Penalized likelihoods: Fast algorithms and risk bounds
Luo, Xi
ProQuest Dissertations and Theses; 2009; Dissertations & Theses @ Yale University

Abstract

Penalized Likelihoods: Fast Algorithms and Risk Bounds

Xi Luo

2009

We present a general iterative algorithm for seeking the optimizer $\hat{f}$ of the penalized likelihood criterion $\log 1/\mathit{likelihood}(f) + \mathit{pen}(f)$. Attention is given to optimization of a linear combination of terms selected from a library consisting of a possibly very large number of candidate variables or functions. We determine the computational accuracy of the optimizer depending on the number of iterations taken. This analysis also provides statistical risk bounds, measured in squared Hellinger distance and Kullback-Leibler divergence, by showing a variable-complexity covering property. We specialize these results to show fast algorithms and risk bounds with a penalty that is the ℓ_1 norm of the coefficients times a suitable multiplier λ. Examples we study here include ℓ_1 penalized least squares and Gaussian covariance matrix estimation. We show adaptive risk bounds for least squares with proper choices of penalizing parameters as well as a modified criterion to adapt to unknown error variance. Numerical merits and modified variants of our algorithms are also provided. The computation and estimation performance are illustrated using numerical examples, and they compare favorably with other existing methods.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Penalized Likelihoods:

Fast Algorithms and Risk Bounds

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Xi Luo

Dissertation Director: Andrew R. Barron

May, 2009

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

UMI Number: 3361525

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 3361525
Copyright 2009 by ProQuest LLC
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Copyright © 2009 by Xi Luo

All rights reserved.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Contents

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Overview

Contemporary statistical modeling, as arises in many scientific and technological investigations, is strongly impacted by modern data gathering and computational capabilities. One encounters not only a possibly large number of observed variables but also the computational flexibility to consider a vast set of candidate models, so as to seek out one which is parsimonious and accurate, not only empirically on the given data but also theoretically in its statistical risk. Penalized likelihood methods, especially with the penalty favoring simple models, have been widely studied both in theory and computation, and they also have broad applications in modeling large and complex data in biology, social sciences and many other disciplines.

The criterion we consider in this dissertation is to seek the estimator $\hat{f}$ to minimize the penalized likelihood $\log 1/likelihood(f) + pen(f)$. Special attention is given to estimators that use linear combinations of terms from a possibly large class of candidate functions; special cases include but are not restricted to polynomials, trigonometric terms, sigmoids, splines, and wavelets, enlarging the flexibility to model the underlying f from various function classes.

Though we allow various forms of the penalty $pen(f)$, we are particularly attracted to the popular ℓ_1 penalty. There have been considerable research efforts in both theory and computation, especially in the ℓ_1 penalized least squares case. This ℓ_1 penalty has shown promising results for modeling high-dimensional data

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

when the number of variables is far bigger than the sample size. This scenario is now very common in many areas of application including, for example, microarray, fMRI, genome-wide association in biological research and many applications in astrophysics and politics. Important questions concerning the penalized likelihood procedures arise in applications as well as in mathematical statistics, and we hope to attach some of them here.

This dissertation starts by considering a generic density model where the computational accuracy and statistical risk bounds are derived. Interesting cases include least squares regression, Gaussian graphical models, and logistic regression. Corresponding computation and statistical risk results for the first two special models are reviewed, while the logistic model is omitted due to its close connection to the generic one. The proposed framework for obtaining these results is also fairly general, and the potential to extend to other density models arising in applications and theory is very attractive.

Research efforts in theory and practice have diverged. Theoretical results have shed light on the generalization of such estimators to future data, usually under conditions or assumptions that provide little guidance for use in practical applications. On the other hand, statistical practitioners choose to stick to methods preferred from their ample experience working with real data. This dissertation provides explicit procedures for practice as well as corresponding statistical risk bounds with minimal conditions.

Statistical methodologies face a computational challenge in practice when dealing with extremely large data sets, possibly in the hundreds of thousands of variables. Even for the case of least squares with the ℓ_1 penalty, which is a special case of our likelihood criteria, there has been more than a decade of efforts to improve the computational efficiency. Working in Parallel to other theoretical efforts on this prob-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

lem, we provide algorithmic and numerical strategies with favorable computational performance, compared to other existing strategies.

Modern data may come with a structure different than traditional response and explanatory variables, and the generic setting we consider accommodates a vast spectrum of models, beyond the simple regression model. Some structure specialties are studied in this dissertation and there is a continuing interest in the application to models for many other complex data sets.

In this dissertation, we will address these interesting topics in separate chapters. Each chapter will serve as a short version of a future publication, with introduction, results and summary.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 1

Log-density Estimation

1.1 Introduction

Likelihood estimation has been a major component in statistical modeling, and the variant using the ℓ_1 penalty is currently a popular approach. Examples include applications in gene microarray, fMRI, climate studies and many other scenarios where simple and sparse models are favored for scientific investigation purposes, particularly when the sample size is far smaller than the number of variables. Computation of these estimators is challenging when applying this likelihood approach. The maximum likelihood principle requires the computed estimator to be at the optimum, or at least within a close neighborhood. Moreover, there have been recent efforts in studying statistical risk properties of ℓ_1 penalized estimators, mostly in the regression setting. It is also important to understand analogous risk properties in penalized likelihood settings. These motivate us to study the problem of computation and risk bounds of ℓ_1 penalized likelihood.

Concerning computation, algorithms have been proposed for solving the ℓ_1 penalized likelihood criteria in specific settings. For a convex criterion, the existing

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

interior point methods (see [15] for example) may be applied, but with a time complexity of the order $O(p^3)$ where p is the number of parameters of the search, usually related to the number of candidate variables. This kind of complexity is not realistic for p in the tens of thousands. Other alternatives have been proposed to overcome the computational difficulty. For example in a recent popular topic on sparse Gaussian covariance matrix estimation, [24, 64] adapt the interior point method to solve the exact ℓ_1 penalized likelihood, and [1, 35] apply coordinate descent type methods to optimize the dual problem of the same criterion. The latter is efficient for those p in the tens of thousands, but runs into difficulty for ever larger problems. The algorithmic convergence rate of such can only be established locally using the general results by [51, 58]. Though these existing algorithms may tend to work well on specific numerical examples, it is still an open question how close the computed objective is to the global optimum in moderate time. We present in this paper a greedy algorithm to solve the ℓ_1 penalized likelihood principle for a general class of densities, and we also establish results concerning the closeness of our estimator to the global optimum.

Relaxed greedy algorithms have been mostly studied in the case of least squares (see the work [45, 3, 48, 6] and references within). [42] recently introduced a relaxed greedy algorithm (called ℓ_1 penalized greedy pursuit) to solve the ℓ_1 penalized least squares problem. In chapter 3, I show that their algorithm and the modification I introduce is competitive with other methods of ℓ_1 penalized least squares, also known as LASSO [57] or Basis Pursuit [22]. In the present chapter, I generalize the greedy principles shared by those regression counterparts to the ℓ_1 penalized likelihood, and derive corresponding algorithmic bounds. Another related use of greedy algorithms for density estimation is in [66], but that work requires densities from convex smooth families. Our results do not place restriction on smoothness. The log density is

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

modeled as being a superposition of functions in a library of candidate functions, which we argue is quite common in statistic modeling. Though we will mainly focus on the ℓ_1 penalty in this paper, other forms of the penalty could also be considered following the same relaxed greedy selection principle, and similar algorithmic results would apply.

Density estimation is closely related to data compression, following the pioneering work by Shannon [56]. Indeed $\log 1/p(data)$ corresponds to the length of a code for data if p is given. More generally, if there is family of distributions for data, the Minimal Description Length (MDL) principle (e.g. [11] for a review) seeks the choice for which a total codelength is minimal. In particular, this may entail minimization of $\log 1/p(data) + pen(f)$ where the penalty $pen(f)$ is an information theoretic codelength needed to describe f in a family $\mathcal{F}$. An index of resolvability, the optimum sum of relative entropy approximation error and penalty, is used to upper bound the expected redundancy of data compression, as well as statistical risk in model estimation. Indeed, resolvability bounds have been developed in [2, 7, 50, 46, 47, 40] for f in a countable set that discretizes $\mathcal{F}$, and in [62, 5] for finite dimensional families. Recently, the ℓ_1 penalty has been shown to be an also information theoretically valid penalty in [9, 8, 10], where a covering idea is used to extend the results for a countable set. We will summarize these recent developments and present an alternative proof for demonstrating the existence of covering properties.

Here, we are mainly interested in the performance of recovering the population density p_{f^*} rather than good variable selection. This coincides with the MDL thinking of data compression and also is an indication of statistical estimation performance. Moreover, the goal of variable selection is not so transparent in over-complete flexible function fitting we are inclined to think here, whereas in parametric models (our result would also apply) the classification of zero and nonzero variables might

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

be important. Variable selection requires proper assumptions on the correlation structure of variables, which we don't impose in this dissertation.

This chapter is organized as follows. We will present our main greedy algorithm in section 1.2, and discuss the risk bounds in section 1.3. We will include proof analysis in section 1.4 and then summarize. Examples of special cases as well as their numerical performance will be reported in the following chapters.

1.2 Algorithm

We consider *i.i.d.* observations $X_1, \ldots, X_n$ from a density function of the form

$$p_f(x) = \frac{e^{f(x)} p_0(x)}{C_f}$$

where $C_f = \int e^{f(x)} p_0(x) dx$ is the normalizing constant integrating over a known reference density $p_0(x)$. We are particularly interested in the estimation of the log density function $f(x)$ by a linear combination of element functions $h(x)$ from a library of functions, which we call a dictionary $\mathcal{H}$. Formally, we consider the estimator of the form $f_\theta(x) = \sum_h \theta_h h(x)$ summing over all $h \in \mathcal{H}$. We impose a ℓ_1 penalty $\|\theta\|_1 = \sum_h a_h |\theta_h|$, where a_h is a customarily chosen scale of h to be specified later. It is then natural to consider minimizing the sum of minus per sample log likelihood and a ℓ_1 penalty term, that is

$$L(X, f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{p_\theta(X_i)} + \lambda_n \|\theta\|_1 \tag{1.2.1}$$

where $\lambda_n \geq 0$ is an appropriate penalizing parameter, and we write p_θ for p_f when $f = f_\theta$. We use $\underline{X} = (X_1, \ldots, X_n)^T$ to refer to all the data.

