Variational properties of the Square Root LASSO: Smoothness, uniqueness, explicit solutions

CMS S23 Mathematics of Machine Learning

Aaron Berk
Postdoctoral researcher
McGill University
aaronberk.ca

3 June 2023

Acknowledgements

Collaborators

- Simone Brugiapaglia (Concordia)
- Tim Hoheisel (McGill)

Funding

CENTRE DE RECHERCHES MATHÉMATIQUES

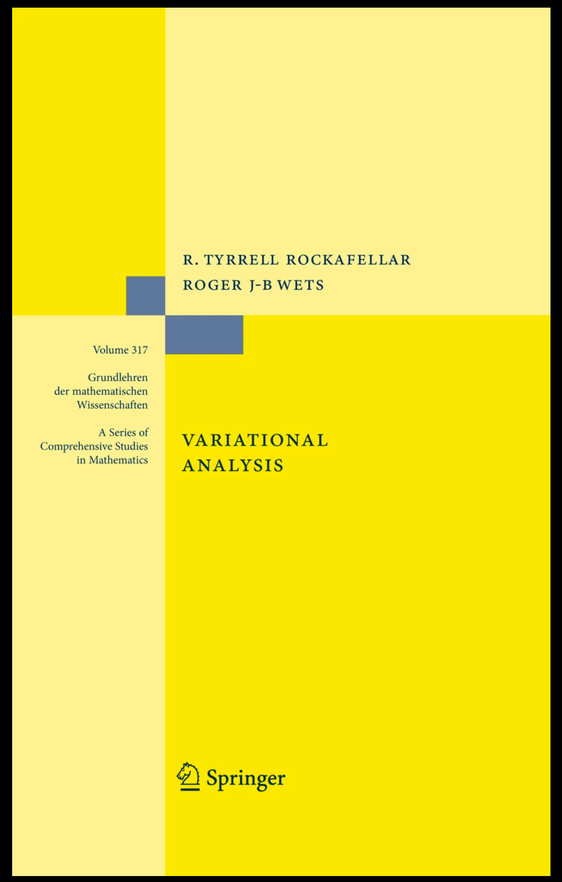
IVADO

NSERC CRSNG

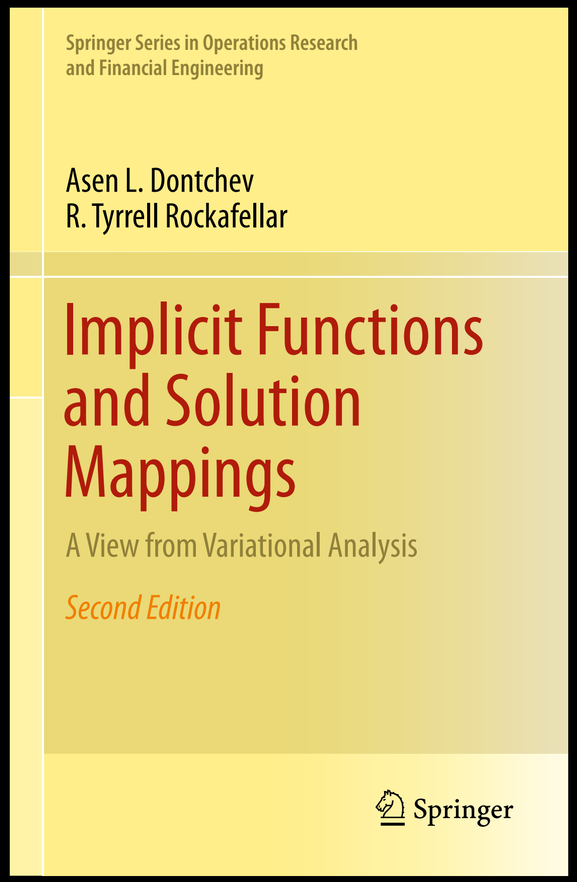
Variational Analysis

Key References

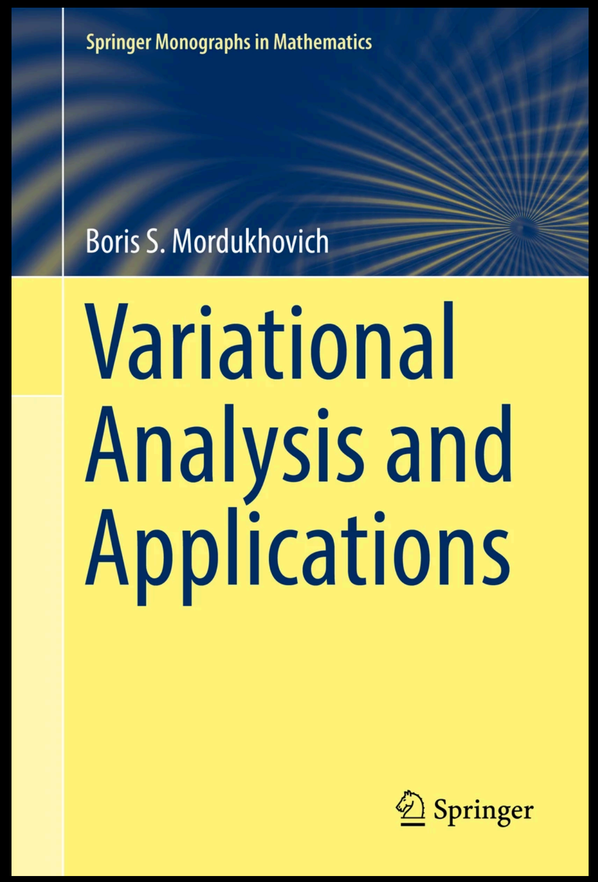
[Rockafellar & Wets, 2009]



[Dontchev & Rockafellar, 2014]



[Mordukhovich, 2018]



The LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) is:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1 \qquad \begin{cases} A \in \mathbb{R}^{m \times n} \\ b \in \mathbb{R}^m \\ \lambda > 0 \end{cases} \qquad \text{(UC)}$$

The LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) is:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1 \qquad \begin{cases} A \in \mathbb{R}^{m \times n} \\ b \in \mathbb{R}^m \\ \lambda > 0 \end{cases} \qquad \text{(UC)}$$

Data fidelity

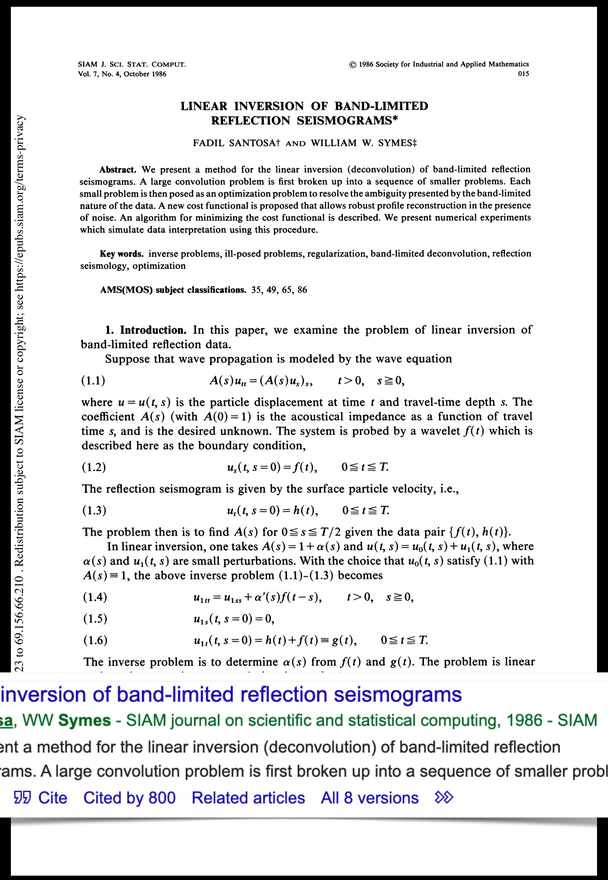
Tuning parameter

Regularization

The LASSO

More than 35 years of history

[Santosa & Symes, 1986]



SIAM J. SCI. STAT. COMPUT.
Vol. 7, No. 4, October 1986

015

LINEAR INVERSION OF BAND-LIMITED REFLECTION SEISMOGRAMS*

FADIL SANTOSA† AND WILLIAM W. SYMES‡

Abstract. We present a method for the linear inversion (deconvolution) of band-limited reflection seismograms. A large convolution problem is first broken up into a sequence of smaller problems. Each small problem is then posed as an optimization problem to resolve the ambiguity presented by the band-limited nature of the data. A new cost functional is proposed that allows robust profile reconstruction in the presence of noise. An algorithm for minimizing the cost functional is described. We present numerical experiments which simulate data interpretation using this procedure.

Key words. inverse problems, ill-posed problems, regularization, band-limited deconvolution, reflection seismology, optimization

AMS(MOS) subject classifications. 35, 49, 65, 86

1. Introduction. In this paper, we examine the problem of linear inversion of band-limited reflection data.

Suppose that wave propagation is modeled by the wave equation

$$A(s)u_{tt} = (A(s)u_s)_s, \qquad t > 0, \quad s \geqq 0, \tag{1.1}$$

where $u = u(t, s)$ is the particle displacement at time t and travel-time depth s. The coefficient $A(s)$ (with $A(0) = 1$) is the acoustical impedance as a function of travel time s, and is the desired unknown. The system is probed by a wavelet $f(t)$ which is described here as the boundary condition,

$$u_s(t, s = 0) = f(t), \qquad 0 \leqq t \leqq T. \tag{1.2}$$

The reflection seismogram is given by the surface particle velocity, i.e.,

$$u_t(t, s = 0) = h(t), \qquad 0 \leqq t \leqq T. \tag{1.3}$$

The problem then is to find $A(s)$ for $0 \leqq s \leqq T/2$ given the data pair $\{f(t), h(t)\}$.

In linear inversion, one takes $A(s) = 1 + \alpha(s)$ and $u(t, s) = u_0(t, s) + u_1(t, s)$, where $\alpha(s)$ and $u_1(t, s)$ are small perturbations. With the choice that $u_0(t, s)$ satisfy (1.1) with $A(s) \equiv 1$, the above inverse problem (1.1)–(1.3) becomes

$$u_{1tt} = u_{1ss} + \alpha'(s)f(t - s), \qquad t > 0, \quad s \geqq 0, \tag{1.4}$$

$$u_{1s}(t, s = 0) = 0, \tag{1.5}$$

$$u_{1t}(t, s = 0) = h(t) + f(t) \equiv g(t), \qquad 0 \leqq t \leqq T. \tag{1.6}$$

The inverse problem is to determine $\alpha(s)$ from $f(t)$ and $g(t)$. The problem is linear

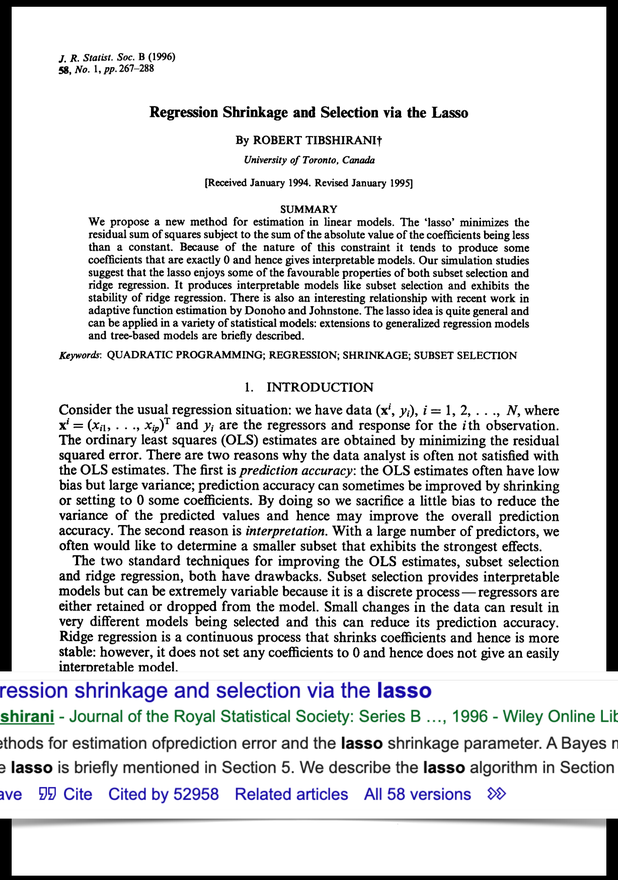
Linear inversion of band-limited reflection seismograms

F **Santosa**, WW **Symes** - SIAM journal on scientific and statistical computing, 1986 - SIAM

We present a method for the linear inversion (deconvolution) of band-limited reflection seismograms. A large convolution problem is first broken up into a sequence of smaller problems.

☆ Save Cite Cited by 800 Related articles All 8 versions »

[Tibshirani, 1996]



J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY

We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Johnstone. The lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.

Keywords: QUADRATIC PROGRAMMING; REGRESSION; SHRINKAGE; SUBSET SELECTION

1. INTRODUCTION

Consider the usual regression situation: we have data $(\mathbf{x}^i, y_i)$, $i = 1, 2, \ldots, N$, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})^T$ and y_i are the regressors and response for the ith observation. The ordinary least squares (OLS) estimates are obtained by minimizing the residual squared error. There are two reasons why the data analyst is often not satisfied with the OLS estimates. The first is *prediction accuracy*: the OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to 0 some coefficients. By doing so we sacrifice a little bias to reduce the variance of the predicted values and hence may improve the overall prediction accuracy. The second reason is *interpretation*. With a large number of predictors, we often would like to determine a smaller subset that exhibits the strongest effects.

