 U_{t-1} + \alpha_{i,t}\widehat{N}_t(\theta), \tag{2.64}$$

with the initial condition $\bar{Y}_1(\theta, \alpha_i) = Y_1$. It is easy to see that for $\theta = \theta^*$ we have

$$\widehat{N}_t(\theta^*) = N_t \tag{2.65}$$

and $\bar{Y}_{t-1}$ can be expressed as

$$\bar{Y}_t(\theta^*, \alpha_i) = (-a_1^*)^{t-1} Y_1 + \sum_{k=1}^{t-1} b_1^* (-a_1^*)^{k-1} U_{t-k} + \sum_{k=1}^{t-1} (-a_1^*)^{k-1} \alpha_{i,t-k} N_{t-k+1}. \tag{2.66}$$

We can also form alternative regressors

$$\bar{\varphi}_t(\theta, \alpha_i) \triangleq [-\bar{Y}_t(\theta, \alpha_i), U_t]^{\mathrm{T}}. \tag{2.67}$$

Using these now we are able to define the sign-perturbed sums and the reference sum

$$S_0(\theta) \triangleq R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t(Y_t - \varphi_t^{\mathrm{T}}\theta), \tag{2.68}$$

$$S_i(\theta) \triangleq \bar{R}_n^{-\frac{1}{2}}(\theta, \alpha_i) \sum_{t=1}^{n} \bar{\varphi}_t(\theta, \alpha_i)\alpha_{i,t}(Y_t - \varphi_t^{\mathrm{T}}\theta), \tag{2.69}$$

for $i = 1, \ldots, m-1$, where

$$\bar{R}_n^{-\frac{1}{2}}(\theta, \alpha_i) \triangleq \frac{1}{n} \sum_{t=0}^{n} \bar{\varphi}_t(\theta, \alpha_i)\bar{\varphi}_t^{\mathrm{T}}(\theta, \alpha_i). \tag{2.70}$$

Notice that for $\theta = \theta^*$ we know that $R_n^{-\frac{1}{2}} = \bar{R}_n^{-\frac{1}{2}}(\theta^*, \mathbb{1})$ and $Y_{t-1} = \bar{Y}_{t-1}(\theta^*, \mathbb{1})$, since $\widehat{N}_t(\theta^*) = N_t$ where $\mathbb{1}$ is the all one vector. Calculating the reference sum respective to $\theta^*$ we get

$$S_0(\theta^*) = R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t(Y_t - \varphi_t^{\mathrm{T}}\theta^*) = \bar{R}_n^{-\frac{1}{2}}(\theta^*, \mathbb{1}) \sum_{t=1}^{n} \begin{bmatrix} -\bar{Y}_t(\theta^*, \alpha_i) \\ U_{t-1} \end{bmatrix} N_t \tag{2.71}$$

Evaluating the sign-perturbed sums gives us

$$S_i(\theta^*) = \bar{R}_n^{-\frac{1}{2}}(\theta^*, \mathbb{1}) \sum_{t=1}^{n} \begin{bmatrix} -\bar{Y}_t(\theta^*, \alpha_i) \\ U_{t-1} \end{bmatrix} (\alpha_{i,t} N_t) \tag{2.72}$$

We can express $Y_{t-1}$ and $\bar{Y}_t(\theta^*, \alpha_i)$ as

$$Y_t = (-a_1^*)^{t-1} Y_1 + \sum_{k=1}^{t-1} b_1^*(-a_1^*)^{k-1} U_{t-k} + \sum_{k=1}^{t-1} (-a_1^*)^{k-1} N_{t-k+1} \tag{2.73}$$

$$\bar{Y}_t(\theta^*, \alpha_i) = (-a_1^*)^{t-1} Y_1 + \sum_{k=1}^{t-1} b_1^*(-a_1^*)^{k-1} U_{t-k} + \sum_{k=1}^{t-1} (-a_1^*)^{k-1} \alpha_{i,t-k} N_{t-k+1} \tag{2.74}$$

We can see that symmetry has been kept, because in the equation of $S_i(\theta^*)$ all $N_t$ has been replaced by $\alpha_{i,t}N_t$. It is also clear that $Y_t$ and $\bar{Y}_t(\theta^*, \alpha_i)$ are identically distributed, because we know that the noise has a distribution symmetric around zero. It also implies that $S_i(\theta^*)$ and $S_0(\theta^*)$ are identically distributed. Similarly like in the original system we are going to get a uniformly ordered collection of random variables respective to the $\succ_\pi$ total order which was defined earlier. That is why the confidence region we get excluding

those parameters where $\|S_0(\theta)\|^2$ is among the $q$ largest values of $\{\|S_i(\theta)\|^2\}_{i=0}^{m-1}$ remains exact ergo we can proceed as before.

### 2.3.2 General ARX system

When we consider a more general ARX system we may rely on the same idea. Again we would like to predict alternative $\bar{Y}_t$ such that $\sum_{t=2}^{n} -\bar{Y}_{t-1}\alpha_{i,t}N_t$ has the same distribution as $\sum_{t=2}^{n} -Y_{t-1}N_t$ corresponding to $\theta = \theta^*$. We calculate the residuals just like in the first order case

$$\widehat{N}_t(\theta) \triangleq Y_t - \varphi^T\theta \tag{2.75}$$

for all $t = 1, \ldots, n$. Again, in the initialization process we generate $m-1$ random sign vectors, so that each element of these vectors are $\pm 1$ with exactly $\frac{1}{2}$ probability and independent of any other variable of the system. We perturb the predicted residuals and using these alternative perturbed errors we predict alternative outcomes

$$\bar{Y}_t(\theta, \alpha_i) \triangleq -\sum_{j=1}^{n_a} a_j \bar{Y}_{t-j}(\theta, \alpha_i) + \sum_{j=1}^{n_b} b_j U_{t-j} + \alpha_{i,t}\widehat{N}_t(\theta), \tag{2.76}$$

with the initial conditions of $\bar{Y}_t(\theta, \alpha_i) \triangleq Y_t$, for $t \in \{1 - n_a, \ldots, 0\}$. Alternative regressors can be expressed as

$$\bar{\varphi}_t(\theta, \alpha_i) \triangleq [-\bar{Y}_{t-1}(\theta, \alpha_i), \ldots, -\bar{Y}_{t-n_a}(\theta, \alpha_i), U_{t-1}, \ldots, U_{t-n_b}]^{\mathrm{T}}. \tag{2.77}$$

We compute the sign-perturbed sums and the reference sum as in the first order case. See equation (2.71) and (2.72). We can proceed as for simple linear systems. If necessary we have to use a tie-breaking rule. We may use a random permutation for this purpose. The true parameter, $\theta^*$ is included in the confidence region constructed by this extended SPS method with exactly $1 - \frac{q}{m}$ probability, formally

$$\mathbb{P}(\theta^* \in \Theta_m^q) = 1 - \frac{q}{m}. \tag{2.78}$$

# Chapter 3

# Least Absolute Deviations Criterion

Another important well-known point estimator used in linear regression models is the least absolute deviations (LAD) estimator also called the least absolute errors (LAE), least absolute value (LAV), least absolute residual (LAR), sum of absolute deviations or $L_1 - norm$ condition. These names are used to this specific criterion and also for the optimization techniqe that relies on it. As the names indicate this method minimizes the sum of absolute errors or residuals which are going to be introduced formally later. This chapter is based on a very clear and detailed summary written by Dielman [14].