In a special case when $p_\theta(y|x)$ is proportional to $\exp\left\{-(y - x \cdot \theta)^2 / (2\sigma^2)\right\}$ with

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

a constant σ, the likelihood criterion $L(X, f_\theta)$ is equivalent to the LASSO objective in regression with an appropriately rescaled $\lambda \geq 0$:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i \cdot \theta)^2 + \lambda \|\theta\|_1 .$$

In the regression setting, relaxed greedy algorithms and ℓ_1 penalized greedy pursuit would build up the estimator in a iterative scheme, by adding an optimal weighted covariate while down-weighting the coefficients from the previous iterations. There is interest in generalizing this strategy for penalized likelihood and demonstrating its computational properties.

The greedy algorithm we propose is also an iterative strategy to reduce the objective $L(x, f_\theta)$ by carefully introducing a new term, and we call it greedy likelihood pursuit (GLP). One initializes with an estimator $\hat{f}_0$, usually with $\hat{f}_0 = 0$ for simplicity. Given the previous estimator $\hat{f}_{k-1}$, an enhanced estimator $\hat{f}_k = (1-\alpha_k)\hat{f}_{k-1} + \beta_k h_k$ is constructed with (α_k, β_k, h_k) chosen to minimize

$$L\left(x, (1-\alpha)\hat{f}_{k-1} + \beta h\right)$$

over $\alpha \in [0,1]$, $\beta \in \mathbb{R}$, and $h \in \mathcal{H}$. Repeat such iterations until the desirable computational accuracy is achieved. At each iteration, the minimization is in a lower-dimensional space than the original problem. Thus, there is a reduction of the computation complexity. For a finite dictionary $\mathcal{H}$, the minimization over h is implemented easily by comparing objective values for each choice $h \in \mathcal{H}$ with optimal weights α, β correspondingly.

The accuracy of $\hat{f}_k$ to the global optimal $\hat{f}_\theta$ is measured in the objective value sense. We will establish this computational accuracy bound by an induction argument. First it is obvious to see that the objective value $L(x, \hat{f}_k)$ is not more than

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$L(x, \hat{f}_{k-1})$. By considering the objective difference of our estimator to an arbitrary one (favorably a minimizer), significant improvement in the difference can be shown after each iteration. We interpret the difference in normalizing constants as a moment generating function and the rest of the terms is taken care of easily. The generating function term we identify is bounded by a Hoeffding type inequality when a_h is chosen to be not less than $\|h\|_\infty$. We have the following theorem describing the computational accuracy for k iterations.

Theorem 1.2.1. *For any given data set X and for all $k \geq 1$, the following computational accuracy bound holds for GLP*

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{p_{\hat{f}_k}(x_i)} + \lambda_n v_k \leq \inf_{\theta} \left\{ \frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{p_\theta(x_i)} + \lambda_n \|\theta\|_1 + \frac{2\|\theta\|_1^2}{k+1} \right\}$$

where $v_k = \sum_h a_h \left|\hat{\theta}_{h,k}\right|$ for $\hat{f}_k = \sum \hat{\theta}_{h,k} h$, and $a_h \geq \|h\|_\infty$.

This bound reveals the optimal sum of the objective value and the computational accuracy term for appropriate θ. This is, in particular, not bigger than the sum of the objective at optimal θ^* and the corresponding computational accuracy term $2\|\theta^*\|_1^2/(k+1)$, as one may expect to see.

Clearly, our algorithm will produce an estimator within an order $1/k$ bound above the global minimal. Unlike the convergence rate analysis of other algorithms that make assumptions in a close neighborhood around $L(x, \hat{f})$ or assume the number of iterations k large, our proof does not impose such assumptions and the bound holds for all $k \geq 1$.

This result is particularly useful when considering statistical risk with an approximate minimizer where we can balance the computation accuracy and the risk term. Indeed, the overall risk bound corresponds to a sum of statistical risk for an exact

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

minimizer and a computation accuracy bound. We could seek k carefully so that the computational accuracy is of small order compared to the statistical risk. As we will show in section 1.3, this k for the overall risk consideration can be of significantly smaller order than n and p.

Here we require bounded $\|h\|_\infty$ in order to use the Hoeffding type bound on moment generating functions. Examples include bounded variables, as in categorical variable settings; or bounded function classes, as in neural net fitting. In addition, there is an interest in considering unbounded variables arising in Gaussian inverse covariance matrix estimation, where moment generating function control could be achieved via Berstein type moment conditions. The extended results on this topic will be reported in chapter 4.

1.3 Risk Analysis

The risk analysis of the ℓ_1 penalized estimator minimizing (1.2.1) is built on the early work of [50, 40] where the penalty term arises from a countable class of estimators. For the class of estimators with continuous ℓ_1 norm of coefficients which is uncountable, a simple condition for valid penalty to satisfy a certain covering property is reviewed here, as detailed in [9, 8, 10].

The density distance measure we consider here is the Kullback divergence and a Bhattacharyya, Rényi, Hellinger divergence, which are used in examining the quality of statistical estimates and data compression. The Kullback divergence $D(P_{\underline{X}}\|Q_{\underline{X}}) = E\log p(\underline{X})/q(\underline{X})$ is the total expected redundancy for data $\underline{X}$ described using joint density function $q(\underline{x})$ but governed by a density $p(\underline{x})$. Likewise the Bhattacharyya, Hellinger, Rényi divergence [12, 23, 55] is given by $d(P_{\underline{X}}, Q_{\underline{X}}) = 2\log 1/\int (p(\underline{x})q(\underline{x}))^{1/2}$. We use $D_n(f^*, f)$ and $d_n(f^*, f)$ to denote the divergences

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

between the joint distributions $p_{f^*}(\underline{x})$ and $p_f(\underline{x})$. In the *i.i.d.* modeling case these take the form $D_n(f^*, f) = nD(f^*, f)$ and $d_n(f^*, f) = nd(f^*, f)$, respectively, where $D(f^*, f)$ and $d(f^*, f)$ are the divergences between the single observation distributions $p_{f^*}(x)$ and $p_f(x)$. The divergences measure how well f approximates f^*.

Writing $D(P||Q) = -2E\log(q(\underline{X})/p(\underline{X}))^{1/2}$ and employing Jensen's inequality shows that $D(P\|Q) \geq d(P,Q)$. The relationship to the squared Hellinger distance $H^2(P,Q) = \int(p(\underline{x})^{1/2} - q(\underline{x})^{1/2})^2$ is $d(P,Q) = -2\log(1-\frac{1}{2}H^2)$, which is not less than $H^2(P,Q)$. These divergences upper bound the square of the L_1 distance. Moreover, $d(P,Q)$ is locally equivalent to the Kullback-Leibler divergence when $\log p(\underline{x})/q(\underline{x})$ is upper-bounded by a constant. Moreover, it evaluates to familiar quantities in special cases, e.g., for two normals of mean μ and $\tilde{\mu}$ and variance σ^2, it is $\frac{1}{4}(\mu - \tilde{\mu})^2/\sigma^2$. The most important reason for our use of the Bhattacharyya, Rényi, Hellinger loss function is that it allows clean examination of the risk, without putting any conditions on the density functions $p_f(\underline{x})$.

Statistical risk analysis of minimal complexity estimators like (1.2.1) traditionally deals with penalty $pen(\tilde{f})$ for $\tilde{f}$ in a countable class $\tilde{\mathcal{F}}$, where $\tilde{\mathcal{F}}$ could be discretization of a uncountable class $\mathcal{F}$. The resolvability bound on statistical risk in [50, 40] shows that for $pen(\tilde{f}) \geq 2L_n(\tilde{f})$ where $L_n(\tilde{f})$ satisfies $\sum_{\tilde{f}\in\tilde{\mathcal{F}}} e^{-L_n(\tilde{f})} \leq 1$, the minimal complexity estimator $\hat{f}$ achieving

$$\min_{\tilde{f}\in\tilde{\mathcal{F}}} \left\{ \log \frac{1}{p_{\tilde{f}}(\underline{X}_n)} + pen(\tilde{f}) \right\}$$

has the expected divergence of $p_{\hat{f}}$ and p^* bounded by the index of resolvability, that is

$$\mathbb{E}d_n(f^*, \hat{f}) \leq \inf_{\tilde{f}\in\tilde{\mathcal{F}}} \left\{ D_n(f^*, \tilde{f}) + pen(\tilde{f}) \right\}.$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In particular with *i.i.d.* modeling, the risk satisfies

$$\mathbb{E}d(f^*, \hat{f}) \leq \inf_{\tilde{f} \in \tilde{\mathcal{F}}} \left\{ D(f^*, \tilde{f}) + \frac{pen(\tilde{f})}{n} \right\}.$$

Corresponding results for uncountable $\mathcal{F}$ were developed in [9, 8, 10], allowing application to optimization over real-valued parameters in standard statistical models. In that analysis an important role is played by a measure of the discrepancy between empirical and population values of the log-likelihood ratio at a candidate f. As explained there it is given by

$$dis(f) = \log \frac{p_{f^*}(\underline{X})}{p_f(\underline{X})} - 2 \log \frac{1}{E(p_f(\underline{X})/p_{f^*}(\underline{X}))^{1/2}}.$$

In the proof of statistical risk bounds for the countable case, if an information-theoretically valid penalty $pen(f)$ is added to the discrepancy, then uniformly in f (i.e., even with a data-based $\hat{f}$ in place of a fixed f) the expectation of the penalized discrepancy is positive.

This leads to consideration, in the uncountable case, of penalties which exhibit a similar discrepancy control. We say that a uncountable collection $\mathcal{F}$ with a penalty $pen(f)$ for $f \in \mathcal{F}$ has a *variable–complexity variable–discrepancy* cover suitable for p_{f^*} if there exists a countable $\tilde{\mathcal{F}}$ and $\mathcal{L}(\tilde{f}) = 2L(\tilde{f})$ satisfying $\sum_{\tilde{f}} e^{-L(\tilde{f})} \leq 1$, such that the following condition $(*)$ holds for all $\underline{X}$:

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}} \left\{ dis(\tilde{f}) + \mathcal{L}(\tilde{f}) \right\} \leq \inf_{f \in \mathcal{F}} \left\{ dis(f) + pen(f) \right\}. \qquad (*)$$

This condition satisfies the aim that the penalty in the uncountable case mirrors an information-theoretically valid penalty in the countable case. The condition, above, yields the desired positivity of the expected penalized discrepancy. Indeed, because

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the minimum over the countable $\tilde{f}$ is shown to have non-negative expectation, the minimum over all f in $\mathcal{F}$ will also when $(*)$ holds.