The two standard techniques for improving the OLS estimates, subset selection and ridge regression, both have drawbacks. Subset selection provides interpretable models but can be extremely variable because it is a discrete process — regressors are either retained or dropped from the model. Small changes in the data can result in very different models being selected and this can reduce its prediction accuracy. Ridge regression is a continuous process that shrinks coefficients and hence is more stable: however, it does not set any coefficients to 0 and hence does not give an easily interpretable model.

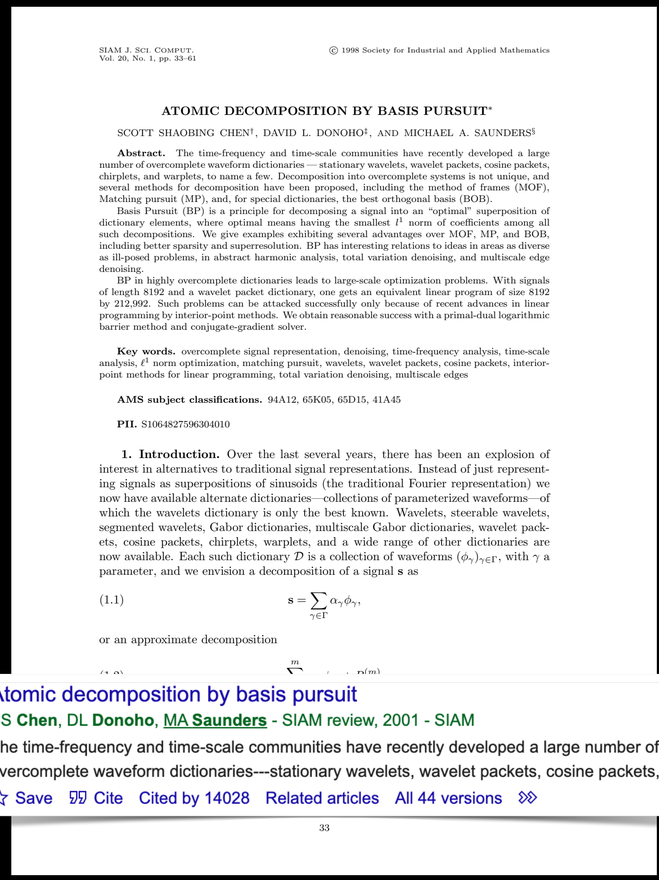
Regression shrinkage and selection via the **lasso**

R **Tibshirani** - Journal of the Royal Statistical Society: Series B …, 1996 - Wiley Online Library

… methods for estimation ofprediction error and the **lasso** shrinkage parameter. A Bayes model for the **lasso** is briefly mentioned in Section 5. We describe the **lasso** algorithm in Section 6. …

☆ Save Cite Cited by 52958 Related articles All 58 versions »

[Chen, Donoho & Saunders, 1998]



SIAM J. SCI. COMPUT.
Vol. 20, No. 1, pp. 33–61

ATOMIC DECOMPOSITION BY BASIS PURSUIT*

SCOTT SHAOBING CHEN†, DAVID L. DONOHO‡, AND MICHAEL A. SAUNDERS§

Abstract. The time-frequency and time-scale communities have recently developed a large number of overcomplete waveform dictionaries — stationary wavelets, wavelet packets, cosine packets, chirplets, and warplets, to name a few. Decomposition into overcomplete systems is not unique, and several methods for decomposition have been proposed, including the method of frames (MOF), Matching pursuit (MP), and, for special dictionaries, the best orthogonal basis (BOB).

Basis Pursuit (BP) is a principle for decomposing a signal into an "optimal" superposition of dictionary elements, where optimal means having the smallest l^1 norm of coefficients among all such decompositions. We give examples exhibiting several advantages over MOF, MP, and BOB, including better sparsity and superresolution. BP has interesting relations to ideas in areas as diverse as ill-posed problems, in abstract harmonic analysis, total variation denoising, and multiscale edge denoising.

BP in highly overcomplete dictionaries leads to large-scale optimization problems. With signals of length 8192 and a wavelet packet dictionary, one gets an equivalent linear program of size 8192 by 212,992. Such problems can be attacked successfully only because of recent advances in linear programming by interior-point methods. We obtain reasonable success with a primal-dual logarithmic barrier method and conjugate-gradient solver.

Key words. overcomplete signal representation, denoising, time-frequency analysis, time-scale analysis, l^1 norm optimization, matching pursuit, wavelets, wavelet packets, cosine packets, interior-point methods for linear programming, total variation denoising, multiscale edges

AMS subject classifications. 94A12, 65K05, 65D15, 41A45

PII. S1064827596304010

1. Introduction. Over the last several years, there has been an explosion of interest in alternatives to traditional signal representations. Instead of just representing signals as superpositions of sinusoids (the traditional Fourier representation) we now have available alternate dictionaries—collections of parameterized waveforms—of which the wavelets dictionary is only the best known. Wavelets, steerable wavelets, segmented wavelets, Gabor dictionaries, multiscale Gabor dictionaries, wavelet packets, cosine packets, chirplets, warplets, and a wide range of other dictionaries are now available. Each such dictionary $\mathcal{D}$ is a collection of waveforms $(\phi_\gamma)_{\gamma \in \Gamma}$, with γ a parameter, and we envision a decomposition of a signal $\mathbf{s}$ as

$$\mathbf{s} = \sum_{\gamma \in \Gamma} \alpha_\gamma \phi_\gamma, \tag{1.1}$$

or an approximate decomposition

$$\sum_{m}$$

Atomic decomposition by basis pursuit

SS **Chen**, DL **Donoho**, MA **Saunders** - SIAM review, 2001 - SIAM

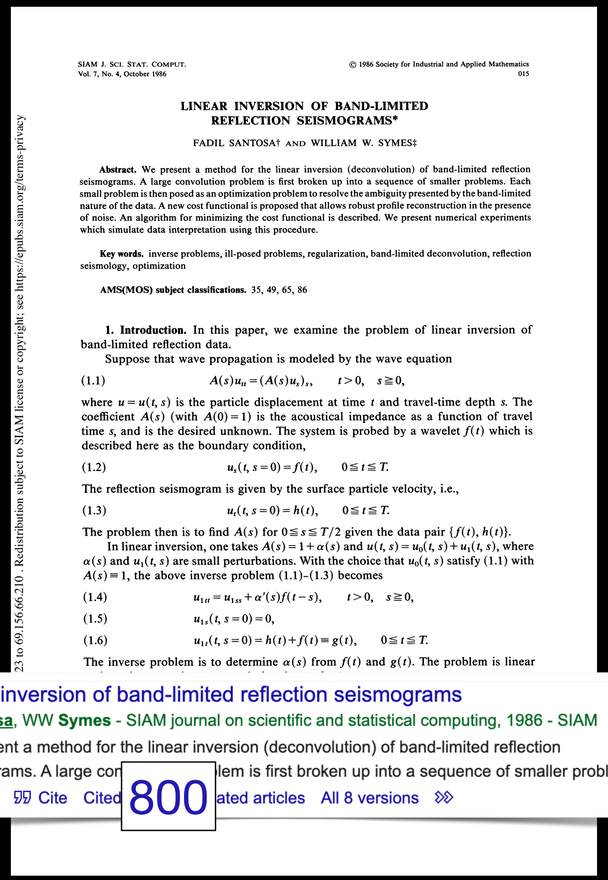
The time-frequency and time-scale communities have recently developed a large number of overcomplete waveform dictionaries---stationary wavelets, wavelet packets, cosine packets, …

☆ Save Cite Cited by 14028 Related articles All 44 versions »

The LASSO

More than 35 years of history

[Santosa & Symes, 1986]



SIAM J. SCI. STAT. COMPUT.
Vol. 7, No. 4, October 1986

LINEAR INVERSION OF BAND-LIMITED REFLECTION SEISMOGRAMS*

FADIL SANTOSA† AND WILLIAM W. SYMES‡

Abstract. We present a method for the linear inversion (deconvolution) of band-limited reflection seismograms. A large convolution problem is first broken up into a sequence of smaller problems. Each small problem is then posed as an optimization problem to resolve the ambiguity presented by the band-limited nature of the data. A new cost functional is proposed that allows robust profile reconstruction in the presence of noise. An algorithm for minimizing the cost functional is described. We present numerical experiments which simulate data interpretation using this procedure.

Key words. inverse problems, ill-posed problems, regularization, band-limited deconvolution, reflection seismology, optimization

AMS(MOS) subject classifications. 35, 49, 65, 86

1. Introduction. In this paper, we examine the problem of linear inversion of band-limited reflection data.

Suppose that wave propagation is modeled by the wave equation

$$A(s)u_{tt} = (A(s)u_s)_s, \qquad t > 0, \quad s \geqq 0, \tag{1.1}$$

where $u = u(t, s)$ is the particle displacement at time t and travel-time depth s. The coefficient $A(s)$ (with $A(0) = 1$) is the acoustical impedance as a function of travel time s, and is the desired unknown. The system is probed by a wavelet $f(t)$ which is described here as the boundary condition,

$$u_s(t, s = 0) = f(t), \qquad 0 \leqq t \leqq T. \tag{1.2}$$

The reflection seismogram is given by the surface particle velocity, i.e.,

$$u_t(t, s = 0) = h(t), \qquad 0 \leqq t \leqq T. \tag{1.3}$$

The problem then is to find $A(s)$ for $0 \leqq s \leqq T/2$ given the data pair $\{f(t), h(t)\}$.

In linear inversion, one takes $A(s) = 1 + \alpha(s)$ and $u(t, s) = u_0(t, s) + u_1(t, s)$, where $\alpha(s)$ and $u_1(t, s)$ are small perturbations. With the choice that $u_0(t, s)$ satisfy (1.1) with $A(s) \equiv 1$, the above inverse problem (1.1)–(1.3) becomes

$$u_{1tt} = u_{1ss} + \alpha'(s)f(t - s), \qquad t > 0, \quad s \geqq 0, \tag{1.4}$$

$$u_{1s}(t, s = 0) = 0, \tag{1.5}$$

$$u_{1t}(t, s = 0) = h(t) + f(t) \equiv g(t), \qquad 0 \leqq t \leqq T. \tag{1.6}$$

The inverse problem is to determine $\alpha(s)$ from $f(t)$ and $g(t)$. The problem is linear

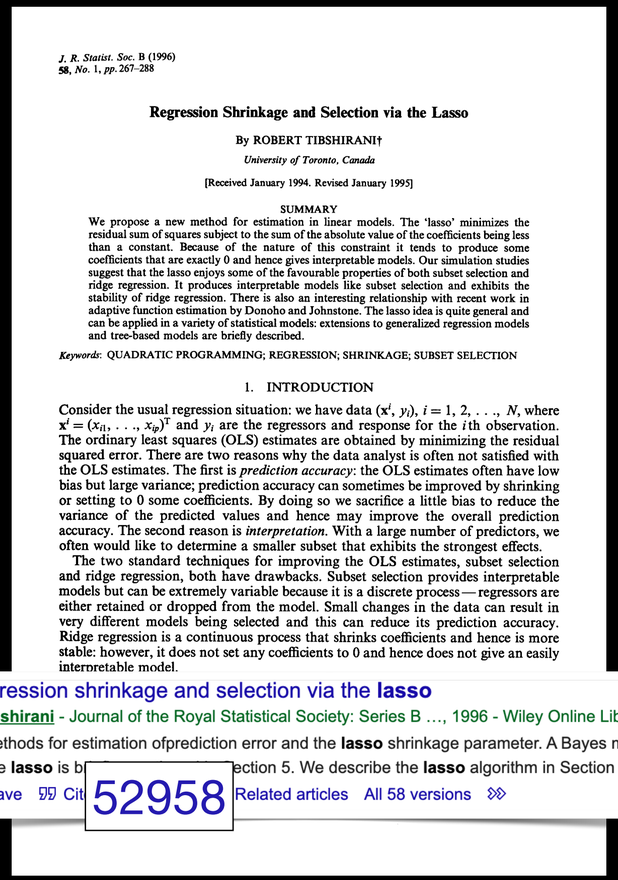
Linear inversion of band-limited reflection seismograms

F **Santosa**, WW **Symes** - SIAM journal on scientific and statistical computing, 1986 - SIAM

We present a method for the linear inversion (deconvolution) of band-limited reflection seismograms. A large cor… …lem is first broken up into a sequence of smaller problems.

☆ Save 99 Cite Cited 800 …ated articles All 8 versions »

[Tibshirani, 1996]



J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY

We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Johnstone. The lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.

Keywords: QUADRATIC PROGRAMMING; REGRESSION; SHRINKAGE; SUBSET SELECTION

1. INTRODUCTION

Consider the usual regression situation: we have data $(\mathbf{x}^i, y_i)$, $i = 1, 2, \ldots, N$, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})^T$ and y_i are the regressors and response for the ith observation. The ordinary least squares (OLS) estimates are obtained by minimizing the residual squared error. There are two reasons why the data analyst is often not satisfied with the OLS estimates. The first is *prediction accuracy*: the OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to 0 some coefficients. By doing so we sacrifice a little bias to reduce the variance of the predicted values and hence may improve the overall prediction accuracy. The second reason is *interpretation*. With a large number of predictors, we often would like to determine a smaller subset that exhibits the strongest effects.

