Optimization and Algorithms in Sparse Regression

Optimization and Algorithms in Sparse Regression

Screening Rules, Coordinate Descent, and Normalization

Johan Larsson

LUND
UNIVERSITY

Thesis for the degree of Doctor of Philosophy

Thesis Advisors
Jonas Wallin and Małgorzata Bogdan

Faculty Opponent
Professor Mário A. T. Figueiredo (Instituto Superior Técnico, Lisbon, Portugal)

To be presented, with the permission of the Lund University School of Economics and Business Administration of Lund University, for public criticism in the Clark Kent lecture hall (Kentsalen) at the Department of Statistics on Sunday, the 34th of December 2024 at 24:00.

Organization
LUND UNIVERSITY
Department of Statistics
Box 7080
SE–220 07 Lund
Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2024-05-24

Sponsoring organization

Author(s)
Johan Larsson

Title and subtitle
Optimization and Algorithms in Sparse Regression: Screening Rules, Coordinate Descent, and Normalization

Abstract
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Key words
power, victory, awesomeness

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
978-91-8104-076-0 (print)
978-91-8104-077-7 (pdf)

Recipient's notes

Number of pages
50

Price

Security classification

Optimization and Algorithms in Sparse Regression

Screening Rules, Coordinate Descent, and Normalization

Johan Larsson

LUND
UNIVERSITY

A doctoral thesis at a university in Sweden takes either the form of a single, cohesive research study (monograph) or a summary of research papers (compilation thesis), which the doctoral student has written alone or together with one or several other author(s).

In the latter case the thesis consists of two parts. An introductory text puts the research work into context and summarizes the main points of the papers. Then, the research publications themselves are reproduced, together with a description of the individual contributions of the authors. The research papers may either have been already published or are manuscripts at various stages (in press, submitted, or in draft).

Cover illustration front: The elastic net path for a data set of diabetes patients.

Cover illustration back: Picture showing my research (Paper v).

Funding information: The thesis work was financially supported by my rich uncle.

Lund University School of Economics and Business Administration, Department of Statistics

ISBN: 978-91-8104-076-0 (print)
ISBN: 978-91-8104-077-7 (electronic)

Printed in Sweden by Media-Tryck, Lund University, Lund 2024

There's a point when you go with what you've got.
Or you don't go.

—Joan Didion

Contents

Acknowledgements

I owe my deepest gratitude to my supervisors, Jonas Wallin and Małgorzata Bogdan, who have been attentive and supportive throughout the ordeal.

I also want to thank my colleagues, especially my fellow PhD students, for the camaraderie and the support.

Most of all, I want to thank my family. My parents, my siblings, my partner, and my children. For raising and bearing with me though the work and the stress.

Abstract

Need more languages? Go to preamble.tex and add them to the usepackage[…]babel line. Install the corresponding packages on your system.

List of Publications

This thesis is based on the following publications.

I Johan Larsson, Małgorzata Bogdan, and Jonas Wallin. "The Strong Screening Rule for SLOPE". in: *Advances in Neural Information Processing Systems 33*. 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. Vol. 33. Virtual: Curran Associates, Inc., Dec. 6, 2020–12, pp. 14592–14603. ISBN: 978-1-71382-954-6

II Johan Larsson. "Look-Ahead Screening Rules for the Lasso". In: *22nd European Young Statisticians Meeting - Proceedings*. 22nd European Young Statisticians Meeting. Ed. by Andreas Makridis, Fotios S. Milienos, Panagiotis Papastamoulis, Christina Parpoula, and Athanasios Rakitzis. Athens, Greece: Panteion University of Social and Political Sciences, Sept. 6, 2021, pp. 61–65. ISBN: 978-960-7943-23-1

III Johan Larsson and Jonas Wallin. "The Hessian Screening Rule". In: *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. New Orleans, USA: Curran Associates, Inc., Nov. 28–Dec. 9, 2022, pp. 15823–15835. ISBN: 978-1-71387-108-8

IV Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagréou, Tom Dupré la Tour, Ghislain Durif, Cassio F. Dantas, Quentin Klopfenstein, Johan Larsson, En Lai, Tanguy Lefort, Benoit Malézieux, Badr Moufad, Binh T. Nguyen, Alain Rakotomamonjy, Zaccharie Ramzi, Joseph Salmon, and Samuel Vaiter. "Benchopt: Reproducible, Efficient and Collaborative Optimization Benchmarks". In: *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by S. Koyejo, S.

Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. New Orleans, USA: Curran Associates, Inc., Nov. 28–Dec. 9, 2022, pp. 25404–25421. ISBN: 978-1-71387-108-8

v Johan Larsson, Quentin Klopfenstein, Mathurin Massias, and Jonas Wallin. "Coordinate Descent for SLOPE". in: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. AISTATS 2023. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. Valencia, Spain: PMLR, Apr. 25–27, 2023, pp. 4802–4821

vi PAPER ON THE WAY

Introduction

Everything should be made as simple as possible, but not simpler.
—Albert Einstein

1 Background

With modern advances in science and technology, statistical models and the data on which they are fit are becoming increasingly complex. In many disciplines, such as bioinformatics, it is common to measure observations across many more variables than was previously the case. And so data sets are growing in size and becoming ever more high-dimensional. In some fields, this growth in complexity has been paralleled with more effective methods with which to collect observations, as in, for instance, the field of crowd science. But in other areas this this is still a costly endeavor. In bioinformatics, for example, ethical concerns and rising requirements on the quality of data have only raised the costs of data collection. As a result, the data collected in these fields is becoming *wider*: the ratio between the number of variables and the number of observations is increasing. This has challenged classical statistical methods such as classical linear regression, which break down in this setting. And even if there are indeed methods that can be used to model such data, without reducing the number of parameters, the resulting models are often not interpretable. This is the problem that sparse regression methods attempt to solve.

Sparse regression methods select only a subset of the features[1] in the data set, setting the regression coefficients of the remaining ones to zero. The validity of doing so hinges on the assumption that the true model is in fact sparse, and this is called the *sparsity assumption*, which can be motivated by the *bet-on-sparsity principle*: Assume that the underlying model is sparse and use a sparse method to model it. If the assumption is

[1] A feature is the values of a variable or transformations of a variable for all the observations in the data. It is synonymous with *predictor* and *regressor*.

correct, then our method has a chance of doing well. But if the assumption is incorrect, then our method will not work—but no other method would.

The success of neural networks and other methods that model high-dimensional data well but do not enforce sparsity might seem to defy the validity of this principle. But this is not the case: if the true model truly is sparse, then we will always do better with a sparse method.

Sparse regression methods have a long-standing history in statistics, starting in the 60's with the advent of stepwise regression methods, which are defined by their use of iterative procedures in which features are added or removed from a candidate model in steps. Typically, these algorithms use hypothesis tests to determine whether to add or remove a feature. The first popular stepwise method was invented by Efroymson [Efr60], which is a type of forward-stepwise algorithm. Efroymson's algorithm starts with an intercept-only model and then adds or removes features of the design matrix incrementally into the model. Later methods include backward-stepwise regression [RS68] and best-subset selection [BKM67], the latter considering all possible subsets of the features.

Stepwise methods are now considered problematic for several reasons, such as yielding R^2 values that are too high, p-values that are too small, and arbitrary variable selection when the variables are collinear[2].

Another important reason for why they are not used as much today is that they are computationally expensive since they need to consider all possible subsets of the features[3] This reason was part of the motivation for the development of regularized methods, which instead of solving standard regression problems on subsets of the features penalize the objective with a penalty that induces sparsity, and which can be fit using much more efficient methods. In this short report we will discuss these methods, the optimization algorithms underlying them, and their history. We will do so by focusing on two methods: the lasso and sorted ℓ_1 penalized estimation (SLOPE).

2 Summary of the Papers

2.1 Paper 1

In this paper, we address the challenge of extracting relevant features from data sets where the number of observations, n, is significantly smaller than the number of predictors, p. We focus on the Sorted L-One Penalized Estimation (SLOPE)—a generalization

[2] See Harrell [Har15, Chapter 4.3] for a detailed discussion on this topic.