Equivalent to condition $(*)$, the following characterization $(**)$ is convenient. For each f in $\mathcal{F}$ there is an associated representer $\tilde{f}$ in $\tilde{\mathcal{F}}$ for which

$$pen(f) \geq \log \frac{p_f(\underline{X})}{p_{\tilde{f}}(\underline{X})} - 2 \log \frac{E(p_f(\underline{X})/p_{f^*}(\underline{X}))^{1/2}}{E(p_{\tilde{f}}(\underline{X})/p_{f^*}(\underline{X}))^{1/2}} + 2L(\tilde{f}) \,. \qquad (**)$$

The idea is that if $\tilde{f}$ is close to f then the discrepancy difference is small. Then the complexity of such $\tilde{f}$ along with the discrepancy difference assesses whether a penalty $pen(f)$ is suitable. Nevertheless, the minimizer in $\tilde{\mathcal{F}}$ depends on the data and accordingly we allow the representer $\tilde{f}$ of f to also have such dependence. With this freedom, in cases of interest, the variable complexity cover condition indeed holds for all $\underline{X}$, though it would suffice for our purposes that $(*)$ hold in expectation.

Condition $(**)$ specifies that there be a cover with variable distortion plus complexity rather than a fixed distance and fixed cardinality. This is analogous to the distortion plus rate trade off in Shannon's rate-distortion theory. In our treatment, the distortion is the discrepancy difference (which does not need to be a metric), the codebook is the cover $\tilde{\mathcal{F}}$, the codelengths are the complexities $L(\tilde{f})$. Valid penalties $pen(f)$ exceed the minimal sum of distortion plus complexity.

In the case that $pen(f) = pen(f_\theta) = \lambda \|\theta\|_1$, we focus on in this paper, condition $(**)$ requires that a suitable λ multiplying ℓ_1 norm of continuously parametrized θ would indeed exceed the complexity of $\tilde{f}$ associated with some discretized θ plus the distortion. One may demonstrate the existence of such representor $\tilde{f}$ in a countable cover by showing a representor with random samples of functions would be small enough for the penalty to satisfy the requirement $(**)$. An alternative is to use a greedy algorithm to select those coefficients so as to construct the representor $\tilde{f}$ di-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

rectly, and indeed the computational accuracy argument characterizing the distortion would also rely on similar sampling ideas.

As demonstrated in section 1.2, greedy algorithms on the likelihood would select an estimator in k iterations such that the distance to the target objective is of order $1/k$. The proof of such algorithms demonstrates that a sequence of estimators produced by a deterministic schedule of α_k and β_k for $k \geq 1$ is good enough to achieve the improvement at each iteration, and then the overall computational accuracy bound after any k iterations. Using such a fixed schedule greedy algorithm, the coefficients of the estimator at any k is uniquely determined by the order of functions to be included, which is a product of α_k and β_k for various k. The complexity or codelength for such a estimator is proportional to the number of iterations.

Examining the condition $(**)$, the discrepancy difference is to compare the log-likelihood ratio of f and $\tilde{f}$ to its population counterpart. This difference is shown to be not bigger than a quantity over the number of iterations if $\tilde{f}$ is indeed constructed by a greedy algorithm. Since the complexity is also determined by the number of iterations, the overall bound on the distortion plus complexity is achieved by balancing the two terms, and our penalty in the condition is at least such for an appropriate λ large enough.

Recall that the *i.i.d.* data each follows the density function

$$p_f(x) = \frac{e^{f(x)} p_0(x)}{C_f}.$$

Examining the difference in discrepancies at any estimator $f = f_\theta$ and a representing $\tilde{f}$ we see that both $p_0(x)$ and c_f cancel out. What remains in the discrepancy

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

difference is

$$dis(f, \tilde{f}) = dis(\tilde{f}) - dis(f)$$
$$= \sum_{i=1}^{n} (f(X_i) - \tilde{f}(X_i)) + 2n \log E \exp\{\tfrac{1}{2}(\tilde{f}(X) - f(X))\}$$

where the expectation is with respect to a distribution for X constructed to have density which is the normalized pointwise affinity $p_a(x) = [p_{f^*}(x) p_f(x)]^{1/2} / A(f^*, f)$.

The representor $\tilde{f}$ close to f is also sought similarly by a greedy algorithm as in the likelihood case. We also bound the discrepancy difference by a fixed schedule greedy algorithm where we allow optimizing over h only, with α_k and β_k to be fixed functions of k. For a clean result with small constants, we customarily choose the schedule $\alpha_k = 2/(k+1)$ and $\beta_k = \alpha_k \|\theta\|_1 / a_h$ where $f = f_\theta$ in the discrepancy difference. Similar to the computational accuracy result, the estimator $\tilde{f}_k$ produced by a fixed schedule greedy algorithm for discrepancy will have the discrepancy difference bounded by a term over k.

Theorem 1.3.1. *For any given data set X and for all $k \geq 1$, the greedy algorithm for discrepancy has the following bound for the difference*

$$dis(f_\theta, \tilde{f}_k) \leq \frac{n \|\theta\|_1^2}{k+1}.$$

The codelength of $\tilde{f}_k$ could be crudely coded by $k \log(2M)$ for k terms where $M = Card(\mathcal{H})$. A valid ℓ_1 penalty should exceed the optimal tradeoff of the discrepancy difference and the complexity, where the minimizing k occurs roughly at $\|\theta\|_1 \sqrt{n/(2 \log 2M)}$. The minimal value will follow using this minimizing k. With the ℓ_1 penalty satisfying the condition $(**)$ by the above construction, we achieve a similar bound on the statistical risk as the countable case.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Theorem 1.3.2. *The ℓ_1 penalized likelihood estimator $\hat{f} = f_{\hat{\theta}}$ achieving*

$$\min_{\theta} \left\{ \log \frac{1}{p_{f_\theta}(X_n)} + \lambda_n ||\theta||_1 \right\}$$

has the risk bound

$$\mathbb{E}d(f^*, \hat{f}) \leq \inf_{f_\theta} \left\{ D(f^*, f_\theta) + \frac{\lambda_n ||\theta||_1}{n} \right\}$$

for every sample size, provided $\frac{\lambda_n}{n} \geq 2 \left[\frac{2\log(2M)}{n} \right]^{1/2}$.

The bound shown here uses the minimal sum of the approximation error and the ℓ_1 penalty decaying with the sample size. $D(f^*, f_\theta)$ quantifies he error of approximation and $(\lambda_n/n)\|\theta\|_1$ quantifies additional error from finite sample estimation.

If $f^* = f_{\theta^*}$ is in the linear span of the dictionary with a small or moderate $\|\theta^*\|_1$, then the risk is of order $\sqrt{\log(M)/n}$ which is small even with variable size M large as long as the logarithm of M is small. This allows M to be nearly exponentially large in n while still having small risks. Even if f^* does not have a moderate size ℓ_1 norm representation by linear combination in the dictionary, the risk bound expresses the best tradeoff in approximation error and ℓ_1 norm of the approximation suitable for the given sample size.

A similar bound is proven by constructing $\tilde{f}$ as random samples of functions in $\mathcal{H}$ (see [9, 8, 10]). That result will have a saving of constant 2 in λ_n with the same rate, and plus asymptotically smaller order terms due to rounding. What is enlightening in the proof reported here is that the algorithmic results are adapted to prove risk bounds. The former one is considerably transparent in technical details. Nevertheless, by the proof strategy we use here we will handle more general settings in verifying condition (**), for example when considering unbounded functions h

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

arising in the Gaussian covariance matrix estimation.

The risk bound does not restrict the algorithm to compute the exact minimizer of the ℓ_1 penalized likelihood criterion. It is shown in this chapter that the GLP algorithm has the advantage of an explicit guarantee of the closeness to the minimum. The risk bound proof could also incorporate the computational accuracy term, we could then demonstrate the statistical risk of the sequence of approximate estimators produced by our GLP algorithm.

Theorem 1.3.3. *The k step* approximate *estimator $\hat{f}_k$ of the GLP algorithm for the ℓ_1 penalized likelihood objective*

$$\min_{\theta} \left\{ \log \frac{1}{p_{f_\theta}(\underline{X}_n)} + \lambda_n ||\theta||_1 \right\}$$

has the risk bound

$$\mathbb{E}d(f^*, \hat{f}_k) \leq \inf_{f_\theta} \left\{ D(f^*, f_\theta) + \frac{\lambda_n ||\theta||_1}{n} + \frac{2||\theta||_1^2}{k+1} \right\}$$

for every sample size, provided $\frac{\lambda_n}{n} \geq 2\left[\frac{2\log(2M)}{n}\right]^{1/2}$.

Ideally we would like to have k large enough so that the computation accuracy term vanishes in the risk bound, but the improvement from a super accurate estimator is not significant compared to other terms in the risk bound which are of order $[\frac{\log M}{n}]^{1/2}$. In the situation that exact minimization is not possible, we would like to choose $1/k$ to be of the same order to save the computation. This is quite advantageous in practice where n could be in thousands and M could be in hundreds of thousands, the k of our choice is just not more than hundreds.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1.4 Proofs

We now show the proof for the computation as well as the risk bound. The proof of computational accuracy bound is an extension of the proof for the ℓ_1 penalized least squares case [42], generalized to handle the log density estimation problem.