The two standard techniques for improving the OLS estimates, subset selection and ridge regression, both have drawbacks. Subset selection provides interpretable models but can be extremely variable because it is a discrete process — regressors are either retained or dropped from the model. Small changes in the data can result in very different models being selected and this can reduce its prediction accuracy. Ridge regression is a continuous process that shrinks coefficients and hence is more stable: however, it does not set any coefficients to 0 and hence does not give an easily interpretable model.

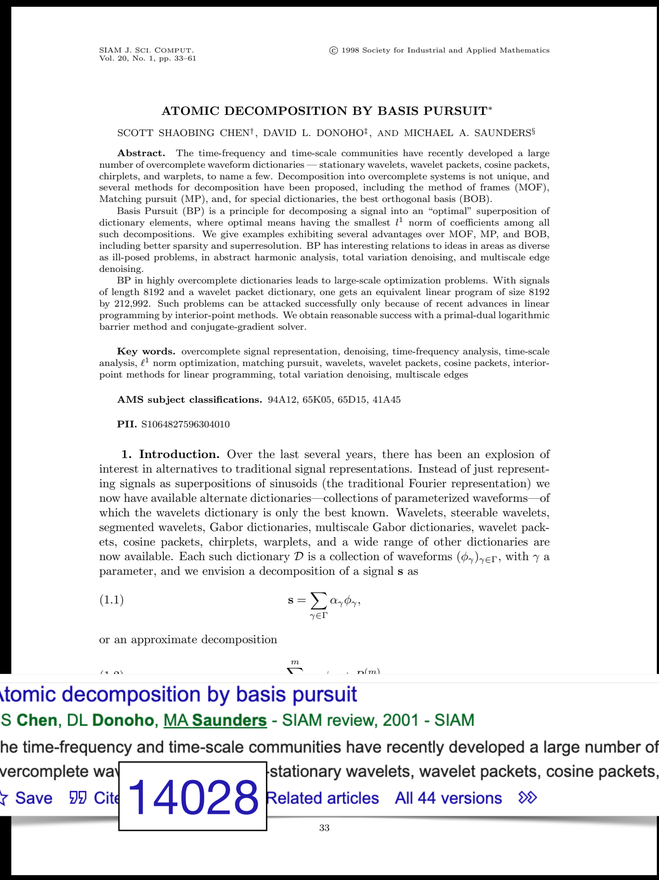
Regression shrinkage and selection via the **lasso**

R **Tibshirani** - Journal of the Royal Statistical Society: Series B …, 1996 - Wiley Online Library

… methods for estimation ofprediction error and the **lasso** shrinkage parameter. A Bayes model for the **lasso** is b… …ection 5. We describe the **lasso** algorithm in Section 6. …

☆ Save 99 Cit 52958 Related articles All 58 versions »

[Chen, Donoho & Saunders, 1998]



SIAM J. SCI. COMPUT.
Vol. 20, No. 1, pp. 33–61

ATOMIC DECOMPOSITION BY BASIS PURSUIT*

SCOTT SHAOBING CHEN†, DAVID L. DONOHO‡, AND MICHAEL A. SAUNDERS§

Abstract. The time-frequency and time-scale communities have recently developed a large number of overcomplete waveform dictionaries — stationary wavelets, wavelet packets, cosine packets, chirplets, and warplets, to name a few. Decomposition into overcomplete systems is not unique, and several methods for decomposition have been proposed, including the method of frames (MOF), Matching pursuit (MP), and, for special dictionaries, the best orthogonal basis (BOB).

Basis Pursuit (BP) is a principle for decomposing a signal into an "optimal" superposition of dictionary elements, where optimal means having the smallest l^1 norm of coefficients among all such decompositions. We give examples exhibiting several advantages over MOF, MP, and BOB, including better sparsity and superresolution. BP has interesting relations to ideas in areas as diverse as ill-posed problems, in abstract harmonic analysis, total variation denoising, and multiscale edge denoising.

BP in highly overcomplete dictionaries leads to large-scale optimization problems. With signals of length 8192 and a wavelet packet dictionary, one gets an equivalent linear program of size 8192 by 212,992. Such problems can be attacked successfully only because of recent advances in linear programming by interior-point methods. We obtain reasonable success with a primal-dual logarithmic barrier method and conjugate-gradient solver.

Key words. overcomplete signal representation, denoising, time-frequency analysis, time-scale analysis, l^1 norm optimization, matching pursuit, wavelets, wavelet packets, cosine packets, interior-point methods for linear programming, total variation denoising, multiscale edges

AMS subject classifications. 94A12, 65K05, 65D15, 41A45

PII. S1064827596304010

1. Introduction. Over the last several years, there has been an explosion of interest in alternatives to traditional signal representations. Instead of just representing signals as superpositions of sinusoids (the traditional Fourier representation) we now have available alternate dictionaries—collections of parameterized waveforms—of which the wavelets dictionary is only the best known. Wavelets, steerable wavelets, segmented wavelets, Gabor dictionaries, multiscale Gabor dictionaries, wavelet packets, cosine packets, chirplets, warplets, and a wide range of other dictionaries are now available. Each such dictionary $\mathcal{D}$ is a collection of waveforms $(\phi_\gamma)_{\gamma \in \Gamma}$, with γ a parameter, and we envision a decomposition of a signal $\mathbf{s}$ as

$$\mathbf{s} = \sum_{\gamma \in \Gamma} \alpha_\gamma \phi_\gamma, \tag{1.1}$$

or an approximate decomposition

$$\sum_{m}$$

Atomic decomposition by basis pursuit

SS **Chen**, DL **Donoho**, MA **Saunders** - SIAM review, 2001 - SIAM

The time-frequency and time-scale communities have recently developed a large number of overcomplete wav… …stationary wavelets, wavelet packets, cosine packets, …

☆ Save 99 Cite 14028 Related articles All 44 versions »

The LASSO

More than 35 years of history

[Santosa & Symes, 1986] [Tibshirani, 1996] [Chen, Donoho & Saunders, 1998]

[Berk, Brugiapaglia & Hoheisel, 2022]

LASSO reloaded: a variational analysis perspective with applications to compressed sensing*

Aaron Berk†, Simone Brugiapaglia‡, and Tim Hoheisel§

Abstract. This paper provides a variational analysis of the unconstrained formulation of the LASSO problem, ubiquitous in statistical learning, signal processing, and inverse problems. In particular, we establish smoothness results for the optimal value as well as Lipschitz and smoothness properties of the optimal solution as functions of the right-hand side (or *measurement vector*) and the regularization parameter. Moreover, we show how to apply the proposed variational analysis to study the sensitivity of the optimal solution to the tuning parameter in the context of compressed sensing with subgaussian measurements. Our theoretical findings are validated by numerical experiments.

Key words. Variational analysis, LASSO, compressed sensing, coderivative, graphical derivative, metric regularity

AMS subject classifications. 49J53, 62J07, 90C25, 94A12, 94A20

Linear inversion of band-limited reflection seismograms

F Santosa, WW Symes - SIAM journal on scientific and statistical computing, 1986 - SIAM

We present a method for the linear inversion (deconvolution) of band-limited reflection seismograms. A large cor…lem is first broken up into a sequence of smaller problems.

☆ Save 99 Cite Cited 800 …ated articles All 8 versions »

Regression shrinkage and selection via the lasso

R Tibshirani - Journal of the Royal Statistical Society: Series B …, 1996 - Wiley Online Library

… methods for estimation ofprediction error and the **lasso** shrinkage parameter. A Bayes model for the **lasso** is b… …ection 5. We describe the **lasso** algorithm in Section 6. …

☆ Save 99 Cit… 52958 Related articles All 58 versions »

Atomic decomposition by basis pursuit

SS Chen, DL Donoho, MA Saunders - SIAM review, 2001 - SIAM

The time-frequency and time-scale communities have recently developed a large number of overcomplete wav… …stationary wavelets, wavelet packets, cosine packets, …

☆ Save 99 Cite 14028 Related articles All 44 versions »

The Square Root LASSO

The Square Root LASSO is defined by:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\| + \lambda \|x\|_1 \qquad \begin{cases} A \in \mathbb{R}^{m \times n} \\ b \in \mathbb{R}^m \\ \lambda > 0 \end{cases} \qquad \text{(SR)}$$

The Square Root LASSO

The Square Root LASSO is defined by:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\| + \lambda \|x\|_1 \qquad \begin{cases} A \in \mathbb{R}^{m \times n} \\ b \in \mathbb{R}^m \\ \lambda > 0 \end{cases} \qquad \text{(SR)}$$

Data fidelity

Tuning parameter

Regularization

The Square Root LASSO

The (set-valued) solution mapping for (SR) is given by:

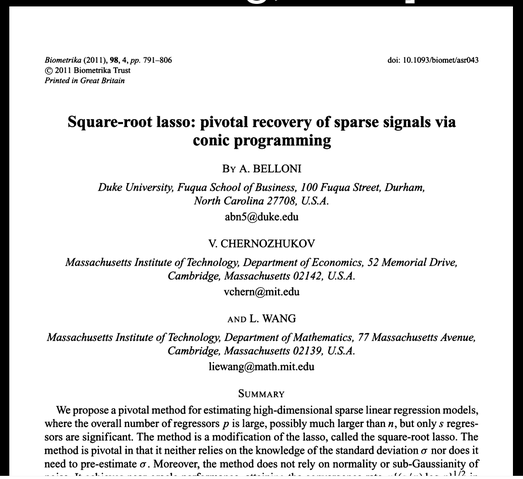
$$S : \mathbb{R}^m \times \mathbb{R}_{++} \rightrightarrows \mathbb{R}^n$$

$$(b, \lambda) \mapsto \underset{x \in \mathbb{R}^n}{\arg\min} \{\|Ax - b\| + \lambda \|x\|_1\}$$

The Square Root LASSO

Almost a teenager

[Belloni, Chernozhukov & Wang, 2011]



Biometrika (2011), **98**, 4, *pp.* 791–806 doi: 10.1093/biomet/asr043
© 2011 Biometrika Trust
Printed in Great Britain

Square-root lasso: pivotal recovery of sparse signals via conic programming

BY A. BELLONI

Duke University, Fuqua School of Business, 100 Fuqua Street, Durham, North Carolina 27708, U.S.A.
abn5@duke.edu

V. CHERNOZHUKOV

Massachusetts Institute of Technology, Department of Economics, 52 Memorial Drive, Cambridge, Massachusetts 02142, U.S.A.
vchern@mit.edu

AND L. WANG

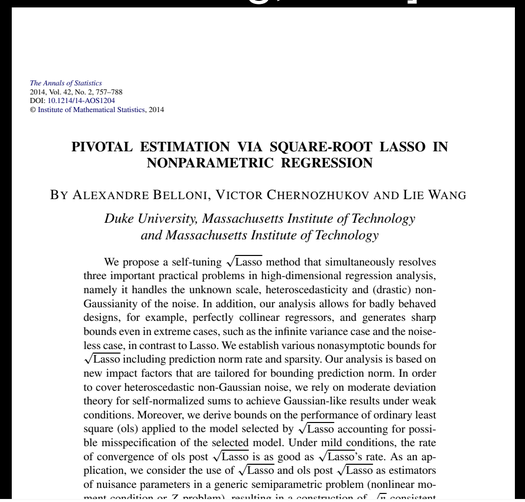
Massachusetts Institute of Technology, Department of Mathematics, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, U.S.A.
liewang@math.mit.edu

SUMMARY

We propose a pivotal method for estimating high-dimensional sparse linear regression models, where the overall number of regressors p is large, possibly much larger than n, but only s regressors are significant. The method is a modification of the lasso, called the square-root lasso. The method is pivotal in that it neither relies on the knowledge of the standard deviation σ nor does it need to pre-estimate σ. Moreover, the method does not rely on normality or sub-Gaussianity of

Square-root lasso: pivotal recovery of sparse signals via conic programming
A Belloni, V Chernozhukov, L Wang - Biometrika, 2011 - academic.oup.com
We propose a pivotal method for estimating high-dimensional sparse linear regression models, where the overall number of regressors p is large, possibly much larger than n, but only s regressors are significant. The method is a modification of the lasso, called the square-root lasso. The method is pivotal in that it neither relies on the knowledge of the standard deviation σ nor does it need to pre-estimate σ. Moreover, the method does not rely on normality or sub-Gaussianity of noise. It achieves near-oracle performance, attaining the …
☆ Save 99 Cite Cited by 660 Related articles All 20 versions »

[Belloni, Chernozhukov & Wang, 2014]



The Annals of Statistics
2014, Vol. 42, No. 2, 757–788
DOI: 10.1214/14-AOS1204

PIVOTAL ESTIMATION VIA SQUARE-ROOT LASSO IN NONPARAMETRIC REGRESSION

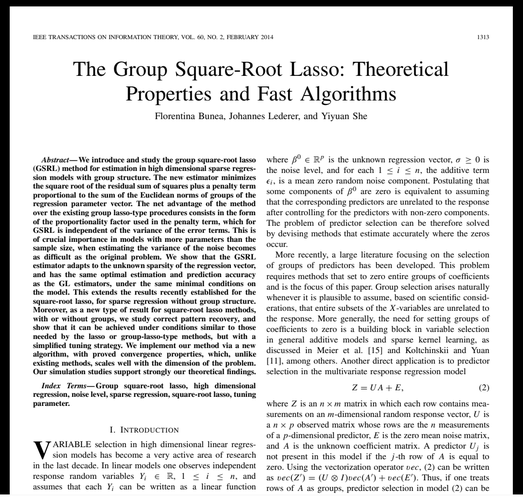
BY ALEXANDRE BELLONI, VICTOR CHERNOZHUKOV AND LIE WANG