[3] Bertsimas, King, and Mazumder [BKM16] has recently shown that this might not be as much of a problem as previously thought. Apparently, the best-subset selection can be solved relatively efficiently using mixed-integer programming, which means that the problem is solvable in polynomial time.

of the lasso—as a promising method in this context. However, current numerical procedures for SLOPE lack the efficiency that lasso tools possess, especially when estimating a complete regularization path. A key component of lasso's efficiency is predictor screening rules, which allow predictors to be discarded before model estimation. This paper is the first to establish such a rule for SLOPE. We develop a SLOPE screening rule by examining its subdifferential and demonstrate that this rule is a generalization of the strong rule for the lasso. Although our rule is heuristic and may occasionally discard predictors erroneously, we show that such instances are rare and can be easily safeguarded against by a simple check of the optimality conditions. Our numerical experiments reveal that the rule performs well in practice, leading to significant improvements for data in the $p \gg n$ domain, and incurs no additional computational overhead when $n > p$. This paper, therefore, presents a significant advancement in the efficiency of SLOPE, particularly in high-dimensional settings.

2.2 Paper II

In this paper, we focus on the lasso, a widely used method for inducing shrinkage and sparsity in the solution vector of regression problems, especially when the number of predictors outweighs the number of observations. Solving the lasso in such high-dimensional settings can be computationally challenging. However, this challenge can be mitigated through the use of screening rules that discard predictors before fitting the model, resulting in a reduced problem. We introduce a new screening strategy, termed look-ahead screening. This method employs safe screening rules to identify a range of penalty values for which a specific predictor cannot enter the model, thereby screening predictors along the remaining path. Our experiments demonstrate that these look-ahead screening rules outperform the active warm-start version of the Gap Safe rules, marking a significant advancement in the efficiency of solving high-dimensional lasso problems.

2.3 Paper III

In this paper, we address the challenge of predictor screening rules in l1-regularized regression problems, such as the lasso. These rules, which eliminate predictors from the design matrix before fitting a model, have significantly improved the speed of solving such problems. However, current state-of-the-art screening rules struggle with highly-correlated predictors, often becoming overly conservative. To tackle this issue, we introduce a new screening rule: the Hessian Screening Rule. This rule leverages second-order information from the model to provide more accurate screening and higher-quality warm starts. Our proposed rule outperforms all other alternatives we studied on datasets with high correlation for both l1-regularized least-squares (the lasso)

and logistic regression. It also delivers the best performance overall on the real datasets we examined. This paper, therefore, presents a significant advancement in dealing with highly-correlated predictors in l1-regularized regression problems.

2.4 Paper iv

In this paper, we tackle the challenges posed by the rapid development of machine learning research, particularly in the area of numerical validation. Researchers often face a multitude of methods to compare, lack of transparency and consensus on best practices, and the tedious task of re-implementing work. This often results in partial validation, which can lead to incorrect conclusions and hinder research progress. To address these issues, we introduce Benchopt, a collaborative framework designed to automate, reproduce, and publish optimization benchmarks in machine learning across different programming languages and hardware architectures. Benchopt simplifies the benchmarking process by providing a ready-to-use tool for running, sharing, and extending experiments. We demonstrate its wide applicability through benchmarks on three standard learning tasks: ℓ_2-regularized logistic regression, Lasso, and ResNet18 training for image classification. These benchmarks reveal key practical findings that provide a more nuanced view of the state-of-the-art for these problems, emphasizing that the details matter in practical evaluation. We believe that Benchopt will encourage collaborative work in the community and improve the reproducibility of research findings.

2.5 Paper v

In this paper we delve into the Sorted L-One Penalized Estimation (SLOPE), an extension of the renowned lasso regression method. Despite the promising statistical properties of SLOPE, its adoption has been limited due to the inefficiency of existing algorithms in high-dimensional contexts. To overcome this challenge, we introduce a novel, faster algorithm that solves the SLOPE optimization problem.

Our algorithm merges the techniques of proximal gradient descent and proximal coordinate descent, significantly enhancing the efficiency of the SLOPE method. We also shed new light on the directional derivative of the SLOPE penalty and its associated SLOPE thresholding operator, and provide assurances of convergence for our proposed solver. Through comprehensive benchmarks on both simulated and real data, we demonstrate that our method outperforms a host of competing algorithms. This paper is a significant contribution as it broadens the applicability of the SLOPE method in high-dimensional settings, potentially paving the way for its wider use in the field.