1.4.1 Proof of Theorem 1.2.1

It is equivalent to show that for each f in the linear span that

$$\frac{1}{n}\left[\log\frac{p_f(\underline{X}_n)}{p_{\hat{f}_k}(\underline{X}_n)} + \lambda(v_k - V_f)\right] \leq \frac{2V_f^2}{k+1}$$

where $V_f = \|\theta\|_1$ with $f = f_\theta$. The left side of this desired inequality which we shall call e_k is built from the difference in the criterion values at $\hat{f}_k$ and an arbitrary f. It can be expressed as

$$e_k = \frac{1}{n}\sum_{i=1}^{n}[f(X_i) - \hat{f}_k(X_i)] + \log\int p_f(x)e^{\hat{f}_k(x)-f(x)}$$

where the integral arising from the ratio of the normalizers for $p_{\hat{f}_k}$ and p_f. Without loss of generality, making $\mathcal{H}$ closed under sign change, we restrict our attention to positive β. This e_k is evaluated with $\hat{f}_k(x) = (1-\alpha)\hat{f}_{k-1}(x) + \beta h(x)$ and $v_k = (1-\alpha)v_{k-1} + \beta a_h$ at the optimized α, β and h, so it is as least as good as at an arbitrary h with $\beta = \alpha V_f/a_h$. Thus for any h we have that e_k is not more than

$$\frac{1}{n}\sum_{i=1}^{n}[f(X_i) - \bar{\alpha}\hat{f}_{k-1}(X_i) - \alpha V_f h(X_i)/a_h] + \log\int p_f(x)e^{[\bar{\alpha}\hat{f}_{k-1}(x)+\alpha V_f h(x)/a_h - f(x)]} + \bar{\alpha}\lambda[v_{k-1} - V_f],$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where $\bar{\alpha} = (1-\alpha)$. Now we reinterpret the integral using the expectation of $e^{\alpha[vh(x)/a_h - f(x)]}$ with respect to $p(x) = e^{\bar{\alpha}[f_{k-1}(x) - f(x)]} p_f(x)/c$, where c is its normalizing constant. Accordingly, we add and subtract $\log c = \log \int e^{\bar{\alpha}[f_{k-1}(x) - f(x)]} p_f(x)$ which, by Jensen's inequality using $\bar{\alpha} \leq 1$, is not more than $\bar{\alpha} \log \int e^{[f_{k-1}(x) - f(x)]} p_f(x)$. Recognizing that this last integral is what arises in e_{k-1} and distributing f between the terms with coefficients $\bar{\alpha}$ and α, we obtain that e_k is not more than

$$\bar{\alpha} e_k + \alpha \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - vh(X_i)/a_h] + \log \int e^{\alpha[vh(x)/a_h - f(x)]} p(x).$$

This inequality holds for all h so it holds in expectation with a random selection in which each h is drawn with probability $a_h|\theta_h|/v$ where the θ_h are the coefficients in the representation $f(x) = \sum_{h \in \mathcal{H}} \theta_h h(x)$ with $V_f = \sum_h |\theta_h| a_h$. We bring this expectation for random h inside the logarithm, and then inside the integral, obtaining an upper bound by Jensen's inequality. For each x and random h the quantities $[V_f h(x)/a_h - f(x)]$ have mean zero and have range of length not more than $2V_f$ since $a_h \geq \|h\|_\infty$. So by Hoeffding's moment generating function bound, the expectation for random h of $e^{\alpha[vh(x)/a_h - f(x)]}$ is not more than $e^{\alpha^2 v^2/2}$. Thus

$$e_k \leq (1-\alpha) e_{k-1} + \frac{1}{2} \alpha^2 V_f^2$$

for all $0 \leq \alpha \leq 1$, and so in particular with $\alpha = 2/(k+1)$. Also $e_0 \leq 2V_f^2$, so by induction

$$e_k \leq \frac{2V_f^2}{k+1},$$

which is the desired result.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1.4.2 Proof of Theorem 1.3.1

The proof is analogous to the computation analysis shown before but without the penalty term (or correspondingly set $\lambda = 0$). Using a fixed schedule of $\alpha = 2/(k+1)$ and $\beta = \alpha V_f / a_h$ at the k step, the discrepancy difference could be bounded similarly to e_k as

$$dis(f, \tilde{f}_k) \leq \bar{\alpha} dis(f, \tilde{f}_k) + 2n \log E exp\{\frac{1}{2}\alpha [V_f h / a_h - f]\}$$

where the expectation is taken over a density function proportional to $exp\{[\bar{\alpha}\tilde{f}_{k-1} + \alpha f + f^*]/2\}$. The Hoeffding bound is also used here to bound the moment generating function, and we obtain a similar iterative inequality as the one for e_k:

$$dis(f, \tilde{f}_k) \leq \bar{\alpha} dis(f, \tilde{f}_k) + \frac{1}{4} n \alpha^2 V_f^2.$$

Then by induction it follows that

$$dis(f, \tilde{f}_k) \leq \frac{n V_f^2}{k+1}.$$

1.4.3 Proof of Theorem 1.3.2

For completeness, we briefly sketch the proof for the countable cases as was first as shown in [50], as well as in [9, 8, 46, 40]. Adding and subtracting the pointwise redundancy to the Bhattacharyya, Rényi, Hellinger loss for the minimizer $\hat{f}$ in a

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

countable class of estimators, we reorganize terms to show

$$\begin{aligned} d(f^*,\hat{f}) &= 2\log\frac{\left(\frac{p_{\hat{f}}(\underline{X})}{p_{f^*}(\underline{X})}\right)^{1/2}e^{-L(\hat{f})}}{E\left(\frac{p_{\hat{f}}(\underline{X})}{p_{f^*}(\underline{X})}\right)^{1/2}}+\log\frac{p_{f^*}(\underline{X})}{p_{\hat{f}}(\underline{X})}+2L(\hat{f}) \\ &\leq 2\log\sum_f\frac{\left(\frac{p_f(\underline{X})}{p_{f^*}(\underline{X})}\right)^{1/2}e^{-L(f)}}{E\left(\frac{p_f(\underline{X})}{p_{f^*}(\underline{X})}\right)^{1/2}}+\log\frac{p_{f^*}(\underline{X})}{p_{\hat{f}}(\underline{X})}+2L(\hat{f}). \end{aligned}$$

The expectation used here is with respect to p_{f^*} and the last inequality is due to the positivity of summands. The rest of the proof takes expectations again with respect to p_{f^*} on both sides. By Jensen's inequality, the expectation is taken inside the log of the first term on the right side and the whole term then can be discarded, since the summation inside the log is not bigger than 1 when $L(f)$ is a valid codelength satisfying the Kraft inequality. The rest in the upper bound is the expected redundancy which is not more than the index of resolvability.

When condition $(**)$ is satisfied, for each f in uncountable $\mathcal{F}$, there is a $\tilde{f}$ in countable $\tilde{\mathcal{F}}$ such that

$$\log\frac{p_{f^*}(\underline{X})}{p_f(\underline{X})}+pen(f)-d(f^*,f)\geq\log\frac{p_{f^*}(\underline{X})}{p_{\tilde{f}}(\underline{X})}+2L(\tilde{f})-d(f^*,\tilde{f}).$$

Rearranging terms this shows that the Bhattacharyya, Rényi, Hellinger loss for the minimizer $\hat{f}$ in a uncountable $\mathcal{F}$ is bounded by

$$d(f^*,\hat{f})\leq d(f^*,\tilde{f})-\log\frac{p_{f^*}(\underline{X})}{p_{\tilde{f}}(\underline{X})}-2L(\tilde{f})+\log\frac{p_{f^*}(\underline{X})}{p_{\hat{f}}(\underline{X})}+pen(\hat{f}).$$

All terms involving $\tilde{f}$ can be thrown away, upon taking expectation with respect to p_{f^*} for all $\tilde{f}$ in a countable $\tilde{\mathcal{F}}$ using the same argument as above. Then an index of

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

resolvability bound on the rest terms reveals

$$\mathbb{E}d_n(f^*, \hat{f}) \leq \inf_f \{D_n(f^*, f) + pen(f)\}.$$

In the *i.i.d.* modeling case, it specializes to show that the risk satisfies

$$\mathbb{E}d(f^*, \hat{f}) \leq \inf_f \left\{ D(f^*, f) + \frac{pen(f)}{n} \right\}.$$

Now we only need to demonstrate the condition (**) for the ℓ_1 penalty and then the risk result will follow by the argument above. We consider finding $\tilde{f}$ by a greedy algorithm on discrepancy for k steps, then the discrepancy difference is shown by Theorem 1.3.1 bounded by

$$dis(f, \tilde{f}) \leq \frac{nV_f^2}{k+1}$$

and the complexity for such $\tilde{f}$ is $k \log(2M)$, where we code the integer crudely by $k \log 2$ and the function included at each step has a codelength of $\log M$. Altogether, the condition (**) requires that the penalty satisfies

$$pen(f) \geq \frac{nV_f^2}{k+1} + 2k \log(2M).$$

The minimal k is achieved at $k = V_f \sqrt{n/2 \log(2M)} - 1$, and we round up for an integer k for which it reveals a minimal condition

$$pen(f) \geq \lambda_n V_f$$

provided $\lambda_n \geq 2\sqrt{2n \log(2M)}$, which is the required minimal λ_n in the theorem.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1.4.4 Proof of Theorem 1.3.3

The proof for an approximate estimator $\hat{f}_k$ will replace the pointwise redundancy by the minimizer $\hat{f} = f_{\hat{\theta}}$ of the pointwise redundancy plus an extra computational accuracy term as in the computation bound. An index of resolvability bound with computational accuracy modification easily follows.

1.5 Summary

We propose a greedy algorithm for ℓ_1 penalized likelihood estimation and demonstrate the risk bound for such estimators. We obtain an explicit guarantee of the computational accuracy of our iteratively computed estimators, which is inverse proportional to the number of iterations. The risk bound is shown to be optimal sum of the approximation error and the ℓ_1 penalty. The valid penalizing parameter λ_n required to satisfy the risk result has to exceed the order $\sqrt{n \log M}$. In particular the *i.i.d.* modeling case reveals a choice of order $\left[\frac{\log M}{n}\right]^{1/2}$ where the risk bound shares the same rate. The computation and risk results together suggest valid approximate computation when the computational accuracy term is comparable or smaller than the risk terms. The number of steps required using our algorithm for computing such valid approximate estimators is far smaller than the sample size n or the variable size M. We will consider important applications of the computation and risk results in following chapters.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 2

Adaptive ℓ_1 Penalized Least Squares

2.1 Introduction

Consider the problem of estimating a mean response $X\beta$ from the *i.i.d.* data $Y_1, \ldots, Y_n$ following the linear model

$$Y = X\beta + z$$

where X is an $n \times p$ matrix of explanatory variables, β is a p-dimensional vector of interest, and z are independent random errors. We will assume each z_i follows the Gaussian distribution $N(0, \sigma^2)$, though our theory could easily accommodate other distributions as well. The mean vector $X\beta$ presented here is for convenience only and we do not restrict the population mean to have such a linear form. We assess the statistical risk of the estimator $X\hat{\beta}$ by the expected mean square error, for a fixed design X; results for a random design are implied by this work. A popular scenario we are accustomed to think is the $p > n$ situation, and this motivates this work but is not an imposed condition.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

A widely studied technique to handle this situation is the ℓ_1 penalized least squares criterion, also known as LASSO [57] or Basis Pursuit [22], which seeks the coefficients β to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where $\lambda \geq 0$ is an appropriate penalization parameter. For simplicity, we assume the predictors X and response Y are properly standardized with mean zero, and all predictors are normalized to have $(1/n)\sum_{i=1}^{n} x_{ij}^2 = 1$ for every $j = 1, \ldots, p$.