Duke University, Massachusetts Institute of Technology and Massachusetts Institute of Technology

We propose a self-tuning $\sqrt{\text{Lasso}}$ method that simultaneously resolves three important practical problems in high-dimensional regression analysis, namely it handles the unknown scale, heteroscedasticity and (drastic) non-Gaussianity of the noise. In addition, our analysis allows for badly behaved designs, for example, perfectly collinear regressors, and generates sharp bounds even in extreme cases, such as the infinite variance case and the noiseless case, in contrast to Lasso. We establish various nonasymptotic bounds for $\sqrt{\text{Lasso}}$ including prediction norm rate and sparsity. Our analysis is based on new impact factors that are tailored for bounding prediction norm. In order to cover heteroscedastic non-Gaussian noise, we rely on moderate deviation theory for self-normalized sums to achieve Gaussian-like results under weak conditions. Moreover, we derive bounds on the performance of ordinary least square (ols) applied to the model selected by $\sqrt{\text{Lasso}}$ accounting for possible misspecification of the selected model. Under mild conditions, the rate of convergence of ols post $\sqrt{\text{Lasso}}$ is as good as $\sqrt{\text{Lasso}}$'s rate. As an application, we consider the use of $\sqrt{\text{Lasso}}$ and ols post $\sqrt{\text{Lasso}}$ as estimators of nuisance parameters in a generic semiparametric problem (nonlinear moment condition or Z problem), resulting in a construction of $\sqrt{n}$ consistent

Pivotal estimation via **square-root lasso** in nonparametric regression
A Belloni, V Chernozhukov, L Wang - 2014 - projecteuclid.org
… In addition, our analysis allows for badly behaved designs, for example, perfectly collinear regressors, and generates sharp bounds even in extreme cases, such as the …
☆ Save 99 Cite Cited by 145 Related articles All 15 versions

[Bunea, Lederer & She, 2014]



The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms

Florentina Bunea, Johannes Lederer, and Yiyuan She

***Abstract*— We introduce and study the group square-root lasso (GSRL) method for estimation in high dimensional sparse regression models with group structure. The new estimator minimizes the square root of the residual sum of squares plus a penalty term proportional to the sum of the Euclidean norms of groups of the regression parameter vector. The net advantage of the method over the existing group lasso-type procedures consists in the form of the proportionality factor used in the penalty term, which for GSRL is independent of the variance of the error terms. This is of crucial importance in models with more parameters than the sample size, when estimating the variance of the noise becomes as difficult as the original problem. We show that the GSRL estimator adapts to the unknown sparsity of the regression vector, and has the same optimal estimation and prediction accuracy as the GL estimators, under the same minimal conditions on the model. This extends the results recently established for the square-root lasso, for sparse regression without group structure. Moreover, as a new type of result for square-root lasso methods, with or without groups, we study correct pattern recovery, and show that it can be achieved under conditions similar to those needed by the lasso or group-lasso-type methods, but with a simplified tuning strategy. We implement our method via a new algorithm, with proved convergence properties, which, unlike existing methods, scales well with the dimension of the problem. Our simulation studies support strongly our theoretical findings.**

***Index Terms*— Group square-root lasso, high dimensional regression, noise level, sparse regression, square-root lasso, tuning parameter.**

I. INTRODUCTION

VARIABLE selection in high dimensional linear regression models has become a very active area of research in the last decade. In linear models one observes independent response random variables $Y_i \in \mathbb{R}$, $1 \leq i \leq n$, and assumes that each Y_i can be written as a linear function of the noise level, and for each $1 \leq i \leq n$, the additive term ϵ_i, is a mean zero random noise component. Postulating that some components of β^0 are zero is equivalent to assuming that the corresponding predictors are unrelated to the response after controlling for the predictors with non-zero components. The problem of predictor selection can be therefore solved by devising methods that estimate accurately where the zeros occur.

More recently, a large literature focusing on the selection of groups of predictors has been developed. This problem requires methods that set to zero entire groups of coefficients and is the focus of this paper. Group selection arises naturally whenever it is plausible to assume, based on scientific considerations, that entire subsets of the X-variables are unrelated to the response. More generally, the need for setting groups of coefficients to zero is a building block in variable selection in general additive models and sparse kernel learning, as discussed in Meier et al. [15] and Koltchinskii and Yuan [11], among others. Another direct application is to predictor selection in the multivariate response regression model

$$Z = UA + E, \qquad (2)$$

where Z is an $n \times m$ matrix in which each row contains measurements on an m-dimensional random response vector, U is a $n \times p$ observed matrix whose rows are the n measurements of a p-dimensional predictor, E is the zero mean noise matrix, and A is the unknown coefficient matrix. A predictor U_j is not present in this model if the j-th row of A is equal to zero. Using the vectorization operator vec, (2) can be written as $vec(Z) = (U \otimes I)vec(A) + vec(E)$. Thus, if one treats rows of A as groups, predictor selection in model (2) can be

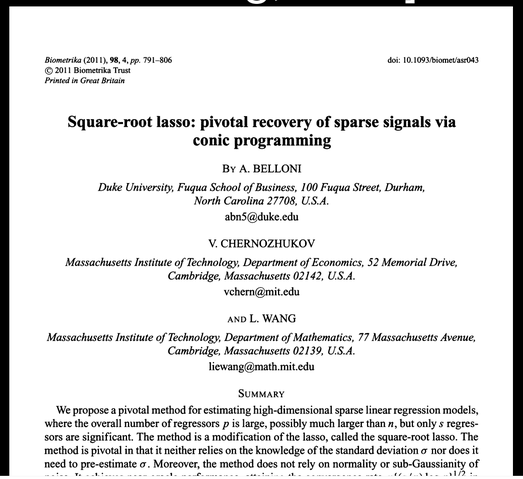
where $\beta^0 \in \mathbb{R}^p$ is the unknown regression vector, $\sigma \geq 0$ is

The group **square-root lasso**: Theoretical properties and fast algorithms
F Bunea, J Lederer, Y She - IEEE Transactions on Information …, 2013 - ieeexplore.ieee.org
… **square-root lasso**, for sparse regression without group structure. Moreover, as a new type of result for **square-root lasso** … to those needed by the **lasso** or group-**lasso**-type methods, but .
☆ Save 99 Cite Cited by 109 Related articles All 9 versions »

The Square Root LASSO

Almost a teenager

[Belloni, Chernozhukov & Wang, 2011]



Biometrika (2011), **98**, 4, *pp.* 791–806 doi: 10.1093/biomet/asr043
© 2011 Biometrika Trust
Printed in Great Britain

Square-root lasso: pivotal recovery of sparse signals via conic programming

BY A. BELLONI

Duke University, Fuqua School of Business, 100 Fuqua Street, Durham, North Carolina 27708, U.S.A.
abn5@duke.edu

V. CHERNOZHUKOV

Massachusetts Institute of Technology, Department of Economics, 52 Memorial Drive, Cambridge, Massachusetts 02142, U.S.A.
vchern@mit.edu

AND L. WANG

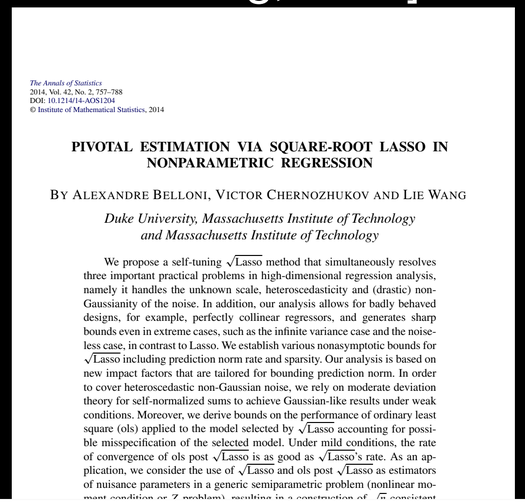
Massachusetts Institute of Technology, Department of Mathematics, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, U.S.A.
liewang@math.mit.edu

SUMMARY

We propose a pivotal method for estimating high-dimensional sparse linear regression models, where the overall number of regressors p is large, possibly much larger than n, but only s regressors are significant. The method is a modification of the lasso, called the square-root lasso. The method is pivotal in that it neither relies on the knowledge of the standard deviation σ nor does it need to pre-estimate σ. Moreover, the method does not rely on normality or sub-Gaussianity of

Square-root lasso: pivotal recovery of sparse signals via conic programming
A Belloni, V Chernozhukov, L Wang - Biometrika, 2011 - academic.oup.com
We propose a pivotal method for estimating high-dimensional sparse linear regression models, where the overall number of regressors p is large, possibly much larger than n, but only s regressors are significant. The method is a modification of the lasso, called the square-root lasso. The method is pivotal in that it neither relies on the knowledge of the standard deviation σ nor does it need to pre-estimate σ. Moreover, the method does not rely on normality or sub… oise. It achieves near-oracle performance, attaining the …
☆ Save 99 Cite 660 Related articles All 20 versions »

[Belloni, Chernozhukov & Wang, 2014]



The Annals of Statistics
2014, Vol. 42, No. 2, 757–788
DOI: 10.1214/14-AOS1204

PIVOTAL ESTIMATION VIA SQUARE-ROOT LASSO IN NONPARAMETRIC REGRESSION

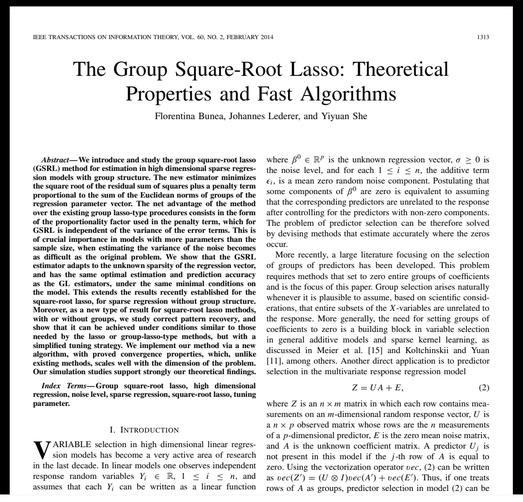
BY ALEXANDRE BELLONI, VICTOR CHERNOZHUKOV AND LIE WANG

Duke University, Massachusetts Institute of Technology and Massachusetts Institute of Technology

We propose a self-tuning $\sqrt{\text{Lasso}}$ method that simultaneously resolves three important practical problems in high-dimensional regression analysis, namely it handles the unknown scale, heteroscedasticity and (drastic) non-Gaussianity of the noise. In addition, our analysis allows for badly behaved designs, for example, perfectly collinear regressors, and generates sharp bounds even in extreme cases, such as the infinite variance case and the noiseless case, in contrast to Lasso. We establish various nonasymptotic bounds for $\sqrt{\text{Lasso}}$ including prediction norm rate and sparsity. Our analysis is based on new impact factors that are tailored for bounding prediction norm. In order to cover heteroscedastic non-Gaussian noise, we rely on moderate deviation theory for self-normalized sums to achieve Gaussian-like results under weak conditions. Moreover, we derive bounds on the performance of ordinary least square (ols) applied to the model selected by $\sqrt{\text{Lasso}}$ accounting for possible misspecification of the selected model. Under mild conditions, the rate of convergence of ols post $\sqrt{\text{Lasso}}$ is as good as $\sqrt{\text{Lasso}}$'s rate. As an application, we consider the use of $\sqrt{\text{Lasso}}$ and ols post $\sqrt{\text{Lasso}}$ as estimators of nuisance parameters in a generic semiparametric problem (nonlinear moment condition or Z problem), resulting in a construction of $\sqrt{n}$-consistent

Pivotal estimation via **square-root lasso** in nonparametric regression
A Belloni, V Chernozhukov, L Wang - 2014 - projecteuclid.org
… In addition, our analysis allows for badly behaved designs, for example, perfectly collinear regressors, … sharp bounds even in extreme cases, such as the …
☆ Save 99 Cite 145 Related articles All 15 versions »

[Bunea, Lederer & She, 2014]



The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms

Florentina Bunea, Johannes Lederer, and Yiyuan She

***Abstract*— We introduce and study the group square-root lasso (GSRL) method for estimation in high dimensional sparse regression models with group structure. The new estimator minimizes the square root of the residual sum of squares plus a penalty term proportional to the sum of the Euclidean norms of groups of the regression parameter vector. The net advantage of the method over the existing group lasso-type procedures consists in the form of the proportionality factor used in the penalty term, which for GSRL is independent of the variance of the error terms. This is of crucial importance in models with more parameters than the sample size, when estimating the variance of the noise becomes as difficult as the original problem. We show that the GSRL estimator adapts to the unknown sparsity of the regression vector, and has the same optimal estimation and prediction accuracy as the GL estimators, under the same minimal conditions on the model. This extends the results recently established for the square-root lasso, for sparse regression without group structure. Moreover, as a new type of result for square-root lasso methods, with or without groups, we study correct pattern recovery, and show that it can be achieved under conditions similar to those needed by the lasso or group-lasso-type methods, but with a simplified tuning strategy. We implement our method via a new algorithm, with proved convergence properties, which, unlike existing methods, scales well with the dimension of the problem. Our simulation studies support strongly our theoretical findings.**

***Index Terms*— Group square-root lasso, high dimensional regression, noise level, sparse regression, square-root lasso, tuning parameter.**

I. INTRODUCTION

VARIABLE selection in high dimensional linear regression models has become a very active area of research in the last decade. In linear models one observes independent response random variables $Y_i \in \mathbb{R}$, $1 \leq i \leq n$, and assumes that each Y_i can be written as a linear function where $\beta^0 \in \mathbb{R}^p$ is the unknown regression vector, $\sigma \geq 0$ is the noise level, and for each $1 \leq i \leq n$, the additive term ϵ_i, is a mean zero random noise component. Postulating that some components of β^0 are zero is equivalent to assuming that the corresponding predictors are unrelated to the response after controlling for the predictors with non-zero components. The problem of predictor selection can be therefore solved by devising methods that estimate accurately where the zeros occur.