The LAD criterion was first applied by Boscovich R. J. [5] in the middle of the $18^{th}$ century. In this paper it was used to fit a line to observation data. The method remained prior until Legendre announced the LS criterion in 1805 [17]. After that, LAD took a secondary role in solving regression problems. The main reason of this was that LS has a unique analitical solution which is not computationally demanding to find. Another factor was that Gauss [15] and Laplace made a great developement on the method of LS in terms of probability theory. Many of these important results are expounded in the previous chapter.

Although LS is more known there are many advantagous property of the LAD estimator campared to LS. In this chapter I am going to review some of the fundamental results related to LAD and point out the differences and similarities between the LAD and the LS estimator.

First of all in this section the linear regression problem we would like to solve is going to be introduced and possible methods to find a solution are going to be proposed. Later on some of the LAD's important properties are going to be presented.

## 3.1 Linear programming problem

Consider the same linear regression model as in Chapter 2. Again we have a linear system with known structure but unknown true parameter $\theta^*$ so that

$$Y_t = \varphi_t^T \theta^* + N_t. \tag{3.1}$$

The notation is the same as it was before. We have $n$ observations and our goal is to estimate the true parameter.
For each $\theta$ we can calculate the prediction error or residual as before using (2.2).

$$\widehat{N}_t(\theta) \triangleq Y_t - \varphi_t^T \theta \tag{3.2}$$

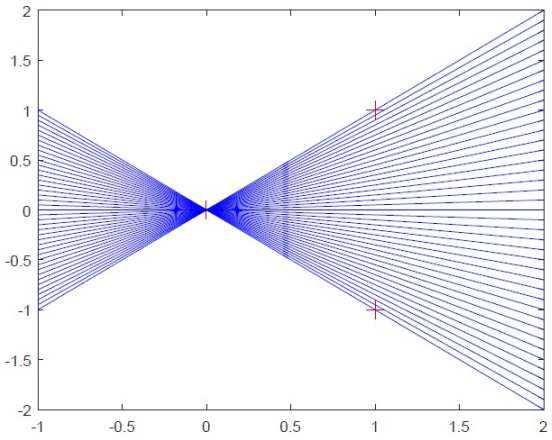
Now we are going to estimate the true parameter by minimizing the sum of the absolute residuals made in each data point. There is no square function introduced this time. Consequently the LAD estimator is

$$\widehat{\theta}_n \triangleq \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{t=1}^{n} \|Y_t - \widehat{Y}_t(\theta)\|_1. \tag{3.3}$$

Notice that we cannot take the derivative with respect to $\theta$ as we did in case of the LS criterion because this time the function that we want to minimize is not differentiable. As we see there is not necessarily a simple analitical solution. Uniqueness is another questionable property of this estimator. It is relatively easy to construct examples that do not have only one solution. You can see one very simple counterexample below on Figure 3.1 based on three observations searching for a two dimensional parameter. As we can see there could be infinitely many estimators. In these cases tie-breaking rules are applied to pick one parameter whenever it is needed.

Even though there is not always a unique analitical solution to this optimization problem, there are methods dealing with it. Notice that our regression problem can be reformulated into a linear programming problem, where the LAD estimator can be found

Figure 3.1: LAD estimator is not always unique. Each line minimizes the residuals in $L1$-norm.




as the optimum of the following optimization problem:

$$\text{minimize} \quad \sum_{t=1}^{n}(d_t^+ + d_t^-) \tag{3.4}$$

$$\text{subject to} \quad \widehat{Y}_t - (\varphi_t^T\widehat{\theta}_n + d_t^+ - d_t^-) = 0 \tag{3.5}$$

$$\forall t = 1, \ldots, n \tag{3.6}$$

where $d_t^+$, $d_t^- \geq 0$, and $\varphi_t$ are the input vectors and $\widehat{\theta}_n$ is the LAD estimator for $n$ observations. The signs of the estimators' coordinates are unrestricted. The $d_t^+$ and $d_t^-$ are the positive and negative residuals associated with the $t^{th}$ observation.

To solve this linear programming (LP) problem we can use the simplex method, which was developed by Dantzig in 1947 [13]. Charnes [7] was the first who used simplex method to solve the LAD regression problem. He and his group applied directly the primal version of the simplex method however it was soon recognised that taking into consideration the special structure of the problem computational efficiency can be improved. Until the 1990s a great variety of algorithm using the simplex method has been developed.

As LP solving improved, nowadays interior point algorithms or ellipsoid algorithms can be applied to solve this problem in polynomial time, even though in practise simplex method is efficient in most of the cases. Portnoy and Kroenker [21] showed that interior point algorithms together with a simple data preprocessing approach can provide a significant improvement in speed. They also noted that simplex-based algorithms can find the LAD estimator in less time than computing LS for a few hundred observations, but

when the size of data gets very large simplex produces a solution much slower.

## 3.2 Properties of the LAD estimator

As it was listed in the first chapter the LS has many advantagous properties. We are going to see that under certain circumstances the LAD estimator has nice properties as well. In this section our aim is to review these therefore proofs are not going to be presented, though they can be found in the referenced papers.

The assumptions we need to take are different than they were before in case of the LS estimator. Sometimes they become relatively complicated.

### 3.2.1 Asymptotic Gaussianity

Koenker and Bassett [3] investigating the asymptotic properties of the LAD estimator proved the following theorem. They needed to make a variety of assumptions.

**C 1.** *Assume that the distribution function (F) of the noise has a median zero.*

**C 2.** *F is continuous and has continous and positive f at median, where f is the density function of the noise.*

**C 3.** *Assume that* $\frac{1}{n}\Phi^T\Phi \to Q$ *as* $n \to \infty$ *where* $Q$ *is a positive definite matrix.*

**Theorem 3.2.1.** *Let* $\{\widehat{\theta}_n\}$ *denote a sequence of unique LAD estimators. Assuming C1, C2 and C3* $\sqrt{n}(\widehat{\theta}_n - \theta^*)$ *converges in distribution to a d dimensional Gaussian random vector variable with mean zero and covariance matrix* $\sigma^2 Q^{-1}$ *where* $\sigma^2$ *is the asymptotic variance of the sample median from random samples of distribution F, i.e.* $\sigma = [2f(0)]^{-1}$.

The proof can be found in [3]. Bassett and Koenker also proved that asymptotic normality holds when the mean residual is zero.

Pollard [20] showed a direct proof of the asymptotic Gaussianity. The technique he used relies on the convexity of the criterion function and this proof is more direct.

### 3.2.2 Consistency

Wu proves strong consistency in [26] under certain conditions for the LAD estimator. He emphasizes that proving the asymptotic theory for this estimate is more difficult as

compared to the LS estimator, because the absolute value function is not differentiable in 0. In the referenced paper Wu provides these conditions for ensuring strong consistency.

**D 1.** $\frac{\lambda_n}{d_n^2 \log n} \to \infty$ *as* $n \to \infty$, *where* $\lambda_n$ *is the smallest eigenvalue of* $\Phi_n$.