An important property of ℓ_1 penalized least squares is that the resulting estimator of β is nonzero only at a few coordinates, and the sparsity is controlled by the choice of penalization parameter λ. The discussions on the optimal choice of λ diverge in the past literature.

In practice, one may conveniently use cross validation to pick λ minimizing the prediction error. The resulting coefficients can have more nonzero coordinates than the population counterparts, and it is known in theory that having extra variables correlated with those in the true model can help the prediction performance. Cross validation is an computationally intensive practice with large data sets having n or p in the hundreds of thousands. This usually involves solving the ℓ_1 penalized least squares criterion for several folds of training data sets with various λ, and the λ with smallest average prediction error on corresponding testing data sets is chosen to produce the fitted coefficients.

The theoretical analysis studies various performance aspects of ℓ_1 penalized least squares under conditions of the design matrix X as well as the choice of λ. The recent literature [36, 26, 16, 17, 21, 39, 38, 52, 67, 68, 42, 59, 14, 18] in this direction demonstrated that ℓ_1 penalized least squares is effective in some nice situations. A

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

typical analysis assumes that a risk optimal λ is given, or relates λ to underlying parameters in a manner leading to a successful analysis. Such assumptions could hardly be checked in practice, and thus hardly address the problem of choosing λ in practical situations.

One popular choice of λ shown by theory is proportional to a constant variance σ with a critical rate $\sqrt{\log p/n}$, for example in [16, 17, 67, 42, 20, 14]. We obtain a similar form of λ with clean constants, and the statistical risk bounds for finite samples. Our results are different in several important ways. We don't require the assumption of an almost independent design matrix; the constants we obtain are explicit and applicable in practice; and the case with unknown σ, common in practice, can be handled in our framework where a risk bound with the estimated σ is obtained.

How to estimate σ alone in the regression setting with $p > n$ is an important question. One may favor a good estimator of σ for use in statistical inference. The estimator we provide is closely connected to the ℓ_1 penalized estimation of β, where the general likelihood principle is considered. Indeed, it is a generalized version of the conventional estimator of the residual sum of squares with the modification of the ℓ_1 penalty.

In the penalized least squares setting, adaptation as well as new penalty forms are important topics in the statistics literature, for example SCAD in [30, 31], Adaptive LASSO in [69, 65], Group LASSO [63] and Fused LASSO [33]. Though we consider here mainly the ℓ_1 penalty, other suitable forms satisfying our information theoretic condition could also be studied in our framework, as well as their adaptation to data.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2.2 Risk Analysis

We first consider a fixed σ that is assumed to be known. The proof for the risk is an application of the log-density estimation results. From now on, to be consistent with the log-density analysis, we use f^* in place of $X\beta$ for the underlying signal, and accordingly we consider the estimator of the form $f_\theta = \sum_h \theta_h h$. We take the empirical ℓ_2 norm squares to be $\|h\|_{\underline{x}}^2 = (1/n)\sum_{i=1}^n h^2(x_i)$, and denote the ℓ_1 norm of the coefficients by $\|\theta\|_1 = \sum_h |\theta| \|h\|_{\underline{x}}$. Here we could apply the density results in the chapter 1, but we use the log density results in [10] to show better constants. Consequently, we show the following theorem on ℓ_1 penalized least squares.

Theorem 2.2.1. *The ℓ_1 penalized least squares estimator $\hat{f} = f_{\hat{\theta}}$ achieving*

$$\min_\theta \left\{ \frac{1}{n}\sum_{i=1}^n (y_i - f_\theta(x_i))^2 + 2\sigma\frac{\lambda_n}{n}||\theta||_1 \right\}$$

has the risk bound

$$\mathbb{E}\|\hat{f} - f^*\|_{\underline{x}}^2 \le 2\inf_\theta \left\{ \|f_\theta - f^*\|_{\underline{x}}^2 + 2\sigma\frac{\lambda_n}{n}\|\theta\|_1 \right\} + \frac{8\sigma^2\log(2p)}{n}$$

for every sample size, provided that $\frac{\lambda_n}{n} \ge \left[\frac{2\log(2p)}{n}\right]^{1/2}$.

The penalizing parameter suggested by theory is $\frac{\lambda_n}{n} = \left[\frac{2\log(2p)}{n}\right]^{1/2}$ which reveals the smallest order of risk bounds, which is also the critical rate for these problems. The constants in the bound are explicit for every sample size and the bound also has the advantage of trading off optimally the approximation error (between f_θ and the underlying f^*) and the ℓ_1 penalty term. For people with more theoretical taste, it is pointed out in the manuscript [42] that an improved rate can be shown, taking advantage of this tradeoff, as long as certain refined covering properties exist.

We now consider the case of unknown σ. The likelihood principle would suggest

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the optimization over σ as well as θ. This is the approach we take to analyze the problem, leading to the following theorem.

Theorem 2.2.2. *The ℓ_1 penalized least squares estimator $\hat{f} = f_{\hat{\theta}}$ achieving*

$$\min_{\theta,\sigma} \left\{ \frac{1}{2\sigma^2}(1+\frac{1}{n})\frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2 + \frac{\lambda_n}{n\sigma}||\theta||_1 + \frac{\log(\sigma^2)}{2}(1+\frac{4}{n}) \right\}$$

has the risk bound

$$\frac{1}{n}\mathbb{E}d(P_{Y|x,f^*,\sigma_*}, P_{Y|x,\hat{f},\hat{\sigma}}) \leq \inf_{\theta,\sigma^2} \left\{ \frac{1}{n}D(P_{Y|x,f^*,\sigma_*}||P_{Y|x,f_\theta,\sigma}) + \frac{\lambda_n}{n\sigma}\|\theta\|_1 + \frac{\|y-f_\theta\|_x}{2\sigma^2 n} + \frac{2\log\sigma^2}{n} \right\} + \frac{2\log(4pn)}{n}$$

for every sample size, provided that $\frac{\lambda_n}{n} \geq (2+\frac{1}{n})\sqrt{\log(2p)/n}$.

Comment on computation for regression: The optimization producing $\hat{\theta}, \hat{\sigma}^2$ is reasonably straightforward. Each value of σ^2 corresponds to a multiplier of the ℓ_1 penalty. For each value of σ^2 one may optimize over θ by standard ℓ_1-penalized least squares algorithms. A particularly fast theoretical method with computational guarantees is the greedy algorithm in the manuscript [42], and new modified proposals with real computational advantages in practice are included in this dissertation in chapter 3. The log-likelihood version has also been discussed where, rather than picking the multiplier by some auxiliary cross-validation method, MDL chooses it (or equivalently chooses the single parameter σ^2) to optimize the above criterion.

Alternatively, we note that for θ the best $\sigma^2 = \sigma^2_{f_\theta}$ solves a quadratic

$$(1+\frac{4}{n})\sigma^2 = \sigma\frac{\lambda_n}{n}\|\theta\|_1 + (1+\frac{1}{n})\|y - f_\theta\|^2_x.$$

Then one may plug in the solution $\sigma^2_{f_\theta}$ and optimize the resulting function of θ.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2.3 Proofs

The risk analysis is built on the earlier result for log-densities and we will only briefly sketch the proofs and results here. We will use $f(x)$ to denote the fitted value $x\beta$ from now on, in accordance with the general theory on $f = \sum_h \theta_h h$ for $h \in \mathcal{H}$. The key step in the proof is to check that the condition $(**)$ holds for a suitable representer $\tilde{f}$. As detailed in [10], we here demonstrate the existence of such a representer $\tilde{f}$ by showing that the condition holds for a random $\tilde{f}$.

The countable set $\tilde{\mathcal{F}}$ of representers is taken to be the set of all functions of the form $\tilde{f}(x) = V\frac{1}{K}\sum_{k=1}^{K} h_k(x)/a_{h_k}$ for terms h_k in $\mathcal{H} \cup -\mathcal{H} \cup \{0\}$, where the number of terms K is in $\{1, 2, \ldots\}$ and the nonnegative multipliers V will be determined from K in a manner we will specify later. We let p be the cardinality of $\mathcal{H} \cup -\mathcal{H} \cup \{0\}$, allowing for h or $-h$ or 0 to be a term in $\tilde{f}$ for each h in $\mathcal{H}$.

The main part of the codelength $L(\tilde{f})$ is $K \log p$ nats to describe the choices of $h_1, \ldots, h_K$. The other part is for the description of K and it is negligible in comparison, but to include it simply, we may use a possibly crude codelength for the integer K such as $K \log 2$. Adding these contributions of $K \log 2$ for the description of K and of $K \log p$ for the description of $\tilde{f}$ given K, we have

$$L(\tilde{f}) = K \log(2p).$$

To establish existence of a representer $\tilde{f}$ of f with the desired properties, we put a distribution on choices of $h_1, \ldots, h_K$ in which each is selected independently, where h_k is h with probability $|\theta_h| a_h / V$ (with a sign flip if θ_h is negative). Here $K = K_f = \lceil V_f/\delta \rceil$ is set to equal V_f/δ rounded up to the nearest integer, where $V_f = \sum_h |\theta_h| a_h$, where a small value for δ will be specified later. Moreover, we set $V = K\delta$, which is V_f rounded up to the nearest point in a grid of spacings δ. When

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

V_f is strictly less than V there is leftover an event of probability $1 - V_f/V$ in which h_k is set to 0.

As f varies, so does the complexity of its representers. Yet for any one f, with $K = K_f$, each of the possibilities for the terms h_k produces a possible representer $\tilde{f}$ with the same complexity $K_f \log 2p$.

The key property of our random choice of $\tilde{f}(x)$ representing $f(x)$ is that, for each x, it is a sample average of i.i.d. choices $Vh_k(x)/a_{h_k}$. Each of these terms has expectation $f(x)$ and variance $V \sum_h |\theta_h| h^2(x)/a_h - f^2(x)$ not more than V^2.

As the sample average of K such independent terms, $\tilde{f}(x)$ has expectation $f(x)$ and variance $(1/K)$ times the variance given for a single draw. We will also need expectations of exponentials of $\tilde{f}(x)$ which is made possible by the representation of such an exponential of sums as the product of the exponentials of the independent summands.