More recently, a large literature focusing on the selection of groups of predictors has been developed. This problem requires methods that set to zero entire groups of coefficients and is the focus of this paper. Group selection arises naturally whenever it is plausible to assume, based on scientific considerations, that entire subsets of the X-variables are unrelated to the response. More generally, the need for setting groups of coefficients to zero is a building block in variable selection in general additive models and sparse kernel learning, as discussed in Meier et al. [15] and Koltchinskii and Yuan [11], among others. Another direct application is to predictor selection in the multivariate response regression model

$$Z = UA + E, \qquad (2)$$

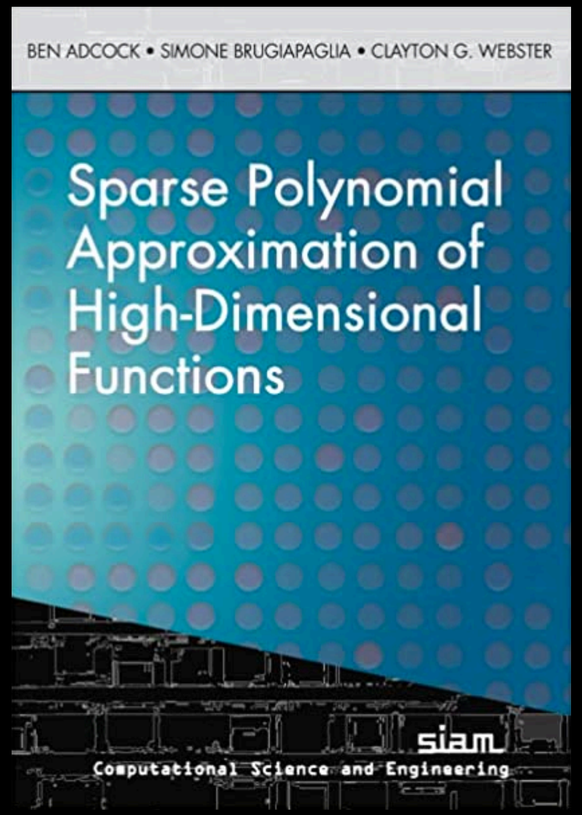
where Z is an $n \times m$ matrix in which each row contains measurements on an m-dimensional random response vector, U is a $n \times p$ observed matrix whose rows are the n measurements of a p-dimensional predictor, E is the zero mean noise matrix, and A is the unknown coefficient matrix. A predictor U_j is not present in this model if the j-th row of A is equal to zero. Using the vectorization operator vec, (2) can be written as $vec(Z) = (U \otimes I)vec(A) + vec(E)$. Thus, if one treats rows of A as groups, predictor selection in model (2) can be

The group **square-root lasso**: Theoretical properties and fast algorithms
F Bunea, J Lederer, Y She - IEEE Transactions on Information …, 2013 - ieeexplore.ieee.org
… **square-root lasso**, for sparse regression without group structure. Moreover, as a new type of result for **square-root l**… se needed by the **lasso** or group-**lasso**-type methods, but .
☆ Save 99 Cite Cit… 109 …lated articles All 9 versions »

The Square Root LASSO

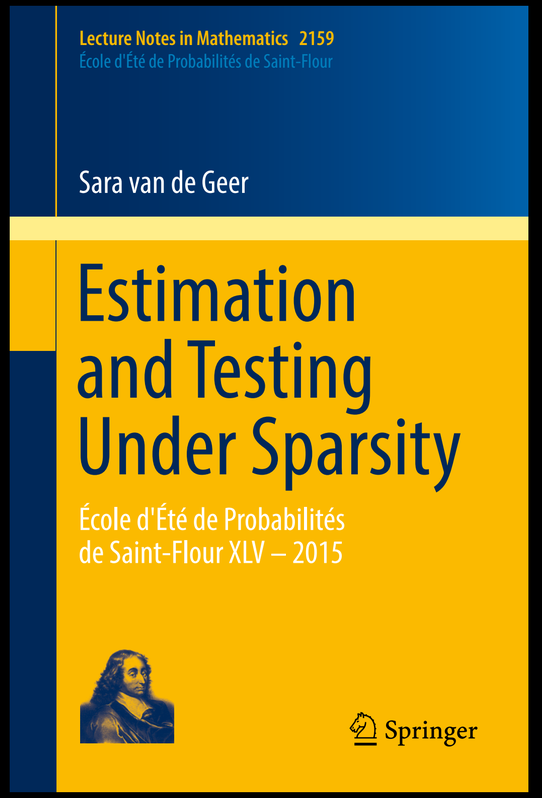
Applications

[Adcock, Brugiapaglia & Webster, 2022]



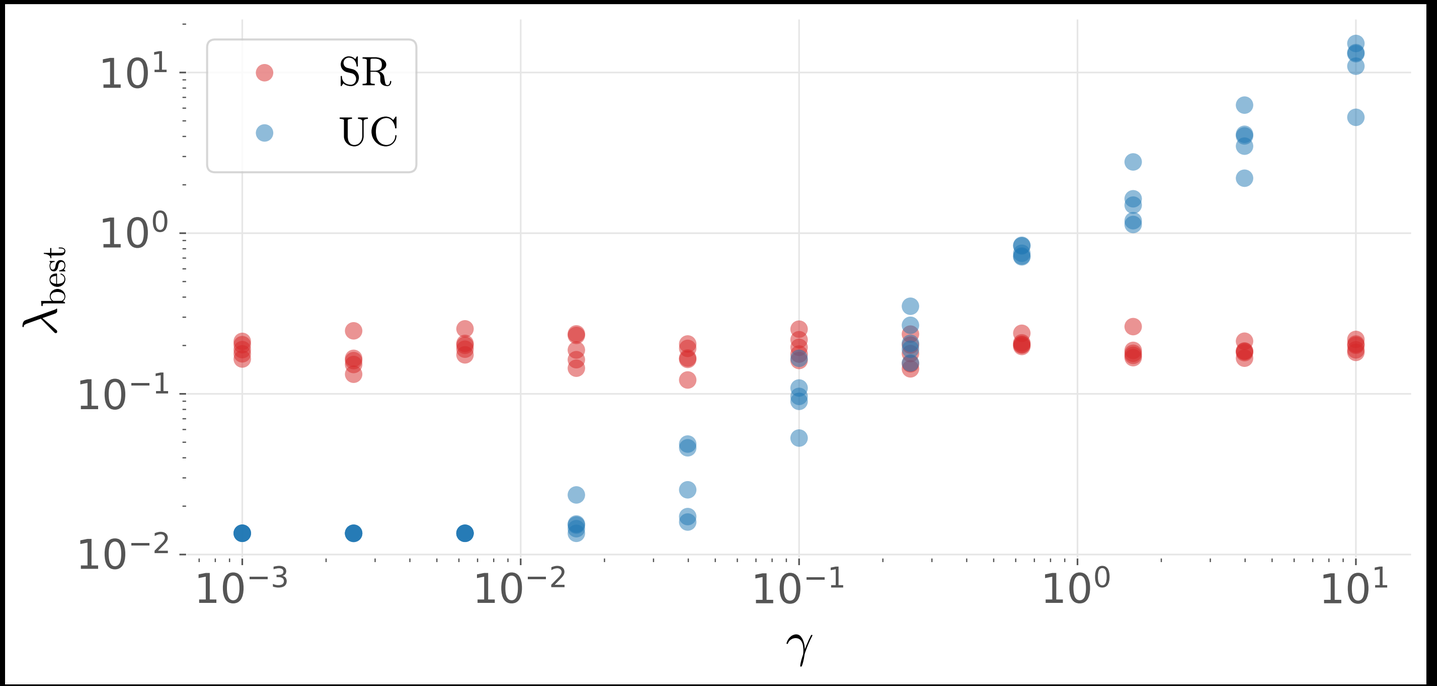
[Adcock & Hansen, 2021]

[van de Geer, 2015]

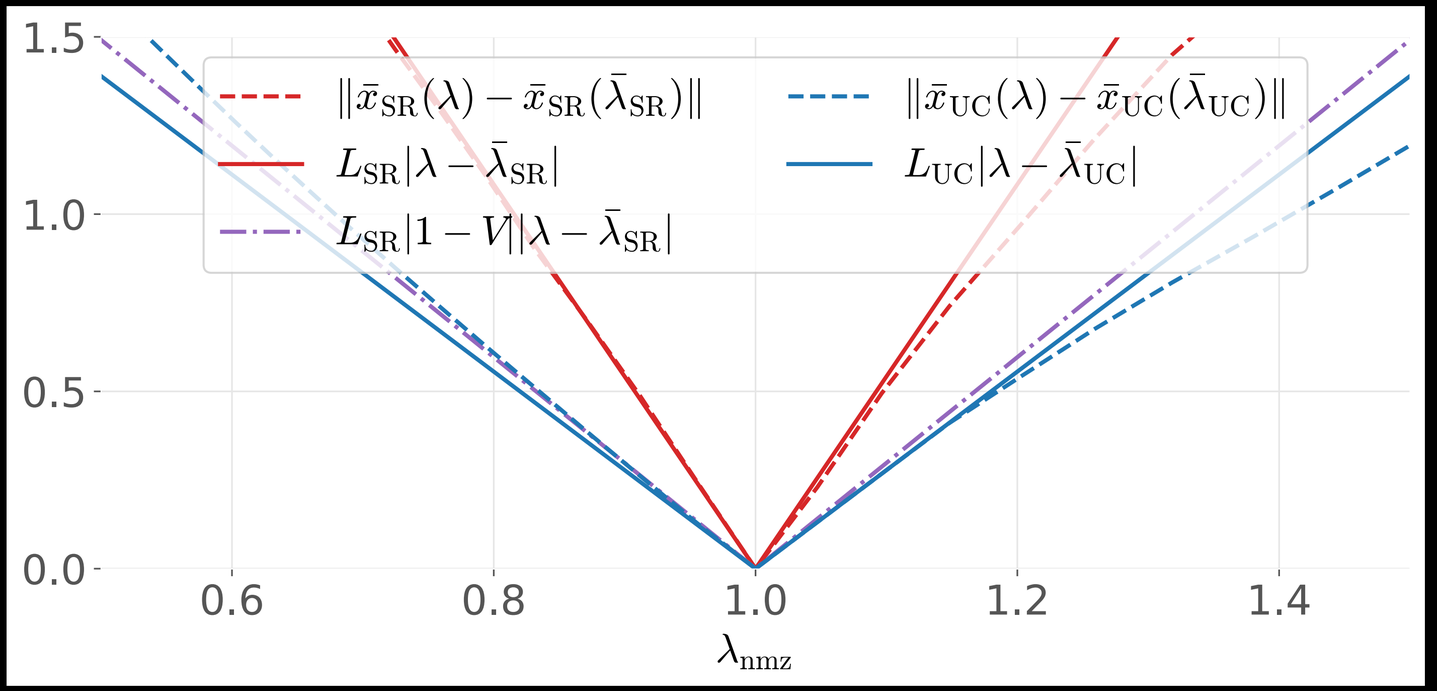


Motivation

Top:
Best parameter choice vs. noise scale. Five realizations each. (*m, n, s*) = (50, 100, 5)



Bottom:
Lipschitz behaviour for (SR) and (UC).
$L_{\mathrm{UC}} \approx L_{\mathrm{SR}}|1 - V|$



Outline

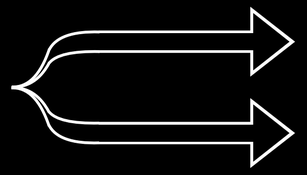
Implication ordering

$$S : (b, \lambda) \mapsto \arg\min_{x \in \mathbb{R}^n} \{\|Ax - b\| + \lambda \|x\|_1\}$$

The Strong Assumption

S is continuously differentiable at $(\bar{b}, \bar{\lambda})$

The Intermediate Assumption



S is Lipschitz and directionally differentiable at $(\bar{b}, \bar{\lambda})$

$S(\bar{b}, \bar{\lambda})$ admits an explicit formula

The Weak Assumption

$$S(\bar{b}, \bar{\lambda}) = \{\bar{x}\}$$

Technical background

Sherman-Morrison-Woodbury

Lemma (Sherman-Morrison-Woodbury). *Let* $M \in \mathbb{R}^{m \times s}$ *with* $\operatorname{rank} M = s$, *and* $v \in \mathbb{R}^m$ *with* $\|v\| = 1$. *For* $W := M^\top (\mathbb{I} - vv^\top) M$*:*

(a) W *is invertible if and only if* $v \notin \operatorname{rge} M$, *and*

$$W^{-1} = (M^\top M)^{-1} + \frac{M^\dagger v (M^\dagger v)^\top}{1 - v^\top M M^\dagger v}.$$

(b) In the invertible case, we have

$$\lambda_{\max}(W^{-1}) \leqslant \frac{1}{\sigma_{\min}(M)^2} + \frac{\|M^\dagger v\|^2}{1 - v^\top M M^\dagger v} < \frac{1}{\sigma_{\min}(M)^2} + \frac{1}{1 - v^\top M M^\dagger v}.$$

Technical background

Fenchel-Rockafellar duality

Proposition. *The (Fenchel-Rockafeller) dual problem to (SR) is*

$$\max_{y\in\mathbb{R}^m}\langle b,y\rangle \quad \text{such that} \quad A^\top y\in\lambda\mathbb{B}_\infty, \quad y\in\mathbb{B}. \tag{SR$^\circ$}$$

Moreover, the following are equivalent:

1. $\bar{x}$ *solves (SR),* $\bar{y}$ *solves (SR$^\circ$).*
2. $\|A\bar{x}-b\|+\lambda\|\bar{x}\|_1=\langle b,\bar{y}\rangle$ *and* $\bar{y}$ *is feasible for (SR$^\circ$).*
3. $A^\top\bar{y}\in\lambda\partial\|\cdot\|_1(\bar{x})$, $-\bar{y}\in\partial\|(\cdot)-b\|(A\bar{x})$.
4. $\bar{x}\in N_{\lambda\mathbb{B}_\infty}(A^\top\bar{y})$, $b-A\bar{x}\in N_{\mathbb{B}}(\bar{y})$.