**D 2.** *There exists a constant* $k > 1$ *such that* $\frac{d_n}{n^{k-l}} \to 0$.

**D 3.** $N_1, N_2, \ldots, N_n$ *are independent random variables and* $med(N_i) = 0, i = 1, \ldots, n$.

**D 4.** *There exist constants* $C_1 > 0, C_2 > 0$ *such that* $\mathbb{P}(-h < N_i < 0) > C_2 h$ *and* $\mathbb{P}(O < N_i < h) > C_2 h$ *for all* $i = 1, \ldots$ *and* $h \in (0, C_1)$.

**Theorem 3.2.2.** *[26] Assume D1, D2, D3 and D4, then* $\lim_{n\to\infty} \widehat{\theta}_n = \theta^*$ *almost surely.*

In this paper it is also proved that under these conditions $\widehat{\theta}_n \to \theta^*$ rapidly with an exponential rate.

In another paper he and Bai [2] list a variety of conditions and possible assumptions which implies weak consistency. For example this theorem holds.

**D 5.** *The disturbances are independent and come from distribution functions* $F_i$ *each with median zero.*

**D 6.** *There exist positive constants* $p \in (0, 1/2)$ *and* $\delta > 0$ *such that for each* $i = 1, 2, \ldots$

$$min\{\mathbb{P}(N_i > \delta), \mathbb{P}(N_i < -\delta)\} > p$$

**Theorem 3.2.3.** *[2] Assume D5 and D6, then* $\Phi_n^{-1} \to 0$ *as* $n \to \infty$ *is a necessary condition for weak consistency, where* $\Phi = \sum_{t=1}^{n} \varphi_t \varphi_t^T$.

### 3.2.3 Unbiasedness

LAD estimator is unbiased if the conditional distribution of the vector of errors is symmetric given the matrix of regressors. When there is not a unique solution we must use a tie-breaking rule to ensure unbiasedness. For further information see Andrews' [1] results.

### 3.2.4 Comparison between LS and LAD

Though LAD criterion was developed some decades earlier the use of LS became prior due to its uniqueness and easy to compute analitical solution. Nowadays there are efficient methods to find the LAD estimator as well so it can be applied in many cases. Both estimates have good properties such as consistency, asymptotic normality and unbiasedness. In comparison we can say that LAD is more robust than LS. The use of LAD is superior when we deal with heavy-tailed error distribution, because LAD is not so sensitive to outliers as LS. In other words whenever a median is more efficient than the mean as a parameter of a distribution LAD is preferable.

## 3.3 Another version of the SPS method

The non-asymptotic method of SPS can be modified to a more robust version. It was developed by Algo Care, Csáji Balázs and Erik Weyer [6]. We are going to see that even the strongest assumption that we made on the noise term, that it has to be symmetric about zero, can be relaxed. Making small changes in the algorithm implies that we can use this modified SPS when the noise is asymmetric. Now the only restriction we need to make is that the disturbance has a median zero, which is a weaker assumption than symmetricity was. Here we need to know these medians for every $N_t$ and they need to be the same for all noise terms. When the medians are not zero but identical and known we can shift every noise term and this way reduce the problem back to the normal zero median case.

### 3.3.1 LAD-SPS

The normal SPS method constructs an exact confidence region around the LS estimator. This variant of the SPS called the LAD-SPS builds the region around the LAD estimator. It means that if the region is non-empty than the LAD estimator is going to be included. LAD was introduced in the previous section as

$$\widehat{\theta}_n \triangleq \underset{\theta\in\mathbb{R}^d}{\operatorname{argmin}} \sum_{t=1}^{n} \|Y_t - \widehat{Y}_t(\theta)\|_1. \tag{3.7}$$

Its most important properties were discussed and a comparison has been made to the LS before.

In this section we are going to use the $\widehat{\theta}_n$ notation only for the LAD estimator based on $n$ observations. From now on the LS estimator will be denoted by $\theta_{LS}$ to avoid any misunderstanding.

A small change in the original algorithm is going to construct the confidence region around the LAD estimator instead of the LS estimator. The initialization process remains the same as it was before. Hence, the level of the built confidence region stays user-chosen. The only difference we have to make is replacing the original $S_0$ reference sum and $S_i$ sign-perturbed sums where $i = 1, \ldots, m-1$ by new type of sums defined this way

$$Z_0(\theta) \triangleq R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t \operatorname{sign}(\widehat{N}_t(\theta)) \tag{3.8}$$

$$Z_i \triangleq R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \alpha_{i,t} \varphi_t \operatorname{sign}(\widehat{N}_i(\theta)) \tag{3.9}$$

where

$$\widehat{N}_t(\theta) \triangleq Y_t - \varphi^T \theta. \tag{3.10}$$

As we can see we use the same "shaping matrix", because the errors of LAD estimators are also asymptotically Gaussian [20] with the covariance matrix $R^{-1}$, when $R = \lim_{n\to\infty} R_n$ exists and is invertible.

As earlier, the region contains a $\theta$ parameter if and only it is true for this parameter that the norm of the reference sum, $\|Z_0(\theta)\|$, is among the $q$ largest in the strict order on $\{\|Z_i(\theta)\|^2\}_{i=0}^{m-1}$. Remember that we use a tie-breaking rule if it is necessary. First, notice that $Z_0(\widehat{\theta}_n) = 0$ because $\frac{1}{n}\sum_{t=1}^{n} \varphi_t \operatorname{sign}(N_t(\theta))$ is the subgradient of the mean of the absolute deviation error $\frac{1}{n}\sum_{t=1}^{n} \|Y_t - \varphi_t\theta\|$. Therefore the LAD estimator is included in the confidence region.

If we evaluate the reference sum for the true parameter we get the following

$$Z_0(\theta^*) = R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t \operatorname{sign}(\widehat{N}_t(\theta^*)) = R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t \operatorname{sign}(N_t) \tag{3.11}$$

since

$$\widehat{N}_t(\theta^*) = Y_t - \varphi^T \theta^* = N_t. \tag{3.12}$$

Notice that $\operatorname{sign}(N_t) = \pm 1$ with $\frac{1}{2}$ probability implying that $\operatorname{sign}(N_t) = \alpha_{i,t}$ in distribution. Now evaluate the sign-perturbed sums for the true parameter.

$$Z_i(\theta^*) = R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t \alpha_{i,t} \operatorname{sign}(\widehat{N}_t(\theta^*)) = R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t \alpha_{i,t} \operatorname{sign}(N_t) \tag{3.13}$$

It is easy to see that $\operatorname{sign}(N_t) = \alpha_{i,t} \operatorname{sign}(N_t)$ in distribution thus there is no reason why any sign-perturbed sum should be greater or less than the reference sum in the 2-norm. In other words the $Z_i(\theta^*)$ for all $i = 0, \ldots, m-1$ are identically distributed and they are independent as well so after sorting these vectors all orderings are equally likely. The formal proof is similar to the one we had for the original SPS method, though this case is a bit easier. Consequently the true parameter is going to be included in the confidence region with exactly $1 - \frac{q}{m}$ probability. Let $\Theta_m^q$ denote the confidence region built by the

LAD-SPS using $m-1$ sign-perturbed sums, then formally

$$\mathbb{P}(\theta^* \in \Theta_m^q) = 1 - \frac{q}{m} \tag{3.14}$$

meaning that the exact confidence result holds true.