2.3.1 Proof of Theorem 2.2.1

Consider the linear regression case with fixed design first. At each x_i we seek a fit $f(x_i)$ to a corresponding outcome Y_i. We use the Gaussian model of independent outcomes $Y_1, \ldots, Y_n$ with joint density function

$$p_f(\underline{y}|\underline{x}) = \frac{1}{(2\pi\sigma^2)^{(n/2)}} exp\left\{ -\frac{\sum_{i=1}^n (y_i - f(x_i))^2}{2\sigma^2} \right\}.$$

The case of fixed (known) variance σ^2 is considered first. In this Gaussian regression setting, the divergence $d(P_{\underline{Y}|\underline{x},f^*}, P_{\underline{Y}|\underline{x},f})$ for fixed $\underline{x}$ can be written explicitly as

$$\frac{1}{4\sigma^2} \sum_{i=1}^n (f(x_i) - f^*(x_i))^2.$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Then in accordance with (**) we check the validity of a penalty $pen(f)$ by verifying, for a suitable representer $\tilde{f}$, that

$$pen(f) \geq 2L(\tilde{f}) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(y_i - \tilde{f}(x_i))^2 - (y_i - f(x_i))^2\right] - \frac{1}{4\sigma^2}\sum_{i=1}^{n}\left[(f^*(x_i) - \tilde{f}(x_i))^2 - (f^*(x_i) - f(x_i))^2\right].$$

In this section we adapt the general strategy developed in the previous section to the regression setting to demonstrate that the ℓ_1 penalty on coefficients with suitable multipliers is also an information-theoretic penalty for regression. The result presented here is fascinating as it also reveals what penalty parameter λ should employed for ℓ_1 penalized regression to be justifiable for the MDL interpretation and statistical risk analysis.

We allow the weights a_h in this section to be the empirical ℓ_2 norm $\|h\|_{\underline{x}}$ where $\|h\|_{\underline{x}}^2 = \frac{1}{n}\sum_i h^2(x_i)$ instead of $\|h\|_\infty$ in the density case. We no longer need a bounded range condition nor an appeal to the Hoeffding inequality. The same sampling strategy for generating a random $\tilde{f}$ also applies here.

We bound similarly the minimum over $\tilde{\mathcal{F}}$ of the complexity-penalized discrepancy difference by the quantity obtained by the sample average of randomly selected h_k. For the discrepancy difference, adding and subtracting $f(x_i)$ in each square, the squared terms of $y_i - f(x_i)$ and $f^*(x_i) - f(x_i)$ cancel out when expanding out the squares and their cross product terms with $(f(x_i) - \tilde{f}(x_i))$ vanish in expectation under the random $\tilde{f}(x_i)$. What remains for the expected discrepancy difference is the expectation of

$$\frac{1}{4\sigma^2}\sum_{i=1}^{n}(\tilde{f}(x_i) - f(x_i))^2.$$

Each summand $(\tilde{f}(x_i) - f(x_i))^2$ for fixed x_i under random $\tilde{f}$ has mean not more than

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the $(1/K)$ times the bound $V\sum_h |\theta_h| h^2(x_i)/a_h$ on the variance given for a single draw h. The aggregated bound over x_i yields

$$\frac{V}{4\sigma^2 K}\sum_{i=1}^{n}\sum_{h}|\theta_h|h^2(x_i)/a_h = \frac{nVV_f}{4\sigma^2 K}$$

where n appearing in the equality is by the fact that $\sum_{i=1}^{n} h^2(x_i)/a_h^2 = n$ for each h.

Now the discrepancy difference plus twice the complexity penalty is bounded by

$$2K\log(2p) + \frac{nVV_f}{4\sigma^2 K}.$$

With our choice of $K = \lceil V_f/\delta \rceil = V/\delta$ not more than $V_f/\delta + 1$, we show that the penalty of the form

$$pen(f) \geq \lambda V_f + C$$

is valid as long as λ is not smaller than $2V_f(\log 2p)/\delta + nV_f\delta/(4\sigma^2)$ and $C = 2\log(2p)$. Setting $\delta = 2\sigma(2(\log 2p)/n)^{(1/2)}$ to optimize the bound for λ, the critical value is $\lambda^* = (2n\log(2p))^{(1/2)}/\sigma$ and our analysis shows that ℓ_1 is valid as long as the penalty parameter exceeds λ^*.

2.3.2 Proof of Theorem 2.2.2

Next we generalize the result to the unknown σ case. Following the MDL principle we are motivated to estimate $(\hat{f}, \hat{\sigma}^2)$ by optimizing

$$\frac{1}{2\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{\sigma}\frac{\lambda_n}{n}\|\theta\|_1 + \frac{pen(\sigma^2)}{n}$$

where the first two terms form the $-\frac{1}{n}\log$ *likelihood* and the next term is the penalty used in the fixed σ^2 case. With $pen(f, \sigma^2)$ similar to this, we show that such an

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

optimization indeed satisfies the requirement (**) for validity of statistical risk analysis. For the representer $(\tilde{f}, \tilde{\sigma}^2)$ in a countable cover, we adapt the same strategy of random K-term $\tilde{f}$ and use for $\tilde{\sigma}^2$ of a logarithmic discretization of σ^2, that is, $\log \tilde{\sigma}^2 = \lfloor (\log \sigma^2)/\epsilon \rfloor \epsilon = K'\epsilon$ with the choice of ϵ to be specified and K' an integer. We set the codelength in this case to be $L(\tilde{f}, \tilde{\sigma}^2) = K \log(2p) + 2\log(K'+1)$ where we crudely encode K' by $2\log(K'+1)$ for simplicity. The Bhattacharyya, Renyi, Hellinger divergence $d(P_{\underline{Y}|\underline{X},f^*,\sigma_*}, P_{\underline{Y}|\underline{X},f,\sigma})$ can be written explicitly as

$$\frac{1}{2(\sigma^2 + {\sigma_*}^2)} \sum_{i=1}^{n} (f(x_i) - f^*(x_i))^2 + \log \frac{\sigma^2 + \sigma_*^2}{2\sqrt{\sigma^2 \sigma_*^2}}.$$

Now checking (**) involves the difference of these divergences at (f, σ^2) and at $(\tilde{f}, \tilde{\sigma}^2)$ as well as the differences in the log-likelihood. Some of the resulting terms in the difference are negative (and can be dropped) because of our choice of $\tilde{\sigma}^2$ not more than σ^2 (by rounding down). What remains to verify is that

$$pen(f, \sigma^2) \geq 2L(\tilde{f}, \tilde{\sigma}^2) + \sum_{i=1}^{n} \left[\frac{(y_i - \tilde{f}(x_i))^2}{2\tilde{\sigma}^2} - \frac{(y_i - f(x_i))^2}{2\sigma^2} \right] - \sum_{i=1}^{n} \left[\frac{(f^*(x_i) - \tilde{f}(x_i))^2}{2(\sigma_*^2 + \tilde{\sigma}^2)} - \frac{(f^*(x_i) - f(x_i))^2}{2(\sigma_*^2 + \sigma^2)} \right].$$

To show existence of a suitable representer $\tilde{f}$ we bound again the sample average version. The same bound for $(\tilde{f}(x_i) - f(x_i))^2$ is used and we drop all non-positive terms for cleanness. The discrepancy difference plus twice the complexity is then bounded by

$$\frac{1}{2\sigma^2} \left[(e^\epsilon - 1) \sum_{i=1}^{n} (y_i - f(x_i))^2 + e^\epsilon \frac{nVV_f}{K} \right] + 2K \log(2p) + 2\log(K'+1).$$

With $K' = \lfloor (\log \sigma^2)/\epsilon \rfloor \leq (\log \sigma^2)/\epsilon$ and $K = V/\delta \leq V_f/\delta + 1$, we set $\delta =$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$2\sigma((\log 2p)/n)^{1/2}e^{-\epsilon/2}$ to optimize the bound first assuming fixed ϵ. For simplicity, we pick $\epsilon = 1/(2n)$ to optimize over ϵ crudely and use $e^{1/2n} < 1+1/n$ to simplify the multiplying constants. The resulting satisfactory penalty requirement takes the form

$$pen(f, \sigma^2) \geq \frac{1}{2\sigma^2}\|y - f\|_{\underline{x}}^2 + \frac{\lambda V_f}{\sigma} + 2\log\sigma^2 + 2\log(4pn),$$

valid as long as $\lambda \geq (2 + \frac{1}{n})\sqrt{n\log(2p)}$, where we denote $\|y - f\|_{\underline{x}}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2$.

2.4 A Numerical Example

Consider a simulation case with sparse signals first. Many other examples as well as real data sets are currently being studied, and could be added in the future for a full journal paper.

The design matrix we study is generated from correlated Gaussians, where the correlation between x_j and $x_{j'}$ is $1/2^{|j-j'|}$. We consider a sparse β with almost all zeros except that $\beta_1 = 3$, $\beta_2 = 1.5$, and $\beta_5 = 2$. The normal error z has a variance such that the signal-to-noise ratio is 3.

We compare the performance of our adaptive procedures, for λ with known and unknown σ, to the one with λ picked solely by cross validation. The measures of estimation performance illustrated here include the Euclidean distance of fitted value, the same distance for β, and the Hamming distance for identifying the support of β as well.

As pointed by several authors, for example [21, 67, 20] (under assumptions), a good estimation of the support using ℓ_1 penalized least squares could lead to an improved estimation of the fitted value if an ordinary least squares on the identified support is used as a second stage procedure to recompute the estimated β. Based on

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the set of nonzero coefficients produced by ℓ_1 penalized least squares, with λ chosen by cross validation or our procedures, we regress on the same set to recalculate the coefficients as a second stage correction. We will also impose such a second stage procedure for our adaptive estimation as well as for cross validation.

We compare the performance with large $p = 2000$ and varying n, and we run the simulation for 1000 repetitions. We will use CV to denote the results by cross validation, and APR(σ) and APR for our adaptive penalized regression procedures with known and unknown σ, respectively.