Corollary. *If there exists a solution* $\hat{x}$ *of (SR) with* $A\hat{x}\neq b$*, then:*

1. *(SR$^\circ$) has a unique solution* $\bar{y}$ *with* $\|\bar{y}\|=1$.
2. *Every solution* $\bar{x}$ *of (SR) satisfies* $b-A\bar{x}=\|A\bar{x}-b\|\bar{y}$.

Technical background

Fenchel-Rockafellar duality

Proposition. *The (Fenchel-Rockafeller) dual problem to (SR) is*

$$\max_{y \in \mathbb{R}^m} \langle b, y \rangle \quad \text{such that} \quad A^\top y \in \lambda \mathbb{B}_\infty, \quad y \in \mathbb{B}. \tag{SR$^\circ$}$$

Moreover, the following are equivalent:

1. $\bar{x}$ *solves (SR),* $\bar{y}$ *solves (SR$^\circ$).*
2. $\|A\bar{x} - b\| + \lambda \|\bar{x}\|_1 = \langle b, \bar{y} \rangle$ *and* $\bar{y}$ *is feasible for (SR$^\circ$).*
3. $A^\top \bar{y} \in \lambda \partial \| \cdot \|_1(\bar{x})$, $-\bar{y} \in \partial \|(\cdot) - b\|(A\bar{x})$.
4. $\bar{x} \in N_{\lambda \mathbb{B}_\infty}(A^\top \bar{y})$, $b - A\bar{x} \in N_{\mathbb{B}}(\bar{y})$.

Optimality conditions!

Proof via Rockafellar & Wets (2009) Example 11.41 & invoking strong duality.

Corollary. *If there exists a solution* $\hat{x}$ *of (SR) with* $A\hat{x} \neq b$*, then:*

1. *(SR$^\circ$) has a unique solution* $\bar{y}$ *with* $\|\bar{y}\| = 1$.
2. *Every solution* $\bar{x}$ *of (SR) satisfies* $b - A\bar{x} = \|A\bar{x} - b\|\bar{y}$.

Technical background

Fenchel-Rockafellar duality

Proposition. *The (Fenchel-Rockafeller) dual problem to (SR) is*

$$\max_{y \in \mathbb{R}^m} \langle b, y \rangle \quad \text{such that} \quad A^\top y \in \lambda \mathbb{B}_\infty, \quad y \in \mathbb{B}. \tag{SR$^\circ$}$$

Moreover, the following are equivalent:

1. $\bar{x}$ *solves (SR),* $\bar{y}$ *solves (SR$^\circ$).*
2. $\|A\bar{x} - b\| + \lambda \|\bar{x}\|_1 = \langle b, \bar{y} \rangle$ *and* $\bar{y}$ *is feasible for (SR$^\circ$).*
3. $A^\top \bar{y} \in \lambda \partial \| \cdot \|_1(\bar{x})$, $-\bar{y} \in \partial \|(\cdot) - b\|(A\bar{x})$.
4. $\bar{x} \in N_{\lambda \mathbb{B}_\infty}(A^\top \bar{y})$, $b - A\bar{x} \in N_{\mathbb{B}}(\bar{y})$.

Optimality conditions!

Proof via Rockafellar & Wets (2009) Example 11.41 & invoking strong duality.

Corollary. *If there exists a solution* $\hat{x}$ *of (SR) with* $A\hat{x} \neq b$*, then:*

1. *(SR$^\circ$) has a unique solution* $\bar{y}$ *with* $\|\bar{y}\| = 1$.
2. *Every solution* $\bar{x}$ *of (SR) satisfies* $b - A\bar{x} = \|A\bar{x} - b\|\bar{y}$.

Uniqueness of the solution mapping

A sufficient condition

Assumption 1 (Weak). *For a solution $\bar{x}$ of (SR) with $I := \operatorname{supp}(\bar{x})$ we have*

1. $\ker A_I = \{0\}$ *and* $b \notin \operatorname{rge} A_I$;
2. $\exists z \in \{\bar{y}\}^{\perp} \cap \ker A_I^{\top}$ *such that* $\left\| A_{I^C}^{\top} (\bar{y} + z) \right\|_{\infty} < \lambda$.

Theorem. *Under The Weak Assumption, $\bar{x}$ is the unique minimizer of (SR).*

Uniqueness of the solution mapping

A sufficient condition

Assumption 1 (Weak). *For a solution $\bar{x}$ of (SR) with $I := \operatorname{supp}(\bar{x})$ we have*

1. $\ker A_I = \{0\}$ *and* $b \notin \operatorname{rge} A_I$;
2. $\exists z \in \{\bar{y}\}^\perp \cap \ker A_I^\top$ *such that* $\left\|A_{I^C}^\top (\bar{y} + z)\right\|_\infty < \lambda$.

Theorem. *Under The Weak Assumption, $\bar{x}$ is the unique minimizer of (SR).*

Proof idea: Let $\mathcal{S} := N_{\mathbb{B}}(\bar{y}) = \mathbb{R}_+\{\bar{y}\}$. Fact: every solution $\bar{x}$ of (SR) solves

$$\min_{x \in \mathbb{R}^m} \psi(x) := \lambda \|x\|_1 - \langle A^\top \bar{y}, x \rangle + \delta_{\mathcal{S}}(b - Ax). \tag{aux}$$

Then, (aux) is polyhedral convex so by (Lemma 2.2, Gilbert, 2017):

$$\begin{aligned} \bar{x} \text{ uniquely solves (aux)} &\iff 0 \in \operatorname{int} \partial\psi(\bar{x}) \\ &\iff \operatorname{span} \partial\psi(\bar{x}) = \mathbb{R}^n \text{ and } 0 \in \operatorname{ri} \partial\psi(\bar{x}). \end{aligned}$$

Uniqueness of the solution mapping

A sufficient condition

¿Necessary?

Assumption 1 (Weak). *For a solution $\bar{x}$ of (SR) with $I := \operatorname{supp}(\bar{x})$ we have*

1. $\ker A_I = \{0\}$ *and* $b \notin \operatorname{rge} A_I$;
2. $\exists z \in \{\bar{y}\}^\perp \cap \ker A_I^\top$ *such that* $\left\|A_{I^C}^\top (\bar{y} + z)\right\|_\infty < \lambda$.

Theorem. *Under The Weak Assumption, $\bar{x}$ is the unique minimizer of (SR).*

Proof idea: Let $\mathcal{S} := N_{\mathbb{B}}(\bar{y}) = \mathbb{R}_+\{\bar{y}\}$. Fact: every solution $\bar{x}$ of (SR) solves

$$\min_{x \in \mathbb{R}^m} \psi(x) := \lambda \|x\|_1 - \langle A^\top \bar{y}, x \rangle + \delta_{\mathcal{S}}(b - Ax). \qquad \text{(aux)}$$

Then, (aux) is polyhedral convex so by (Lemma 2.2, Gilbert, 2017):

$$\begin{aligned} \bar{x} \text{ uniquely solves (aux)} &\iff 0 \in \operatorname{int} \partial\psi(\bar{x}) \\ &\iff \operatorname{span} \partial\psi(\bar{x}) = \mathbb{R}^n \text{ and } 0 \in \operatorname{ri} \partial\psi(\bar{x}). \end{aligned}$$

Explicit solution formula

Under the Intermediate assumption

Assumption 2 (Intermediate)**.** *For a minimizer $\bar{x}$ of (SR) we have $A\bar{x} \neq b$ and for*

$$J := \left\{ i \in [n] : \left| A_i^\top \frac{b - A\bar{x}}{\|A\bar{x} - b\|} \right| = \lambda \right\},$$

we have $\ker A_J = \{0\}$ *and* $b \notin \operatorname{rge} A_J$.

Proposition. *The Intermediate Assumption implies The Weak Assumption.*

Proposition. *Let $\bar{x}$ be a solution of (SR) such that The Intermediate Assumption holds at $\bar{x}$. Then $\bar{x}$ is the unique solution and*

$$\bar{x} = L_J \left(B A_J^\top (\mathbb{I} - \bar{y}\bar{y}^\top) b \right), \qquad B := \left[A_J^\top (\mathbb{I} - \bar{y}\bar{y}^\top) A_J \right]^{-1}.$$

Explicit solution formula

Under the Intermediate assumption

Assumption 2 (Intermediate)**.** *For a minimizer $\bar{x}$ of (SR) we have $A\bar{x} \neq b$ and for*

$$J := \left\{ i \in [n] : \left| A_i^\top \frac{b - A\bar{x}}{\|A\bar{x} - b\|} \right| = \lambda \right\},$$

we have $\ker A_J = \{0\}$ *and* $b \notin \operatorname{rge} A_J$.

Proposition. *The Intermediate Assumption implies The Weak Assumption.*

Proposition. *Let $\bar{x}$ be a solution of (SR) such that The Intermediate Assumption holds at $\bar{x}$. Then $\bar{x}$ is the unique solution and*

$$\bar{x} = L_J \left(B A_J^\top (\mathbb{I} - \bar{y}\bar{y}^\top) b \right), \qquad B := \left[A_J^\top (\mathbb{I} - \bar{y}\bar{y}^\top) A_J \right]^{-1}.$$

Lipschitzness and directional differentiability

Under the Intermediate assumption

Theorem. *Let $(\bar{b}, \bar{\lambda}) \in \mathbb{R}^m \times \mathbb{R}_{++}$ and suppose The Intermediate Assumption holds at $\bar{x} := S(\bar{b}, \bar{\lambda})$.*

1. *S is locally Lipschitz at $(\bar{b}, \bar{\lambda})$ with (local) Lipschitz modulus*

$$L \leqslant \left[\frac{1}{\sigma_{\min}(A_J)^2} + \frac{1}{1 - \|A_J A_J^\dagger \bar{y}\|} \right] \cdot \left[\sigma_{\max}(A_J) + \left\| \frac{A_J^\top (A\bar{x} - \bar{b})}{\bar{\lambda}} \right\| \right].$$

2. *S is directionally differentiable at $(\bar{b}, \bar{\lambda})$ and the directional derivative $S'((\bar{b}, \bar{\lambda}); (\cdot, \cdot)) : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^n$ is locally Lipschitz. Moreover, for $(q, \alpha) \in \mathbb{R}^m \times \mathbb{R}$ there exists $K = K(q, \alpha) \subseteq J$ with $\operatorname{supp}(\bar{x}) \subseteq K$ such that*

$$S'((\bar{b}, \bar{\lambda}); (q, \alpha)) = L_K \left(B \left(A_K^\top (\mathbb{I} - \bar{y}\bar{y}^\top) q + \frac{\alpha}{\lambda} A_K^\top (A\bar{x} - \bar{b}) \right) \right),$$

where $B := \left(A_K^\top A_K\right)^{-1} + \dfrac{A_K^\dagger \bar{y} (A_K^\dagger \bar{y})^\top}{1 - \bar{y}^\top A_K A_K^\dagger \bar{y}}.$

Lipschitzness and directional differentiability

Under the Intermediate assumption

Theorem. *Let $(\bar{b}, \bar{\lambda}) \in \mathbb{R}^m \times \mathbb{R}_{++}$ and suppose The Intermediate Assumption holds at $\bar{x} := S(\bar{b}, \bar{\lambda})$.*

1. *S is locally Lipschitz at $(\bar{b}, \bar{\lambda})$ with (local) Lipschitz modulus*

$$L \leqslant \left[\frac{1}{\sigma_{\min}(A_J)^2} + \frac{1}{1 - \|A_J A_J^\dagger \bar{y}\|} \right] \cdot \left[\sigma_{\max}(A_J) + \left\| \frac{A_J^\top (A\bar{x} - \bar{b})}{\bar{\lambda}} \right\| \right].$$

2. *S is directionally differentiable at $(\bar{b}, \bar{\lambda})$ and the directional derivative $S'((\bar{b}, \bar{\lambda}); (\cdot, \cdot)) : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^n$ is locally Lipschitz. Moreover, for $(q, \alpha) \in \mathbb{R}^m \times \mathbb{R}$ there exists $K = K(q, \alpha) \subseteq J$ with $\operatorname{supp}(\bar{x}) \subseteq K$ such that*

$$S'((\bar{b}, \bar{\lambda}); (q, \alpha)) = L_K \left(B \left(A_K^\top (\mathbb{I} - \bar{y}\bar{y}^\top) q + \frac{\alpha}{\lambda} A_K^\top (A\bar{x} - \bar{b}) \right) \right),$$

where $B := \left(A_K^\top A_K\right)^{-1} + \dfrac{A_K^\dagger \bar{y} (A_K^\dagger \bar{y})^\top}{1 - \bar{y}^\top A_K A_K^\dagger \bar{y}}.$

Continuous differentiability

The Strong assumption

Assumption 3 (Strong)**.** *For a minimizer* $\bar{x}$ *of (SR) with* $I := \operatorname{supp}(\bar{x})$*:*

1. $\ker A_I = \{0\}$ *and* $b \notin \operatorname{rge} A_I$*;*
2. $\|A_{I^C}^\top (b - A\bar{x})\|_\infty < \lambda \|b - A\bar{x}\|$.