### 3.3.2 Consistency

In the original version of SPS method consistency was a great property of the constructed regions under mild assumptions (B1, B2,B3,B4 and B5). We hoped that similar theorem holds in case of LAD-SPS. Simulation showed us that LAD-SPS can be consistent because the regions are shrinking around the true parameter as $n$ goes to infinity (see 4.8).

It is easy to see that we need to assume that the LAD is consistent, because the LAD estimator is always included in the region unless it is empty. Satisfactory assumptions for LAD's consistency were listed before. Wu's article deals with this topic [26]. Beside that it is reasonable to assume B1, B2, B3 and B4. In addition, suppose that

**B 8.** *$N_t$ has a median zero and continous density function at 0.*

**Theorem 3.3.1.** *Assume B1, B2, B3, B4 and B8. Furthermore suppose that the LAD estimator is consistent. Then pointwise consistency is true. For all $\theta \neq \theta^*$*

$$\mathbb{P}\Big(\bigcap_{k=1}^{\infty}\bigcup_{n=k}^{\infty}\{\theta \in \widehat{\Theta}_n\}\Big) = 0. \tag{3.15}$$

*Proof.* We look at the following quantity $\|Z_0(\theta)\|^2 - \|Z_i(\theta)\|^2$ for a fixed $\theta \neq \theta^*$. We would like to see that there is an $N_i$ so that if $n > N_i$ this quantity is greater than a positive $\varepsilon$. In that case $\|Z_0(\theta)\|^2$ dominates over $\|Z_i(\theta)\|^2$. If it is true for all $i \in \{1, \ldots, m-1\}$, then we can use the maximum of $N_i$, because for all $n > max\{N_i\}$ $\|Z_0(\theta)\|^2$ will be the largest. So eventually $\theta$ will be excluded.

First we are going to see that $\|S_i(\theta)\|^2$ tends to zero as $n$ goes to infinity similarly to the proof in [25]. We know that

$$Z_i(\theta) = R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t \alpha_{i,t} \operatorname{sign}(\widehat{N}_t(\theta)). \tag{3.16}$$

Let denote

$$\Gamma_n \triangleq \frac{1}{n}\sum_{t=1}^{n} \varphi_t \alpha_{i,t} \operatorname{sign}(\widehat{N}_t(\theta)). \tag{3.17}$$

If we look at $\alpha_{i,t}\operatorname{sign}(\widehat{N}_t(\theta))$ we see that it has the same distribution as $\alpha_{i,t}$. Both variables take values $\pm 1$ with $\frac{1}{2}$ probability. We may check the Kolmogorov criterion for every coordinate of $\Gamma_n$

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}[\varphi_{t,k}^2(\alpha_{i,t}\operatorname{sign}(\widehat{N}_t(\theta)))^2]}{t^2} = \frac{\mathbb{E}[\varphi_{t,k}^2 1]}{t^2} \leq \sum_{t=1}^{\infty} \frac{\|\varphi_t\|^2}{t}\frac{1}{t} \leq \sqrt{\sum_{t=1}^{\infty}\frac{\|\varphi\|^4}{t^2}}\sqrt{\sum_{t=1}^{\infty}\frac{1}{t^2}} < \infty \tag{3.18}$$

by using Cauchy-Schwarz inequality and B4. Therefore, we can apply the Strong Law of Large Numbers and we get that $\Gamma_n \to 0$ *almost surely*. We know that $\liminf_{n\to\infty} \lambda_{min}(R_n) > 0$ (see B3) so we get that $R_n^{-\frac{1}{2}}\Gamma_n = Z_i(\theta) \xrightarrow{a.s.} 0$. Consequently

$$\|Z_i(\theta)\|^2 \xrightarrow{a.s.} 0. \tag{3.19}$$

Just like in the original version of SPS we would like to prove that $\|Z_0(\theta)\|^2 \nrightarrow 0$ when $\theta \neq \theta^*$.

Our intuition is that for any $\theta \neq \widehat{\theta}_n$ (LAD estimate) the expected value of the vector $\sum_{t=1}^{n}\varphi_t \operatorname{sign}(\widehat{N}_t(\theta))$ should not be zero therefore its norm converges to a positive value. In [12] it was shown that if $\mathbb{E}\left[\sum_{t=1}^{n}\varphi_t \operatorname{sign}(\widehat{N}_t(\theta))\right]$ equals to zero then $\theta$ also minimizes the $L_1$-norm function. They used B8 there. Consequently $\theta = \widehat{\theta}_n$. It means that the subgradient of the LAD estimator cannot be zero for any other parameter than $\widehat{\theta}_n$.

Again we may use the Strong Law of Large Numbers. The quantity of $\frac{1}{n}\sum_{t=1}^{n}\varphi_t \operatorname{sign}(N_t(\theta))$ almost surely converges to a nonzero vector. We know that $R_n$ is singular and converges to a positive definite matrix (see B3) therefore $R_n^{-\frac{1}{2}}$ has a minimum eigenvalue greater than zero and even $\liminf \lambda_{min}\left(R_n^{-\frac{1}{2}}\right)$ is positive. It follows that $R_n^{-\frac{1}{2}}\sum_{t=1}^{n}\varphi_t \operatorname{sign}(N_t(\theta))$ almost surely converges to a nonzero vector. It also means that its norm has to be positive. More precisely $\|Z_0(\theta)\| \to \varepsilon > 0$. So we got that eventually all $\theta \neq \theta^*$ is going to be excluded from the confidence region. □

## 3.4 LAD-SPS for ARX Systems

In this section we would like to extend the LAD-SPS to ARX systems. As we are going to see this extension is similar to the one that was presented for the original SPS. Now, our goal is to construct an exact confidence region around the LAD estimator for finitely many data points.

Let us consider the following autoregressive exogenous (ARX) stochastic system

$$Y_t + \sum_{j=1}^{n_a} a_j^* Y_{t-j} \triangleq \sum_{j=1}^{n_b} b_j^* U_{t-j} + N_t, \tag{3.20}$$

where $Y_t$ is the output, $U_t$ is the exogenous input and $N_t$ is the noise affecting the system at time $t$. Random variables $\{Y_t\}$, $\{U_t\}$ and $\{N_t\}$ are real-valued. We assume that the inputs are observed and the orders $n_a$ and $n_b$ are known. Regarding the noise, we only assume that $\{N_t\}$ is a sequence of independent random variables, which are also independent of the inputs, $\{U_t\}$, and each $N_t$ is distributed *symmetrically* around zero. Though mediangale distribution of the disturbance is not enough this time, symmetricity is still a very mild condition on the noise as this assumption is met by most of the well-known distribution (Gaussian, Laplace, Cauchy etc.).

The available sample is (w.l.o.g.) assumed to be (re-index and drop superfluous data to achieve this)

$$Y_{1-n_a}, Y_{1-n_a+1}, \ldots, Y_n, U_{1-n_b}, U_{1-n_b+1}, \ldots, U_{n-1}. \tag{3.21}$$

We cannot use the LAD-SPS directly because the $\sum_{t=2}^{n} -Y_{t-1}N_t$ respective to $\theta^*$ does not have the same distribution as $\sum_{t=2}^{n} -Y_{t-1}\alpha_{i,t}N_t$, since $Y_{t-1}$ depends on $N_{t-1}$,$N_{t-2}$ and so on. For this reason we build up alternative trajectories so that these two have the same distribution.