From tables 2.1-2.3, it can be seen for all performance measures that our proposal has great advantages for intermediate sample sizes and is comparable to cross validation for small sample sizes. Cross validation tends to select more variables than the true model as the sample size increases, where our methods outperform it significantly for all sample sizes. The procedure with unknown σ is comparable to the procedure with fixed σ, though the latter one has slight performance improvement. Two-stage correction does improve the estimation performance and our proposals with this correction greatly beats cross validation across all sample sizes.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

n	50	100	200	400
CV	23(24)	34(43)	46(71)	69(130)
APR(σ)	**2(2.2)**	**1(0.78)**	0(0.43)	0(0.27)
APR	3(2.9)	2(1.5)	**0(0.021)**	**0(0.005)**

Table 2.1: Median(mean) of the Hamming distance $\sum_{j=1}^{p} |\{\hat{\beta}_j = 0\} - \{\beta_j^* = 0\}|$.

n	50	100	200	400
CV	**4.6(0.92)**	**3.3(0.75)**	2.6(0.96)	2.1(1.2)
APR(σ)	4.9(0.85)	3.4(0.58)	**2.3(0.39)**	**1.4(0.24)**
APR	8.7(0.14)	7(0.39)	4.9(0.36)	2.6(0.23)
		Two-stage Correction		
CV-ts	5.4(0.88)	4.4(0.64)	4.1(0.74)	4(1.3)
APR(σ)-ts	**3.8(1.5)**	**1.8(0.91)**	1.1(0.6)	0.62(0.34)
APR-ts	8.5(0.74)	5.4(0.82)	**0.8(0.53)**	**0.48(0.22)**

Table 2.2: Mean(sd) of $\|\hat{\beta} - \beta\|^2$. All numbers reported here are multiplied by 100.

n	50	100	200	400
CV	**28(6.3)**	**16(5.4)**	9.5(4.5)	4.6(2.7)
APR(σ)	33(5.6)	17(2.7)	**8.1(1.3)**	**3.2(0.54)**
APR	65(3.8)	37(1.9)	18(0.99)	6.1(0.49)
		Two-stage Correction		
CV-ts	35(4.9)	23(4)	15(3.4)	8.6(2.7)
APR(σ)-ts	**21(7.2)**	**7.4(3.7)**	3.2(1.8)	1.1(0.63)
APR-ts	61(11)	22(3.6)	**2.2(1.4)**	**0.87(0.38)**

Table 2.3: Mean(sd) of $\|X\hat{\beta} - X\beta^*\|^2$. All numbers reported are multiplied by 100.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 3

Computation of ℓ_1 Penalized Least Squares

3.1 Introduction

Modern technology and computerized sampling enable us to collect humongous data sets in many scientific and technological studies. For example, microarray data have tens of thousands of genes, while Google is web crawling billions of web pages. Statistical inference provides vital tools for answering important questions in those fields, though traditional statistical computation may not be easily applied due to the expensive computational cost. For example, ordinary linear regression requires solving a large linear system. When the data are in the tens of thousands, the regression fit may not be computed in a timely manner or the package may simply refuse to run due to the memory shortage. There are two basic constraints in computational statistics, time and memory costs. Our algorithms achieve a balance of both and they have explicit guarantee of computational accuracy.

Consider an observed data matrix X of size $n \times p$, corresponding to n observations

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and p predictors, and a response vector Y of size n. The LASSO [57] criterion, also known as Basis Pursuit [22], is to fit the coefficient vector β to the ℓ_1 regularized least squares

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3.1.1}$$

where β_0 is the intercept. Here, we do not penalize the intercept and each predictor is standardized to have unit ℓ_2 norm. Without loss of generality, we assume the response and predictors are centered so the optimal intercept is $\beta_0 = 0$ and can be ignored.

Unlike the ridge regression with ℓ_2 penalty, which shrinks the coefficients proportionally towards zero, the LASSO solution shrinks a equal amount of the ordinary least squares fit towards zero, thus it may produce sparse solutions even for large p. Though bias is introduced by the penalty term while reducing the variance, the prediction performance is enhanced as with other shrinkage estimations. The prediction performance is empirically studied in [57] and theoretically justified in [17, 42]. Another feature of LASSO is its ability to do automatic variable selection, even in the situation where the number of observations n is far smaller than the number of predictors p. This has been illustrated in many practical situations and recently studied in theory by [52, 65, 68].

Solving the LASSO criterion in general is not trivial computationally even though the objective is convex in β. Generic convex programming or quadratic programming, such as interior point methods [15] for example, can be used for solving a general design X but is reported to be much slower than other alternative algorithms devised specifically for LASSO, see the survey [32].

Iterative procedures have been well known in statistics for efficiently solving such problems, particularly forward procedures that add one term at a time into the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

model, building up the model from a simple starting one. The computational cost and memory requirements can be greatly reduced when simple calculations are employed at each iteration, as compared to fitting all coordinates simultaneously. For example, Forward Selection, or Forward Stepwise Regression, described in [60], selects the predictor that has the largest absolute correlation with the response y at the first iteration, say x_{j_1}. Regress y on x_{j_1}, using a simple linear model, and the residual vector is treated as the new response. It then projects all other predictors orthogonally to x_{j_1} and repeats the process. The k predictors selected after k steps are used to construct the linear model. However, Forward Selection can be overly greedy in the sense that other useful predictors outside the model have less chance to be selected at a later stage if they are highly correlated with the ones previous selected for the model.

The least angle regression [28], henceforth LARS, is a more cautious version of Forward Selection and its modification solves the LASSO problem. Similar proposals of LARS are in [53, 54]. In the first iteration, LARS employs a smaller coefficient $\hat{\gamma}_{j_1}$ than the ordinary least squares coefficient $\hat{\beta}_{j_1}$ such that the angle of x_{j_1} and a next correlated predictor, say x_{j_2}, is bisected by the residual projected on the linear span of x_{j_1} and x_{j_2}. The LARS fitted vector then moves along the bisector until the projected residual is equally correlated between x_{j_1}, x_{j_2} and a third predictor, say x_{j_3}. It repeats this procedure while adding the corresponding bisector at each iteration until the projected residual bisects the angles between all the predictors in the model. Another procedure similar to LARS is Forward Stagewise Regression. At each iteration, it moves along the direction of the selected predictor chosen by the same correlation criterion as Forward Selection, but with a tiny constant multiplier ϵ instead of the least squares coefficient.

These two procedures are thought to improve upon Forward Selection in the sense

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

that the multiplier for a new predictor to be added to the model is down scaled to give other predictors more chance to be included in the model, perhaps with a larger coefficient, at a later stage. Instead of down-weighting the coefficient of the predictor to be added, we down-weight, or "relax", the previous fit, while applying the best possible coefficient selected by our objective. It is closely related to the relaxed greedy algorithm [45, 6] in approximation theory, which solves the least squares without penalty, but here is mainly devised for the LASSO that includes an ℓ_1 penalty on the coefficients. As with the other two modifications, it avoids the overly greediness of Forward Selection as the other two modifications, and moreover by relaxation we have explicit control over the computational accuracy for any iterative step k.

While LARS is very efficient for LASSO if the data size is moderate, the performance of LARS deteriorates sharply when the data set is on the scale of tens of thousands. As reported in [33] and its precursors [37, 25], the coordinate descent algorithm is very competitive with LARS in general and outperforms it dramatically for large problems. Coordinate descent is easy to implement and its computation can be organized as a closed form calculation at each iteration. However, the number of iterations needed for convergence is unclear. We differ from coordinate descent in two ways at each iteration: we pick the best coordinate (as do other forward procedures) instead of cycling through one predictor at a time; and we down-weight the previous fit by $(1 - \alpha)$ instead of fixing all but one coordinates of the previous fit. The calculation of the down-weighting factor is in a closed form as will be shown later and fewer iterations are required than for coordinate descent, so that the total computation is particularly efficient.

Our algorithms have explicit control of the global computational accuracy for finite k, iterations while a similar analysis of other procedures is not readily accessible to the best of our knowledge. Interior point methods can be shown to converge lin-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

early [15] but may require a constant that is a high order (perhaps three) polynomial of the number of parameters, and the coordinate descent procedure may have limited convergence results using the general theorems in [58]; the global rate is unclear for finite numbers of iterations.

For statistical problems, we argue that extremely accurate solutions are not necessary. The effect of computational improvements in prediction can be dominated by other sources of statistical variability. It is well understood that the prediction error is comprised of computation, approximation and estimation errors [4]. The latter two together are of order $\sqrt{(\log p)/n}$ for the LASSO problem [17, 42]. Therefore, a solution with a computational accuracy matching the order $\sqrt{(\log p)/n}$ is sufficient for the whole prediction error to be of the same order. We establish theoretically that the global computation error of our algorithms is decaying with order $1/k$. With the explicit guarantee on the computation error, we have the potential to solve extremely large problems, where by choosing to solve them approximately, the statistical performance does not suffer from the use of an approximate solution. Other algorithms fail to characterize their performance if not convergent numerically in finite time.

Fast algorithms solving for LASSO are also important for other statistical problems. In particular, some complex problems can be solved by iterative algorithms with the LASSO problem being the inner iteration. For example, [52] proposes to construct a sparse graphical model by regressing each variable on the others using the LASSO model, and [1] shows that the maximum likelihood criterion in the Gaussian inverse covariance estimation can be converted exactly to LASSO by a dual argument. These, for instance, reduce the complexity and memory requirement compared to the original problems. According to their constructions, we solve the original problem by fitting a LASSO model in a loop and iterate over different LASSO problems. Indeed, decomposing a large problem into iterative LASSO models can

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

be very efficient, as shown for the covariance estimation reported in [35]. Accurate solutions may not be needed until near convergence for the outer iterations. Approximate solutions in the early iterations are acceptable since the LASSO problem will vary as the outer iterations move forward. Our algorithms exhibit the speediness and computational accuracy control, and are are therefore promising to be embedded for complex problems.

This chapter is organized as follows. We will formally introduce our algorithms in section 3.2 and state our main theorems on computational accuracy. The basic algorithms solve for a fixed λ but will be extended by a path-following argument in section 3.3. We compare our algorithms with some existing packages on simulated data in section 3.4 and conclude in 3.5.

3.2 Algorithm

In this section, we first introduce the ℓ_1 penalized greedy pursuit algorithm from [42] and its corresponding computation guarantee. I call their ℓ_1 penalized greedy pursuit *approximate* LPGP here to differentiate with the improved versions I develop here. My variants improve over approximate LPGP in two directions, strategy-wise, and I will address extensive numerical proposals next for efficient implementations. One can also optimize over the exact ℓ_1 norm at each iteration rather than using the bound of the norm allowed by the computational theorem. I show that exact optimization of the relaxation parameter over the full range is a computational improvement.