Proposition. *The Strong Assumption implies The Intermediate Assumption.*

Proof. The Strong Assumption implies $I = J$, so $A_I = A_J$ has full rank.

Continuous differentiability

Under the Strong assumption

Theorem. *For $(\bar{b}, \bar{\lambda}) \in \mathbb{R}^m \times \mathbb{R}_{++}$, let $\bar{x} \in \mathbb{R}^n$ be a solution of (SR) with $I := \operatorname{supp}(\bar{x})$ such that The Strong Assumption holds. Then the solution map $S(b, \lambda)$ is continuously differentiable at $(\bar{b}, \bar{\lambda})$ with derivative*

$$DS(\bar{b}, \bar{\lambda})(q, \alpha) = L_I \left(\left[\left(A_I^\top A_I\right)^{-1} + \frac{A_I^\dagger \bar{y} (A_I^\dagger \bar{y})^\top}{1 - \bar{y}^\top A_I A_I^\dagger \bar{y}} \right] \cdot \left[A_I^\top (\mathbb{I} - \bar{y}\bar{y}^\top) q - \frac{\alpha}{\bar{\lambda}} A_I^\top \bar{r} \right] \right),$$

for $\bar{r} := \bar{b} - A\bar{x}$, $\bar{y} := \bar{r}/\|\bar{r}\|$. In particular, S is locally Lipschitz at $(\bar{b}, \bar{\lambda})$ with constant

$$L \leqslant \left[\frac{1}{\sigma_{\min}(A_I)^2} + \frac{1}{1 - \|A_I A_I^\dagger \bar{y}\|} \right] \cdot \left[\sigma_{\max}(A_I) + \left\| \frac{A_I^\top \bar{r}}{\bar{\lambda}} \right\| \right].$$

Continuous differentiability

Under the Strong assumption

Corollary. *For $(\bar{b}, \bar{\lambda}) \in \mathbb{R}^m \times \mathbb{R}_{++}$ let $\bar{x}$ be a solution of (SR) such that The Strong Assumption holds and let $I := \operatorname{supp}(\bar{x})$. Then the solution map $S(\lambda)$ is continuously differentiable at $\bar{\lambda}$ with derivative*

$$DS(\bar{\lambda})(\alpha) = \frac{\alpha}{\bar{\lambda}} \left[A_I^\dagger \bar{r} + \frac{A_I^\dagger \bar{y} (A_I^\dagger \bar{y})^\top A_I^\top \bar{r}}{1 - \bar{y}^\top A_I A_I^\dagger \bar{y}} \right], \quad \forall \alpha \in \mathbb{R},$$

where $\bar{r} := \bar{b} - A\bar{x}$, $\bar{y} := \bar{r}/\|\bar{r}\|$. In particular, S is locally Lipschitz at $\bar{\lambda}$ with constant

$$L \leqslant \frac{1}{\bar{\lambda}} \left\| A_I^\dagger \bar{r} \right\| \cdot |1 - V|^{-1} \leqslant \frac{\|A\bar{x} - \bar{b}\|}{\bar{\lambda} \cdot \sigma_{\min}(A_I) \cdot |1 - V|}, \quad V := \bar{y}^\top A_I A_I^\dagger \bar{y}.$$

Lipschitz sensitivity: (SR) vs. (UC)

A comparison of Lipschitz constants

$$L_{\mathrm{SR}} \leqslant \frac{1}{\lambda} \left\| A^{\dagger}_{I_{\mathrm{SR}}} \left(A\bar{x}_{\mathrm{SR}} - b \right) \right\| \cdot \left| 1 - \bar{y}^{\top} A_{I_{\mathrm{SR}}} A^{\dagger}_{I_{\mathrm{SR}}} \bar{y} \right|^{-1}$$

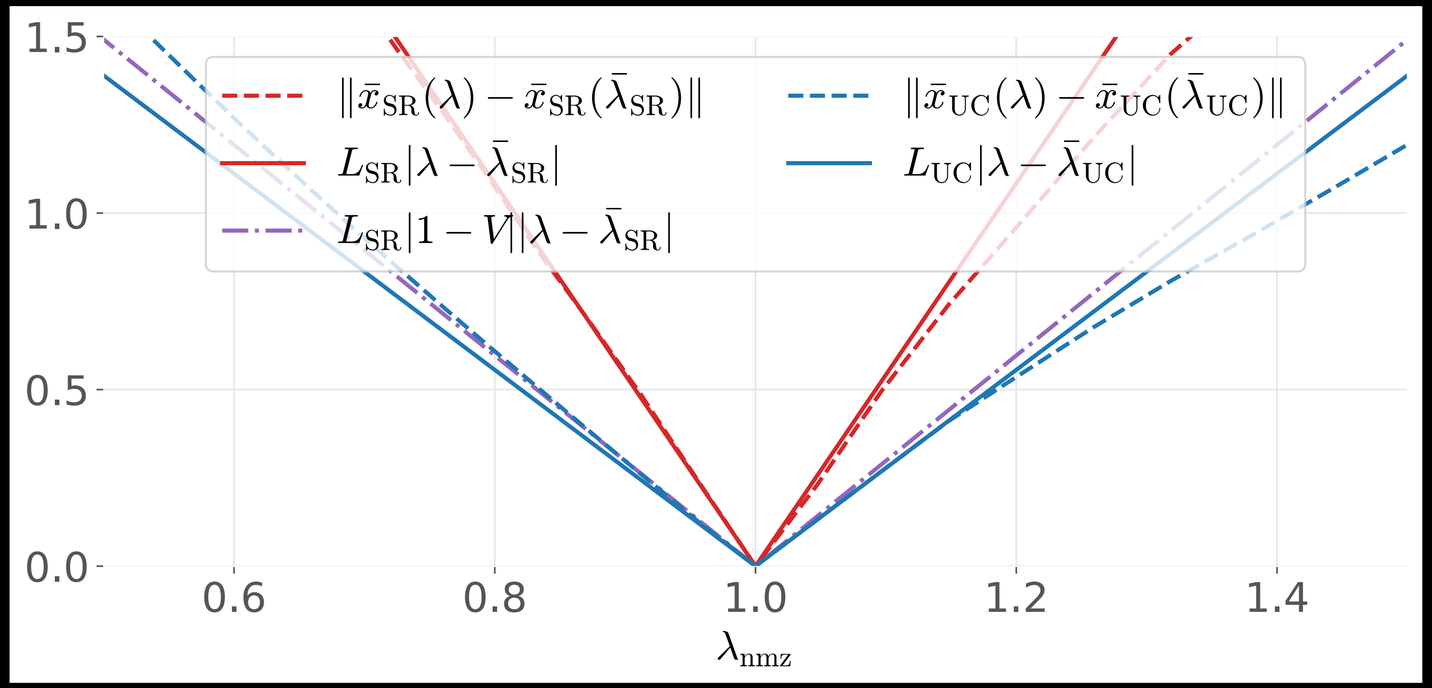
$$L_{\mathrm{UC}} \leqslant \frac{1}{\lambda} \left\| A^{\dagger}_{I_{\mathrm{UC}}} \left(A\bar{x}_{\mathrm{UC}} - b \right) \right\|$$

V

[Berk, Brugiapaglia & Hoheisel, 2021]

Figure:
Lipschitz behaviour for (SR) and (UC).

$$L_{\mathrm{UC}} \approx L_{\mathrm{SR}} |1 - V|$$

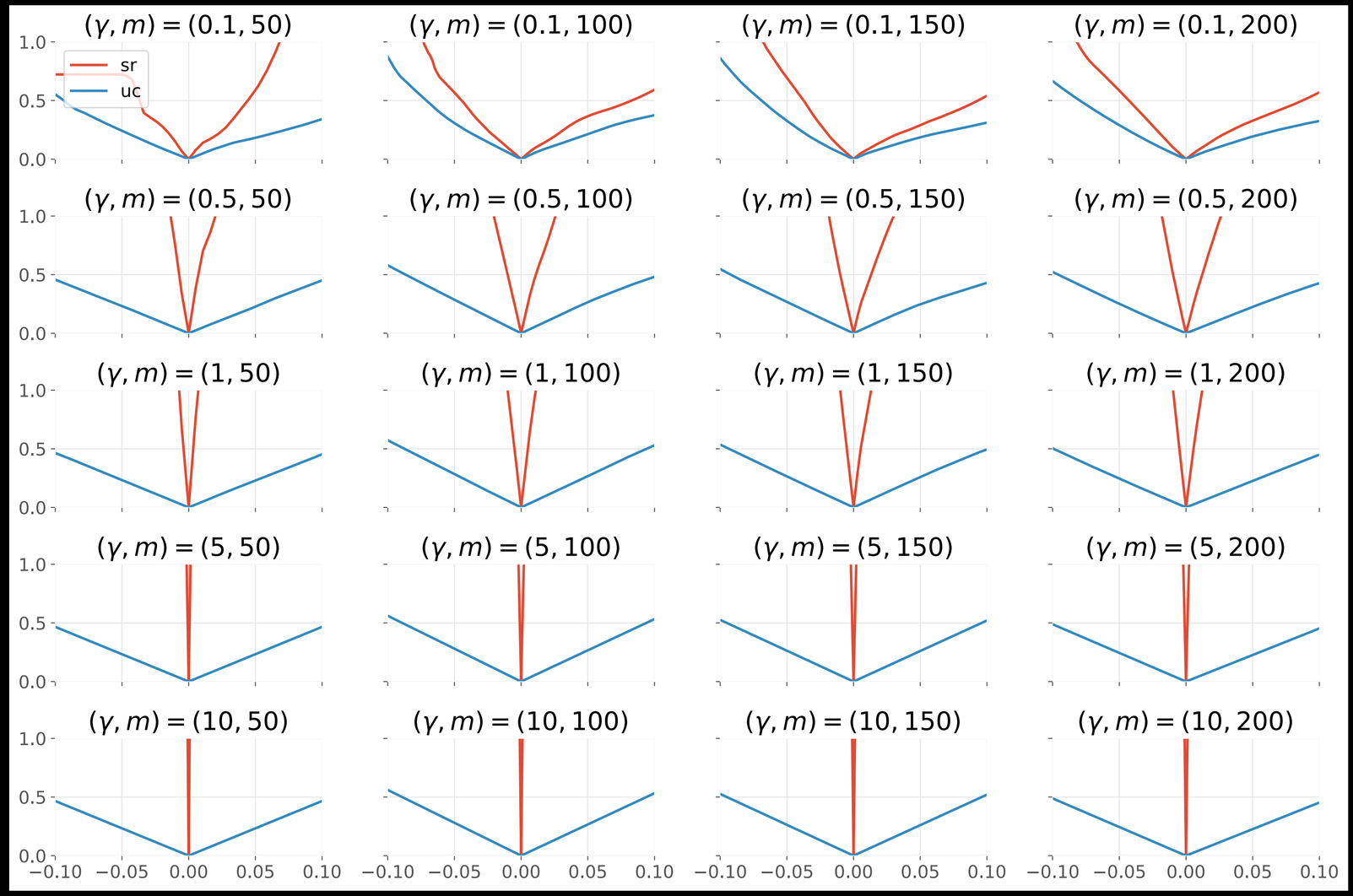


Lipschitz sensitivity: (SR) vs. (UC)

Numerical results

Effect of noise scale on error sensitivity for (UC) and (SR); (n, s) = (200, 7).