For any given $\theta = [a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}]^{\mathrm{T}}$ to find alternative outcomes first we calculate the prediction errors or residuals

$$\widehat{N}_t(\theta) \triangleq Y_t - \varphi_t^{\mathrm{T}}\theta. \tag{3.22}$$

After that we generate $m-1$ random sign vectors so that $\{\alpha_{i,t}\}$ are i.i.d. random signs (take values $\pm 1$ with probability $1/2$ each). Using these random signs we are able to build

up the the alternative perturbed outputs

$$\bar{Y}_t(\theta,\alpha_i) \triangleq -\sum_{j=1}^{n_a} a_j \bar{Y}_{t-j}(\theta,\alpha_i) + \sum_{j=1}^{n_b} b_j U_{t-j} + \alpha_{i,t}\widehat{N}_t(\theta), \tag{3.23}$$

with the initial conditions $\bar{Y}_t(\theta,\alpha_i) \triangleq Y_t$, for $t \in \{1-n_a,\ldots,0\}$. We can calculate $R_n$ and $\{\bar{R}_n(\theta,\alpha_i)\}$, the refence and perturbed covariance estimates reasonably,

$$R_n \triangleq \frac{1}{n}\sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}}, \tag{3.24}$$

$$\bar{R}_n(\theta,\alpha_i) \triangleq \frac{1}{n}\sum_{t=1}^{n} \bar{\varphi}_t(\theta,\alpha_i)\bar{\varphi}_t^{\mathrm{T}}(\theta,\alpha_i), \tag{3.25}$$

where the reference and perturbed regressor vectors $\{\varphi_t\}$ and $\{\bar{\varphi}_t(\theta,\alpha_i)\}$ are defined as

$$\varphi_t \triangleq [-Y_{t-1},\ldots,-Y_{t-n_a},U_{t-1},\ldots,U_{t-n_b}]^{\mathrm{T}}, \tag{3.26}$$

$$\bar{\varphi}_t(\theta,\alpha_i) \triangleq [-\bar{Y}_{t-1}(\theta,\alpha_i),\ldots,-\bar{Y}_{t-n_a}(\theta,\alpha_i),U_{t-1},\ldots,U_{t-n_b}]^{\mathrm{T}}, \tag{3.27}$$

Then, the reference and the sign-perturbed sums, with respect to parameter $\theta$, can be calculated as

$$S_0(\theta) \triangleq R_n^{-\frac{1}{2}} \frac{1}{n}\sum_{t=1}^{n} \varphi_t \operatorname{sign}(\widehat{N}_t(\theta)), \tag{3.28}$$

$$S_i(\theta) \triangleq \bar{R}_n^{-\frac{1}{2}}(\theta,\alpha_i) \frac{1}{n}\sum_{t=1}^{n} \bar{\varphi}_t(\theta,\alpha_i)\,\alpha_{i,t} \operatorname{sign}(\widehat{N}_t(\theta)), \tag{3.29}$$

for $i \in \{1,\ldots,m-1\}$. With these vectors we can proceed as before.

### 3.4.1 Exact confidence

We saw that for $\theta = \theta^*$

$$\widehat{N}_t(\theta^*) = N_t \tag{3.30}$$

It is also clear that $Y_t$ and $\bar{Y}_t(\theta^*, \alpha_i)$ are the same in distribution, since $N_t$ is symmetrically distributed around zero. It is easy to see that

$$Y_t = \bar{Y}_t(\theta^*, \mathbb{1}) \tag{3.31}$$

$$R_n^{-\frac{1}{2}} = \bar{R}_n^{-\frac{1}{2}}(\theta^*, \mathbb{1}). \tag{3.32}$$

Consequently

$$S_0(\theta^*) = R_n^{-\frac{1}{2}} \frac{1}{n} \sum_{t=1}^{n} \varphi_t \operatorname{sign}(\widehat{N}_t(\theta^*)) = \bar{R}_n^{-\frac{1}{2}}(\theta^*, \mathbb{1}) \frac{1}{n} \sum_{t=1}^{n} \bar{\varphi}_t(\theta^*, \mathbb{1}) \operatorname{sign}(\widehat{N}_t) \tag{3.33}$$

which is identically distributed as the sign perturbed sums

$$S_i(\theta^*) = \bar{R}_n^{-\frac{1}{2}}(\theta^*, \alpha_i) \frac{1}{n} \sum_{t=1}^{n} \bar{\varphi}_t(\theta^*, \alpha_i) \operatorname{sign}(\alpha_{i,t} \widehat{N}_t). \tag{3.34}$$

Although the sums are identically distributed they are not necessarily independent of each other. Nevertheless the same arguments works as in the previous sections to show that $\{\|S_i(\theta^*)\|^2\}_{i=1}^{m-1}$ are uniformly ordered, meaning that all permutations are going to be equally likely after sorting the elements. It gives us the wished result

$$\mathbb{P}(\theta^* \in \Theta_m^q) = 1 - \frac{q}{m} \tag{3.35}$$

where $\Theta_m^q$ is the confidence region we get using this extended version of LAD-SPS for an ARX system.

# Chapter 4

# Appendix

## 4.1 Simulation examples

Trying to understand the SPS method as much as possible I made some simulations on computer generated data.

I considered the following second order linear system

$$Y_t = \varphi_t^T \Theta^* + N_t \tag{4.1}$$

where $\Theta^* \triangleq [3, 3]$ and $N_t$ were i.i.d. uniform variables on $[-0.5, 0.5]$ interval. The regressors were randomly generated as well on the $[0, 6]$ interval. I demonstrated the SPS based on $n = 40$, $n = 80$ and $n = 120$ observations with $m = 100$ and $q = 5$ (see 4.1, 4.2 and 4.3). The resulted confidence regions contained the true parameter with exactly 0.95 probability. I did similar experiments using the LAD-SPS based $n = 80$, $n = 120$ and $n = 400$ observations (see 4.6, 4.7 and 4.8). From these figures we can see how the regions shrink around the true parameter. In other figures (4.4, 4.9) we can see the different layers as the confidence level changes. To achieve these layers I switched parameter $q$ to 2, 5, 10 and 20 respectively in the SPS and LAD-SPS functions.

Figure 4.1: 95% confidence region constructed by the original SPS method, n=40, m=100. The noise was uniform on the $[-0.5, 0.5]$ interval. We can also see the true parameter and the LS estimator.