2.6 Paper VI

In this paper, we explore the sensitivity of regularized methods, such as the lasso and ridge regression, to the scales of the features in the data. It's standard practice to normalize features to ensure they share the same scale. While standardization is common for continuous data, binary data, particularly when high-dimensional and sparse, is often not scaled at all. We demonstrate that this choice can significantly impact the estimated model when the binary features are imbalanced, and that these effects also depend on the type of regularization used. Specifically, we show that the size of a feature's corresponding coefficient in the lasso is directly related to its class imbalance, and this effect depends on the normalization used. We propose potential solutions to this issue and discuss the case when data is mixed, containing both continuous and binary features. This paper, therefore, provides valuable insights into the impact of feature scaling on regularized methods and offers practical solutions for handling mixed data.

Bibliography

[BKM16] Dimitris Bertsimas, Angela King, and Rahul Mazumder. "Best Subset Selection via a Modern Optimization Lens". In: *The Annals of Statistics* 44.2 (Apr. 1, 2016), pp. 813–852. ISSN: 0090-5364. DOI: 10.1214/15-AOS1388.

[BKM67] E. M. L. Beale, M. G. Kendall, and D. W. Mann. "The Discarding of Variables in Multivariate Analysis". In: *Biometrika* 54.3-4 (Dec. 1, 1967), pp. 357–366. ISSN: 0006-3444. DOI: 10.1093/biomet/54.3-4.357.

[Efr60] Michael Alin Efroymson. "Multiple Regression Analysis". In: *Mathematical Methods for Digital Computers*. Ed. by Anthony Ralston and Herbert S. Wilf. 1st ed. Vol. 1. New York, USA: John Wiley and Sons, Dec. 1, 1960, pp. 191–203. ISBN: 978-0-471-70686-1.

[Har15] Frank E. Harrell Jr. *Regression Modeling Strategies*. 2nd ed. Springer Series in Statistics. Cham, Switzerland: Springer, Aug. 26, 2015. 582 pp. ISBN: 978-3-319-19425-7. DOI: 10.1007/978-3-319-19425-7.

[Lar+23] Johan Larsson, Quentin Klopfenstein, Mathurin Massias, and Jonas Wallin. "Coordinate Descent for SLOPE". In: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. AISTATS 2023. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. Valencia, Spain: PMLR, Apr. 25–27, 2023, pp. 4802–4821.

[Lar21] Johan Larsson. "Look-Ahead Screening Rules for the Lasso". In: *22nd European Young Statisticians Meeting - Proceedings*. 22nd European Young Statisticians Meeting. Ed. by Andreas Makridis, Fotios S. Milienos, Panagiotis Papastamoulis, Christina Parpoula, and Athanasios Rakitzis. Athens, Greece: Panteion University of Social and Political Sciences, Sept. 6, 2021, pp. 61–65. ISBN: 978-960-7943-23-1.

[LBW20] Johan Larsson, Małgorzata Bogdan, and Jonas Wallin. "The Strong Screening Rule for SLOPE". In: *Advances in Neural Information Processing Systems 33*. 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. Vol. 33. Virtual: Curran Associates, Inc., Dec. 6, 2020–12, pp. 14592–14603. ISBN: 978-1-71382-954-6.

[LW22] Johan Larsson and Jonas Wallin. "The Hessian Screening Rule". In: *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. New Orleans, USA: Curran Associates, Inc., Nov. 28–Dec. 9, 2022, pp. 15823–15835. ISBN: 978-1-71387-108-8.

[Mor+22] Thomas Moreau, Mathurin Massias, Alexandre Gramfort, Pierre Ablin, Pierre-Antoine Bannier, Benjamin Charlier, Mathieu Dagréou, Tom Dupré la Tour, Ghislain Durif, Cassio F. Dantas, Quentin Klopfenstein, Johan Larsson, En Lai, Tanguy Lefort, Benoit Malézieux, Badr Moufad, Binh T. Nguyen, Alain Rakotomamonjy, Zaccharie Ramzi, Joseph Salmon, and Samuel Vaiter. "Benchopt: Reproducible, Efficient and Collaborative Optimization Benchmarks". In: *Advances in Neural Information Processing Systems 35*. 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. New Orleans, USA: Curran Associates, Inc., Nov. 28–Dec. 9, 2022, pp. 25404–25421. ISBN: 978-1-71387-108-8.

[RS68] Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. 1st ed. Boston, USA: Harvard University, 1968. 356 pp. ISBN: 978-0-87584-017-8.

Papers

I

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

papers/paper1.pdf

II

III

IV

V

VI