$\lambda - \bar{\lambda}_{\text{best}}$ vs. $\|\bar{x}(\lambda) - \bar{x}(\bar{\lambda}_{\text{best}})\|$



Numerical results

Experimental setup I

Assumption 1 (Weak). *For a solution* $\bar{x}$ *of (SR) with* $I := \operatorname{supp}(\bar{x})$ *we have*

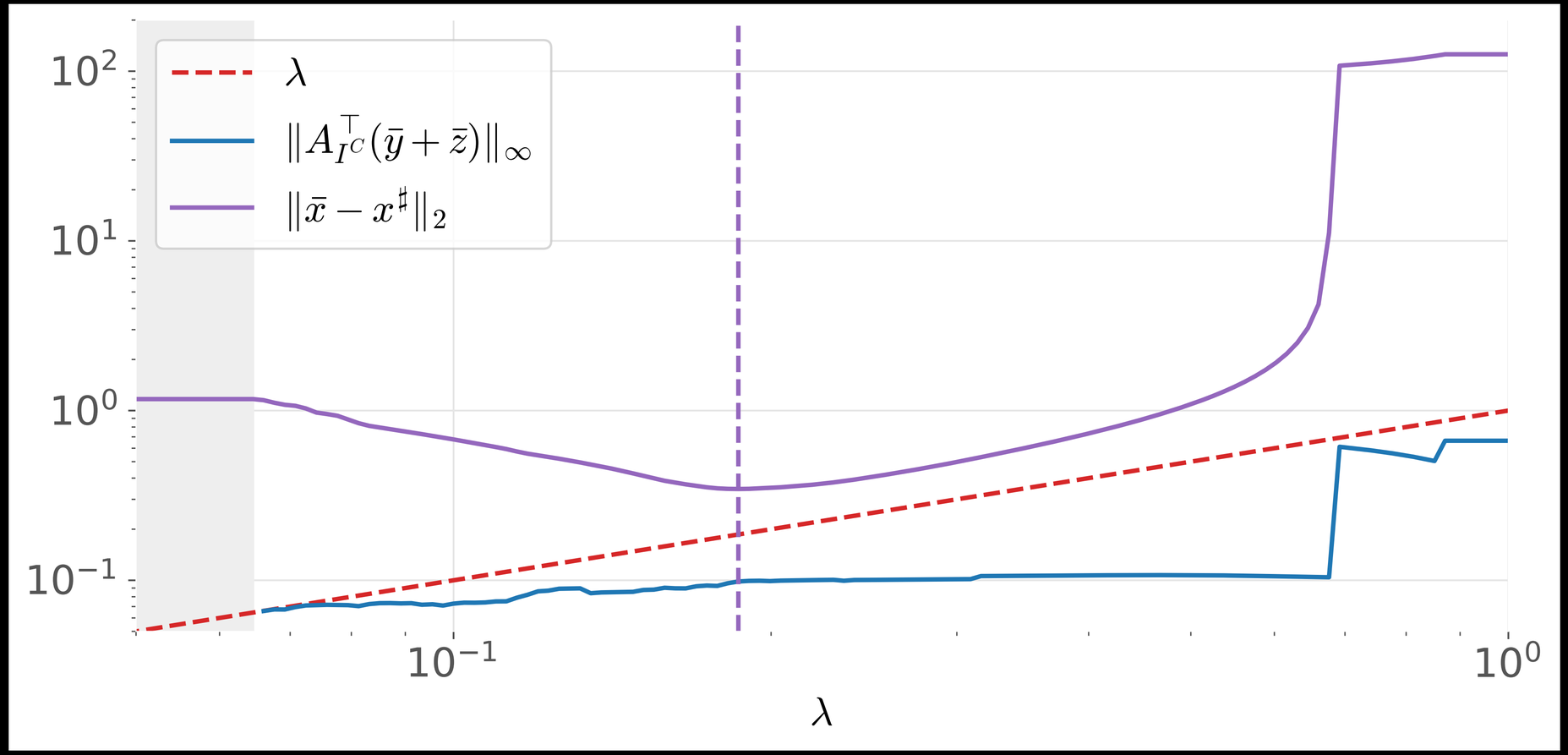
1. $\ker A_I = \{0\}$ *and* $b \notin \operatorname{rge} A_I$;
2. $\exists z \in \{\bar{y}\}^{\perp} \cap \ker A_I^{\top}$ *such that* $\left\|A_{I^C}^{\top}(\bar{y} + z)\right\|_{\infty} < \lambda$.

$$Z^* := \min_{z \in \mathbb{R}^m} \left\{ \|A_{I^C}^{\top}(\bar{y} + z)\|_{\infty} : [A_I \ \bar{y}]^{\top} z = 0 \right\}$$

Observe that The Weak Assuption (ii) is satisfied if $Z^* < \lambda$.

Numerical results

Phase transition plot



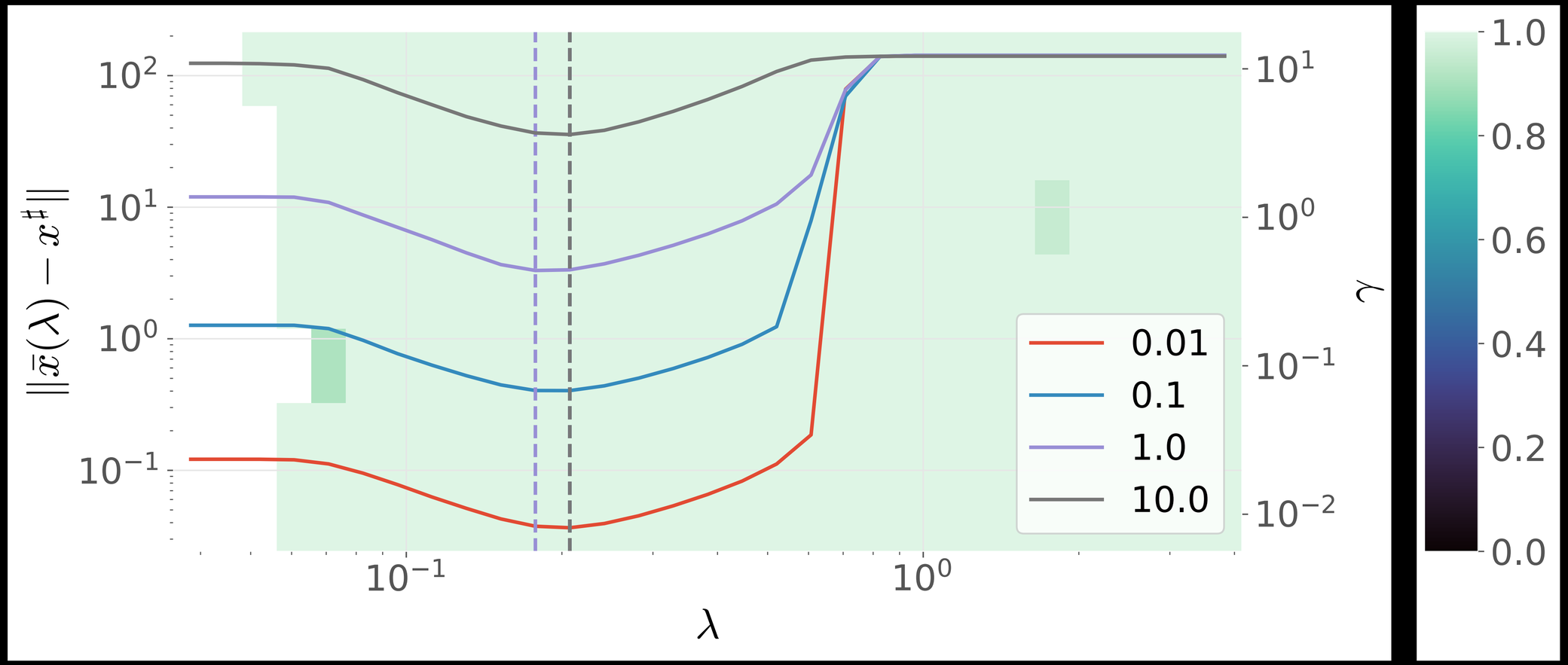
Background: Visualization of sufficiency for (SR) uniqueness & error vs. λ.

(m, n, s, γ) = (100, 200, 2, 0.1). $\bar{z} \in \underset{z \in \mathbb{R}^m}{\arg\min} \left\{ \|A_{I^C}^\top(\bar{y}+z)\|_\infty : [A_I \ \bar{y}]^\top z = 0 \right\}$

Numerical results

Phase transition plot (λ & γ)

$$Z^* := \min_{z \in \mathbb{R}^m} \left\{ \|A_{I^C}^\top (\bar{y} + z)\|_\infty : [A_I \ \bar{y}]^\top z = 0 \right\}$$



Background: Proportion of 20 independent trials for which $Z^* < \lambda$
as a function of λ and γ. **Foreground:** ℓ_2 error vs. λ for 4 different noise scales γ.

Conclusion

Summary

- Convex analytic approach for establishing (SR) uniqueness
- Variational analytic characterization of smoothness of solution mapping under natural assumptions
- Apparent trade-off between oracular agnosticism and parameter sensitivity

Conclusion

Summary and future directions

? Is The Weak Assumption also necessary for uniqueness?

? Sensitivity characterization with respect to a ground truth?

? Analysis for exactly solvable setting?

? Analysis in non-unique/set-valued setting via *Aubin property*?

- Straightforward extension to analyze (A, b, λ).
- Compressed sensing applications for subgaussian measurement matrices.

Thank you!

Extra technical background

Fenchel-Rockafellar duality

Theorem (Theorem 12.2, Rockafellar, 1970). *Let f be a convex function. The conjugate function f^* is proper convex lsc if and only if f is proper, where*

$$f^*(y) := \sup_x \{\langle x, y\rangle - f(x)\}.$$

Theorem (Corollary 31.2.1, Rockafellar, 1970). *Let $f, g : \mathbb{R}^n \to \mathbb{R}_\infty$ be proper convex lsc and $A \in \mathbb{R}^{m\times n}$. Then,*

$$\inf_{x\in\mathbb{R}^n} \{f(x) + g(Ax)\} = \sup_{y\in\mathbb{R}^m} \left\{-g^*(-y) - f^*(A^\top y)\right\}$$

if either of the following conditions is satisfied:

(a) $A \operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g \neq \varnothing$;

(b) $A^\top \operatorname{ri}\operatorname{dom} g^* \cap \operatorname{ri}\operatorname{dom} f^* \neq \varnothing$.

Under (a) the supremum is attained at some y, while under (b) the infimum is attained at some x.

Extra technical background

Polyhedral convex minimization

Lemma (Lemma 2.1, Gilbert, 2017)**.** *Suppose* $f : \mathbb{R}^n \to \mathbb{R}$ *is polyhedral convex. Then*

$$\bar{x} \in \operatorname{ri} \arg\min f \iff 0 \in \operatorname{ri} \partial f(\bar{x}).$$

Lemma (Lemma 2.2, Gilbert, 2017)**.** *Suppose* $f : \mathbb{R}^n \to \mathbb{R}$ *is polyhedral convex. Then*

$$\arg\min f = \{\bar{x}\} \iff 0 \in \operatorname{int} \partial f(\bar{x}).$$

Explicit solution formula for (UC)

Under an analogous Intermediate assumption

Lemma (Lemma 2, Tibshirani, 2013). *For any b, A and $\lambda > 0$, if $\ker A_{\mathcal{E}} = \{0\}$, then $\bar{x}$ solving (UC) is unique and*

$$\bar{x} = L_{\mathcal{E}} \left((A_{\mathcal{E}}^\top A_{\mathcal{E}})^{-1} (A_{\mathcal{E}}^\top b) - \lambda v \right),$$

where v is a subgradient and $\mathcal{E}$ is the (UC) equicorrelation set:

$$\mathcal{E} := \left\{ i \in [n] : \left| A_i^\top (b - A\bar{x}) \right| = \lambda \right\}$$
$$v := \operatorname{sign}\left(A_{\mathcal{E}}^\top (b - A\bar{x}) \right).$$

Main proof idea(s)

For smoothness results

Apply a variational analysis "hammer" in from [Berk, Brugiapaglia & Hoheisel, 2022] — relies on:

- coderivative calculus [Dontchev & Rockafellar, 2014],
- the Mordukhovich criterion [Theorem 3.3(ii), Mordukhovich, 2018] and
- [Theorem 9.56, Rockafellar & Wets, 1998].

Proposition (Proposition 4.11, Berk, Brugiapaglia & Hoheisel, 2022). *Let $(\bar{p}, \bar{x}) \in \mathbb{R}^d \times \mathbb{R}^n$, let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be monotone (locally around $\bar{x}$) and let $f : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable at $(\bar{p}, \bar{x})$ such that $f(\bar{p}, \cdot)$ is monotone (locally at $\bar{x}$). Define $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ by*

$$S(p) = \{x \in \mathbb{R}^n : 0 \in f(p, x) + F(x)\}, \quad \forall p \in \mathbb{R}^d.$$

Assume $(\bar{p}, \bar{x}) \in \operatorname{gph} S$ (i.e., $0 \in f(\bar{p}, \bar{x}) + F(\bar{x})$) and $Q : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ given by $Q := f(\bar{p}, \cdot) + F$ has closed graph with $(\bar{x}, 0) \in \operatorname{gph} Q$ and $\ker D^*Q(\bar{x} \mid 0) = \{0\}$. Then, the following hold:

(a) Q is strongly metrically regular *at $(\bar{x}, 0) \in \operatorname{gph} Q$.*

(b) S is locally Lipschitz at $\bar{p}$.

(c) If F is proto-differentiable *at $(\bar{x}, -f(\bar{p}, \bar{x}))$, then the graphical derivative $DS(\bar{p}|\bar{x})$ is single-valued and locally Lipschitz with*

$$DS(\bar{p})(q) = \{w \in \mathbb{R}^n : 0 \in DG(\bar{p}, \bar{x}|0)(q, w)\}, \quad \forall q \in \mathbb{R}^d,$$

for $G(p, x) := f(p, x) + F(x)$. In particular, S is directionally differentiable at $\bar{p}$ with directional derivative

$$S'(\bar{p}; \cdot) = DS(\bar{p})(\cdot).$$

In addition, S is locally Lipschitz at $\bar{p}$ with modulus

$$L \leqslant \limsup_{p \to \bar{p}} \max_{\|q\| \leqslant 1} \|DS(p)(q)\|.$$

If $DS(\bar{p})$ is linear, then S is differentiable at $\bar{p}$ and the derivative equals the graphical derivative $DS(\bar{p})$.