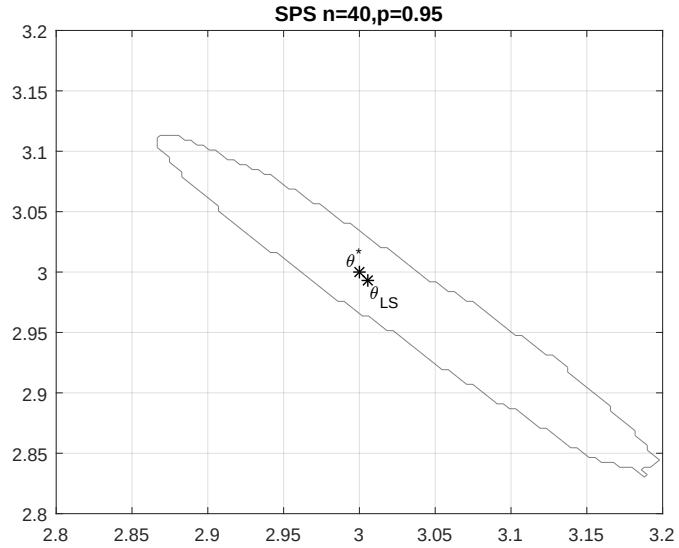



Figure 4.2: 95% confidence region constructed by the original SPS method, n=80, m=100. The noise was uniform on the $[-0.5, 0.5]$ interval. We can also see the true parameter and the LS estimator.

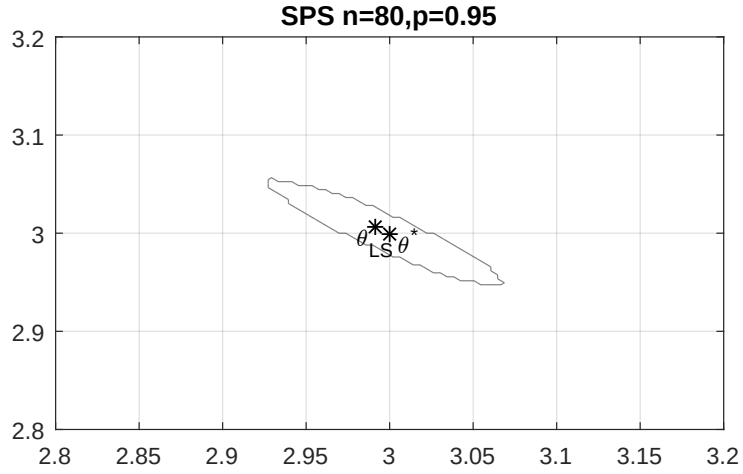


Figure 4.3: 95% confidence region constructed by the original SPS with n=120 and m=100. The true parameter and the LS estimator can also be seen. The noise was uniform on the $[-0.5, 0.5]$ interval.

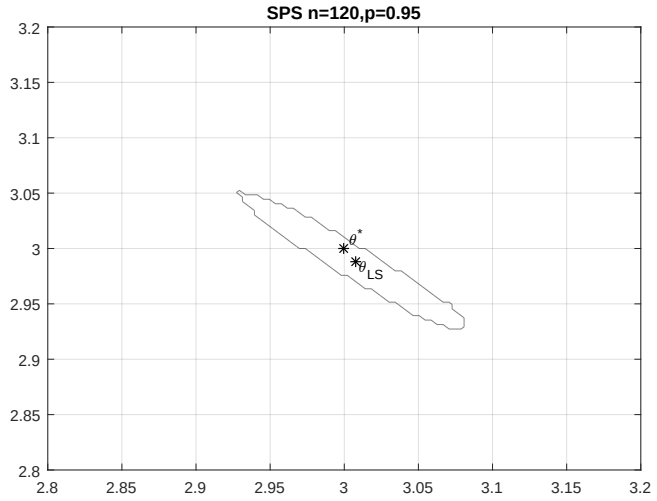



Figure 4.4: The different layers of the confidence region constructed by the SPS are represented here, n=40, m=100 and the noise was uniform on the $[-0.5, 0.5]$ interval. The 80%, the 90%, the 95% and the 98% confidence regions can be seen on the figure.

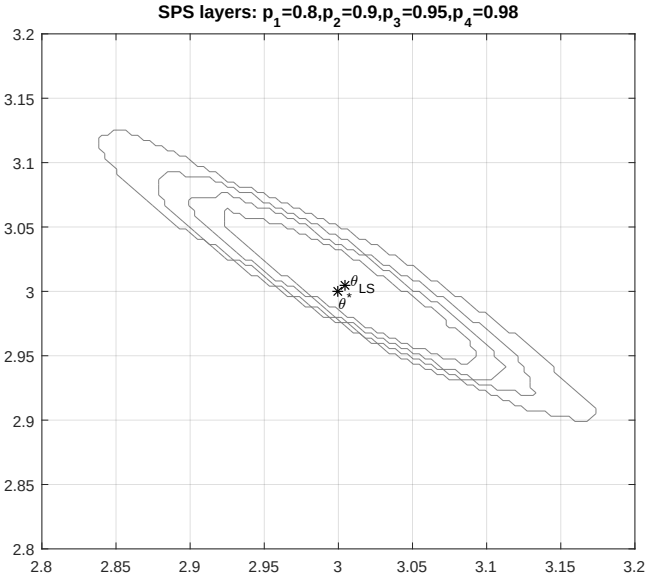


Figure 4.5: In this figure each point is represented with its confidence level. Parameters were n=50, m=100 and the noise was uniform on the $[-0.5, 0.5]$ interval. The darker points are included in the confidence region with higher probability level and the brighter ones are only included in the confidence regions with lower probability level.

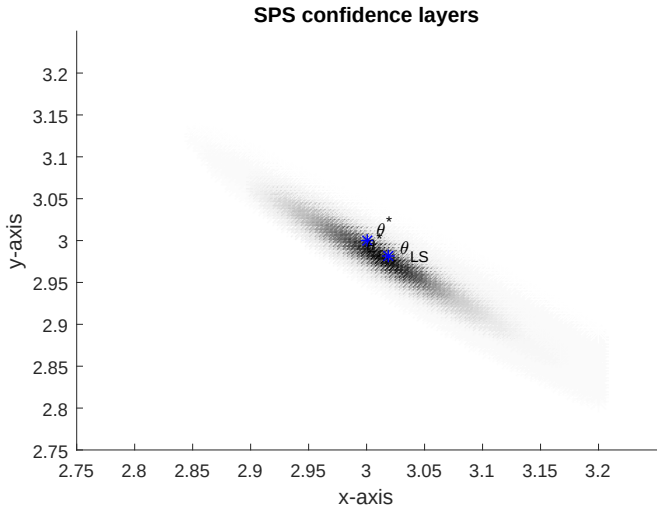



Figure 4.6: 95% confidence region constructed by the LAD-SPS, n=80, m=100. The noise was uniform on the $[-0.5, 0.5]$ interval. We can also see the true parameter.

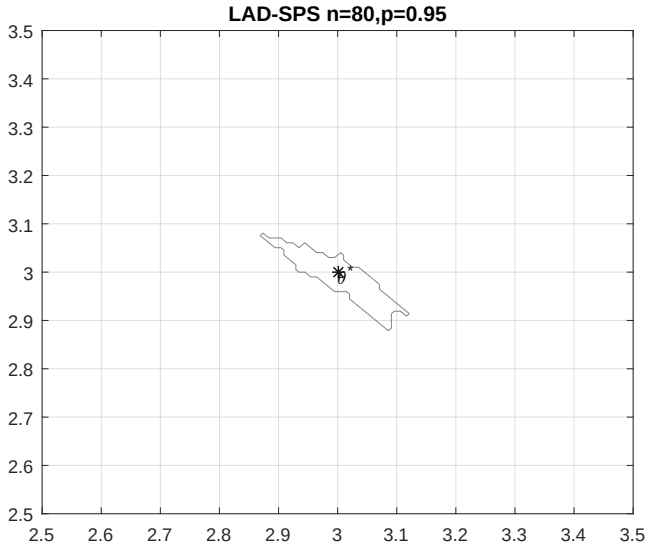


Figure 4.7: 95% confidence region constructed by the LAD-SPS, n=120, m=100. The noise was uniform on the $[-0.5, 0.5]$ interval. We can also see the true parameter.

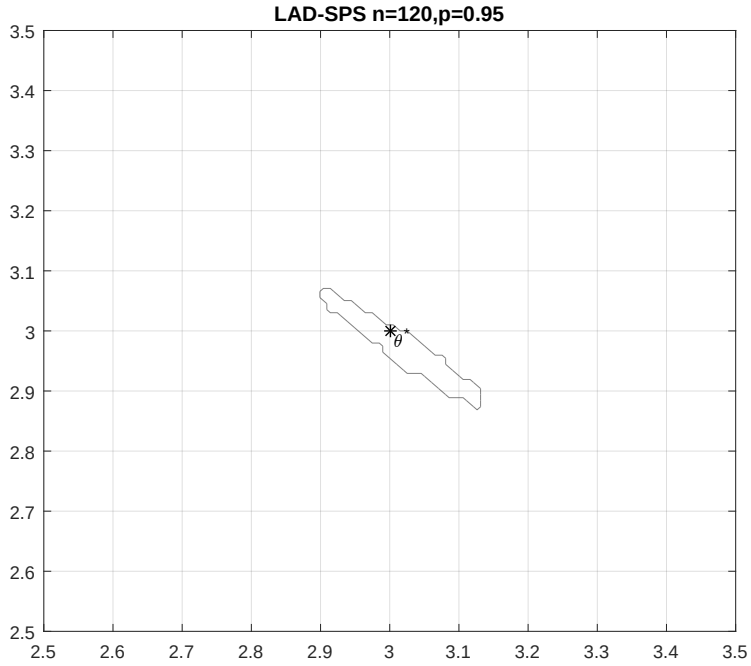



Figure 4.8: 95% confidence region constructed by the LAD-SPS, n=400, m=100. The noise was uniform on the $[-0.5, 0.5]$ interval. We can also see the true parameter.

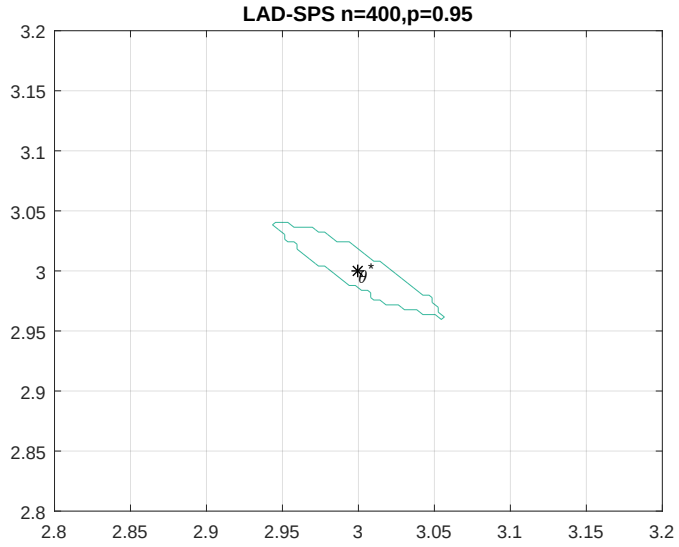
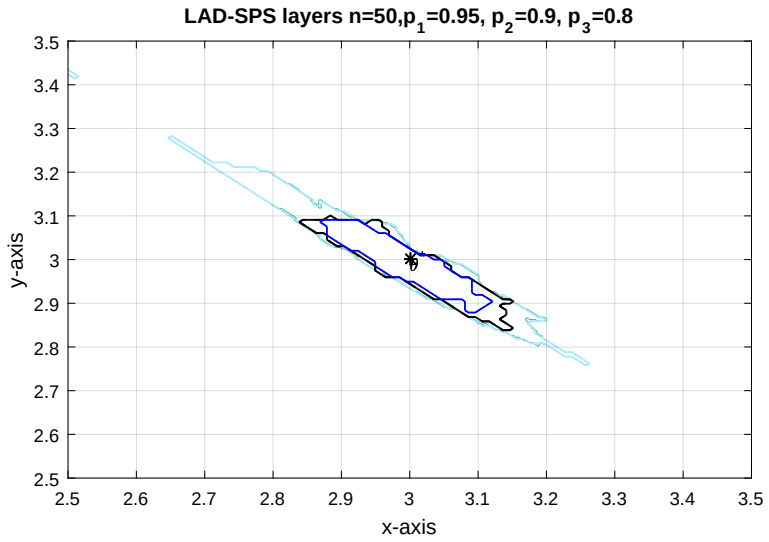


Figure 4.9: The different layers of the confidence region constructed by the LAD-SPS are represented here. Parameters were n=50, m=100 and the noise was uniform on the $[-0.5, 0.5]$ interval. The 80%, the 90%, and the 95% confidence regions can be seen on the figure with different colors.



# Matlab Codes

### 4.1.1 SPS initialization

```
par=[3;3];
n=50;
fi=zeros(2,n);
y=zeros(1,n);
for i=1:n
    fi(1:2,i)=rand(2,1)*6;
    y(1,i)=fi(1:2,i)'*par+rand(1)-0.5;
end
R=1/n*(fi*fi');
Rfel=chol(R);
al=rand(n,99)-0.5;
al=sign(al);
pi=randperm(100);
m=100;
```

### 4.1.2 SPS function

```
function [a]=SPSLS(par,y,fi,al)
   m=100;
    q=5;
    p=1-q/m;
    n=40;
    R=1/n*(fi*fi');
    Rfel=chol(R);
    %randp(1/2);
    %randperm(n);
    d=size(fi,1);
    epsz=(y'-fi'*par);
    RI=inv(Rfel);
    S0=RI*1/n*fi*epsz;
    S=zeros(d,m-1);
```

```
    for  j=1:m-1
      for  i=1:n
        S(1:d,j)=S(1:d,j)+fi(1:d,i)*epsz(i)*al(i,j);
      end
      S(1:d,j)=RI*1/n*S(1:d,j);
    end
    S=[S,S0];
    s=zeros(1,100);
    for  j=1:100
        s(j)=norm  (S(1:d,j));
    end
    s=sort(s);
    norm(S0);
    b=find(s==norm(S0));
    if  b<m-q+1
      a=1;
    else
      a=0;
    %R=1/n*sum(sum(fi*fi'));
    %Rfel=chol(R);
  end
```

### 4.1.3 SPS simulation

```
a=linspace(2.8,3.2,100);
b=linspace(2.8,3.2,100);
c=zeros(100);
for  i=1:100
   for  j=1:100
      c(i,j)=SPSLS([a(i);b(j)],y,fi,al);
   end
end
[X,Y]=meshgrid(a,b);
contour(X,Y,c',1)
```

```
hold on
plot(3,3,'k*')
grid
Theta=inv(fi*fi')*fi*y';
plot(Theta(1),Theta(2),'k*')
text(3,3,'\theta^*')
text(Theta(1),Theta(2),'\theta_{LS}')
title('SPS n=40,p=0.95')
```

### 4.1.4 SPS layers function

```
function [a]=SPSLSplus(par,y,fi,al)
  m=100;
  q=5;
  p=1-q/m;
  n=50;
  R=1/25*(fi*fi');
  Rfel=chol(R);
  %randp(1/2);
  %randperm(n);
  d=size(fi,1);
  epsz=(y'-fi'*par);
  RI=inv(Rfel);
  S0=RI*1/n*fi*epsz;
  S=zeros(d,m-1);
  for j=1:m-1
    for i=1:n
      S(1:d,j)=S(1:d,j)+fi(1:d,i)*epsz(i)*al(i,j);
    end
    S(1:d,j)=RI*1/n*S(1:d,j);
  end
  S=[S,S0];
  s=zeros(1,100);
  for j=1:100
```

```
        s(j)=norm (S(1:d,j));
    end
    s=sort(s);
    norm(S0);
    a=m-find(s==norm(S0));
  end
```

### 4.1.5 SPS layers simulation

```
a=linspace(2.8,3.2,100);
b=linspace(2.8,3.2,100);
figure
hold on
for i=1:100
    for j=1:100
        c=SPSLSplus([a(i);b(j)],y,fi,al)/100;
        plot(a(i),b(j),'*','MarkerSize',10,'MarkerEdgeColor',[1-c 1-
            c 1-c])
        hold on
    end
end
plot(3,3,'b*')
hold on
Theta=inv(fi*fi')*fi*y';
plot(Theta(1),Theta(2),'b*')
title('SPS confidence layers')
text(3,3,'\theta^*')
text(Theta(1),Theta(2),'\theta_{LS}')
xlabel('x-axis')
ylabel('y-axis')
axis([2.75 3.25 2.75 3.25])
```

### 4.1.6 LAD-SPS function

```
function [a]=SPSLAD(par,y,fi,al)
```

```
m=100;
q=5;
p=1-q/m;
n=50;
R=1/n*(fi*fi');
Rfel=chol(R);
%randp(1/2);
%randperm(n);
d=size(fi,1);
epsz=sign(y'-fi'*par);
RI=inv(Rfel);
S0=RI*1/n*fi*epsz;
S=zeros(d,m-1);
for j=1:m-1
   for i=1:n
      S(1:d,j)=S(1:d,j)+fi(1:d,i)*epsz(i)*al(i,j);
   end
   S(1:d,j)=RI*1/n*S(1:d,j);
end
S=[S,S0];
s=zeros(1,100);
for j=1:100
      s(j)=norm (S(1:d,j));
end
s=sort(s);
norm(S0);
b=find(s==norm(S0));
if b<m-q+1
   a=1;
else
   a=0;
%R=1/n*sum(sum(fi*fi'));
%Rfel=chol(R);
end
```

### 4.1.7 LAD-SPS simulation

```
const=100;
a=linspace(2.8,3.2,const);
b=linspace(2.8,3.2,const);
ca=zeros(const);
for i=1:const
   for j=1:const
      ca(i,j)=SPSLAD([a(i);b(j)],y,fi,al);
   end
end
[X,Y]=meshgrid(a,b);
[C,h]=contour(X,Y,ca',1);
contour(X,Y,ca',1)
hold on
plot(3,3,'k*')
title('LAD-SPS n=50,p=0.95')
grid
text(3,3,'\theta^*')
xlabel('x axis')
ylabel('y axis')
```

# Bibliography


[1] D. W. Andrews. A note on the unbiasedness of feasible GLS, quasi-maximum likelihood, robust, adaptive, and spectral estimators of the linear model. *Econometrica: Journal of the Econometric Society*, pages 687–698, 1986.

[2] Z. Bai and Y. Wu. On necessary conditions for the weak consistency of minimum L1-norm estimates in linear models. *Statistics & probability letters*, 34(2):193–199, 1997.

[3] G. Bassett Jr and R. Koenker. Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73(363):618–622, 1978.

[4] A. Bjerhammar. *Application of calculus of matrices to method of least squares: With special reference to geodetic calculations*. Elander, 1951.

[5] R. J. Boscovich. De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impessa. *Bononiensi Scientiarum et Artum Instuto Atque Academia Commentarii*, 4:353–396, 1757.

[6] A. Care, B. Cs. Csáji, and M. C. Campi. Sign-perturbed sums (SPS) with asymmetric noise: Robustness analysis and robustification techniques. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 262–267. IEEE, 2016.

[7] A. Charnes, W. W. Cooper, and R. O. Ferguson. Optimal estimation of executive compensation by linear programming. *Management science*, 1(2):138–151, 1955.

[8] N. Christopeit and K. Helmes. Strong consistency of least squares estimators in linear regression models. *The Annals of Statistics*, pages 778–788, 1980.

[9] H. Cramér. Mathematical methods of statistics. Princeton, NJ. *Princeton University Press*, 16:1964–2, 1946.

[10] B. Cs. Csáji, M. Campi, and E. Weyer. Non-asymptotic confidence regions for the Least-Squares estimate. In *Proceedings of the 16th IFAC Symposium on System Identification*, pages 227 – 232, 2012.

[11] B. Cs. Csáji, M. C. Campi, and E. Weyer. Sign-perturbed sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models. *IEEE Transactions on Signal Processing*, 63(1):169–181, 2015.

[12] B. Cs. Csáji and E. Weyer. Recursive estimation of ARX systems using binary sensors with adjustable thresholds. *IFAC Proceedings Volumes*, 45(16):1185–1190, 2012.

[13] G. B. Dantzig, A. Orden, P. Wolfe, et al. The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics*, 5(2):183–195, 1955.

[14] T. E. Dielman. Least absolute value regression: recent contributions. *Journal of Statistical Computation and Simulation*, 75(4):263–286, 2005.

[15] C. F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss*. sumtibus Frid. Perthes et IH Besser, 1809.

[16] C. Lee. *Linear Regression Model*. Ins.of Economics, NSYSU, Taiwan, 2017.

[17] A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

[18] E. Moore. On the reciprocal of the general algebraic matrix, abstract. *Bull. Amer. Math. Soc.*, 26:394–395, 1920.

[19] R. Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge University Press, 1955.

[20] D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

[21] S. Portnoy, R. Koenker, et al. The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300, 1997.

[22] K. M. Prasad and R. Bapat. The generalized Moore-Penrose inverse. *Linear Algebra and its Applications*, 165:59–69, 1992.

[23] C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pages 235–247. Springer, 1992.

[24] V. Volpe. Identification of dynamical systems with finitely many data points, 2015.

[25] E. Weyer, M. C. Campi, and B. Cs. Csáji. Asymptotic properties of SPS confidence regions. *Automatica*, 82:287–294, 2017.

[26] Y. Wu. Strong consistency and exponential rate of the "minimum L1-norm" estimates in linear regression models. *Computational Statistics & Data Analysis*, 6(3):285–295, 1988.