The approximate LPGP algorithm builds up the fitted coefficient β for the LASSO criterion using iterative steps. Starting from $\beta^{(0)} = 0$, we seek the k^{th} step fit of the form $\beta^{(k)} = (1-\alpha)\beta^{(k-1)} + \gamma I_l$ where the relaxation parameter is $\alpha \in [0,1]$, the new coefficient is $\gamma \in R$ and the index vector I_l is a zero vector except at the l^{th}

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

component, where it is 1, for $l = 1, 2, \ldots, p$. One chooses the optimal $\alpha^{(k)}$, $\gamma^{(k)}$ and $I_{l^{(k)}}$ to minimize the approximate-norm LASSO criterion

$$L^{(k)}(\alpha, \gamma, l) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (1-\alpha)\sum_{j=1}^{p} x_{ij}\beta_j^{(k-1)} - \gamma x_{il}\right)^2 + \lambda\left[(1-\alpha)v^{(k-1)} + |\gamma|\right] \tag{3.2.1}$$

where $v^{(k-1)} = \sum_{j=0}^{p}\left|\beta_j^{(k-1)}\right|$ is the ℓ_1 norm of coefficients at iteration $k-1$. It is approximate LPGP since the term $[(1-\alpha)v^{(k-1)} + |\gamma|]$ representing the ℓ_1 norm is approximate. Indeed, it is the exact ℓ_1 norm of the resulting $\beta^{(k)}$ when $\beta_l^{(k-1)}$ has the same sign as γ or $\beta_l^{(k-1)} = 0$. It differs by $2|\gamma|$ when $\beta_l^{(k-1)}$ has the opposite sign of γ.

Through the iterations, the objective will have a value not larger than that of the previous iteration since the minimal objective achieved at $(\alpha^{(k)}, \gamma^{(k)}, l^{(k)})$ will have the objective value not bigger than $L^{(k)}(1, 0, l)$ for any l, which is not larger than the minimal $L^{(k-1)}$. Moreover, one can bound the objective improvement with respect an arbitrary reference so that the closeness of the approximate LPGP fit to the infinimum can be shown. A simplified version of the result in [42] is reported here for convenience.

Theorem 3.2.1. *The k step fit $\beta^{(k)}$ from the approximate LPGP algorithm achieves that the computational accuracy with difference from the solution that decays with order $1/k$. Indeed*

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j^{(k)}\right)^2 + \lambda v^{(k)} \leq \inf_{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda V_\beta + \frac{4V_\beta^2}{k+1}\right\}$$

where $V_\beta = \sum_{j=1}^{p}|\beta_j|$.

The proof of the theorem is obtained by induction. One bounds the difference in objective values between the k^{th} step $L^{(k)}(\alpha, \gamma, l)$ and an arbitrary reference by a

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

particular choice of α, γ and random l, and then shows that the bounding difference decays with order $1/k$. Any iterative algorithm that achieves an objective value no larger than $L^{(k)}(\alpha, \gamma, l)$ at the k^{th} step will have the computational accuracy as stated in the theorem. The difference between the approximate and exact norm is an insignificant term by construction in the proof. However, the computational cost of tracking the exact norm is low, and we can do better numerically with this exact norm tracking. We advocate a modified algorithm, which we call *exact* LPGP, which tracks the ℓ_1 norm exactly.

The exact LPGP algorithm follows the same iteration strategy as the approximate version except that the objective at each iteration is to minimize the exact LASSO criterion

$$L^{(k)}_{exact}(\alpha, \gamma, l) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (1-\alpha)\sum_{j\neq l} x_{ij}\beta_j^{(k-1)} - \gamma x_{il}\right)^2 + \lambda\left[(1-\alpha)\tilde{v}_l^{(k-1)} + |\gamma|\right]$$

where $\tilde{v}_l^{(k-1)} = \sum_{j\neq l}\left|\beta_j^{(k-1)}\right|$. What is different here is the new fit at the next step, $\tilde{\beta}_j^{(k)} = (1-\alpha)\beta_j^{(k-1)}$ for $j \neq l$, where l is the fixed coordinate index minimizing the criterion, and $\tilde{\beta}_l^{(k)} = \gamma$. Such a modification is necessary only when the previous fit $\beta_l^{(k-1)} \neq 0$; it is equivalent to (3.2.1) when $\beta_l^{(k)} = 0$. In a sparse situation, such a computation overhead for removing the l^{th} nonzero coordinate in the current model is only increased by the order of sparsity. Moreover, by carefully organizing the factors in the objective (to be shown later), we show that only a few simple operations are needed to compute the exact objective.

The exact algorithm has the ability to eliminate a coordinate of the previous iteration if the optimizing γ equals 0. It enjoys the benefit of tracking the exact ℓ_1 norm of our fitted coefficients. This deviates from the basic LPGP algorithm which shrinks a nonzero coordinate by $(1-\alpha)$ at each iteration and variable elimination

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

will take quite a number of steps if a nonzero component is wrongfully included in the early iterations.

We can allow unrestricted $\alpha \in \mathbb{R}$ instead of limiting it to be within the range $[0,1]$. One customarily thinks of $(1-\alpha) \in [0,1]$ first as a relaxation parameter as in [45]. Values in this range play a role in the demonstration of the algorithmic results. By allowing the $\alpha \in R$ it is easier to solve an unconstrained optimization for α, though numerical studies show that the optimal α is within $[0,1]$ in most cases, with only occasional exceptions.

The choice of α can be more restricted since the objective is minimized in the previous iteration and the new fit should not deviate far from the previous one. We show that only positive $1-\alpha$ are feasible as solution for our objective. Indeed, let $L^{(k)}(\alpha)$ denote the minimal objective value of (γ, l) at iteration k (given the fit at $k-1$) with a fixed α. Suppose that it were the case that the optimal $1-\alpha^{(k)} \leq 0$, that is $\alpha^{(k)} \geq 1$, at iteration $k \geq 3$. By minimization and convexity then

$$L^{(k)}(\alpha^{(k)}) \leq L^{(k)}(1) = L^{(1)}(\alpha^{(1)}) \leq L^{(k)}(0) \leq L^{(k-1)}(\alpha^{(k-1)}).$$

Since the objective does not increase at any iteration (for all algorithms in this paper), the above line of inequalities all hold with equality. Objective values on successive iterations should equal each other whenever the optimal $1-\alpha^{(k)} \leq 0$. By ignoring uninteresting cases such as ties, we can that solution holds in the region where $1-\alpha > 0$. We call the algorithms with optimization of α over $(-\infty, 1)$ the *extended* algorithms.

Proposition 3.2.2. *The feasible range for α in the extended LPGP algorithms is $(-\infty, 1)$ when $k \geq 3$.*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Accordingly, the extended approximate LPGP at each iteration minimizes

$$L_{ext}^{(k)}(\alpha,\gamma,l) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (1-\alpha)\sum_{j=1}^{p} x_{ij}\beta_j^{(k-1)} - \gamma x_{il}\right)^2 + \lambda\left[|1-\alpha|\, v^{(k-1)} + |\gamma|\right]$$

over $(\alpha,\gamma) \in (-\infty,1)\times\mathbb{R}$ and $l \in \{1,\ldots p\}$ and the extended exact LPGP minimizes

$$L_{ext-exact}^{(k)}(\alpha,\gamma,l) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (1-\alpha)\sum_{j\neq l} x_{ij}\beta_j^{(k-1)} - \gamma x_{il}\right)^2 + \lambda\left[|1-\alpha|\, \tilde{v}_l^{(k-1)} + |\gamma|\right]$$

over the same ranges for (α,γ,l). We will examine various strategies of each in practice, as well as numerical tricks in the next section.

Our algorithms have smaller objective values at each iteration than the approximate LPGP would give if startted from the same coefficients given by the previous iteration. Thus, the same algorithmic bounds for the approximate LPGP would also hold. Numerically, the approximate algorithm is extremely slow compared to the alternatives we propose here, and it is still an open question how to demonstrate a faster rate or even a smaller constant in the computational bound.

Theorem 3.2.3. *The k step fit $\beta^{(k)}$ from exact LPGP, or extended LPGP, or extended exact LPGP, achieves computational accuracy with difference from the solution that decays with order $1/k$. Indeed,*

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j^{(k)}\right)^2 + \lambda v^{(k)} \leq \inf_{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda V_\beta + \frac{4V_\beta^2}{k+1}\right\}$$

where $V_\beta = \sum_{j=1}^{p}|\beta_j|$.

Our algorithms differ from the coordinate descent algorithm in two ways. First, we pick the best descending coordinate to modify instead of looping through all coordinates; coordinate sweeping according to the index order is inefficient. For

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

example, suppose the set T_1 of true nonzero coefficients are at later indexes in the sweep than the set of true zero coefficient T_0. Some computed coefficients of T_0 may be set to be nonzero in the first few rounds of sweeps and have to be reset to be zero only after subsequent rounds of sweeps. Such an inefficiency may be dramatic when the correlations between the variables are high. The magnitude of computed coefficients in T_0 may be comparable to the true coefficients in T_1 since they are highly correlated. Thus, more time may be needed to correctly adjust the coefficients of T_0 and T_1. We will show in our simulation study that coordinate descent may converge very slowly on highly correlated datasets. Moreover, from the theoretical prospective, our strategy of picking the best coordinate at a time will enable us to prove the computational accuracy bound for any finite k iterations. In contrast, the convergence analysis of coordinate descent falls short in this respect.

Finally, we differ from coordinate descent (and other pure greedy algorithms) by relaxing the previous fit by an optimal factor $(1 - \alpha)$ instead of fixing all but the one coordinate picked for updating. This enables our algorithm to update multiple coordinates (those in the current model, i.e. with nonzero coefficients) in addition to the one picked. This helps obtain more accurate coefficients before considering the addition of variables outside the current model. Indeed, this type of strategy will improve the computation performance in a manner similar to active shooting, which was found to be even faster than coordinate descent by [34]. The active shooting strategy loops through the current set of nonzero coordinates, called the active set, to achieve highly accurate coefficients before looping through the full set of coordinates to enter another variable into the model. Again theoretically, such a relaxation is advantageous for us as well, to achieve a better computational accuracy compared to other pure greedy algorithms.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.3 Computation

In this section, we will discuss the implementation of our algorithms and extend them to compute multiple penalty levels λ.

3.3.1 Analytical Minimization

The most essential computations are the minimization over α and γ for a fixed coordinate l at each iteration, and comparison to determine the minimal objective for varying l. All the algorithms at the k^{th} iteration require minimization of the objective form

$$L_f(\alpha,\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (1-\alpha)\, f_i^{(k-1)} - \gamma x_{il}\right)^2 + \lambda\left[|1-\alpha|\, v_f^{(k-1)} + |\gamma|\right] \qquad (3.3.1)$$

where $f_i^{(k-1)}$ could be either the previous fit $\sum_{j=1}^{p} x_{ij}\beta_j^{(k-1)}$ or the one $\sum_{j\neq l} x_{ij}\beta_j^{(k-1)}$ with l^{th} component removed. The corresponding ℓ_1 norm of the coefficients for $f^{(k-1)}$, denoted as $v_f^{(k-1)}$ is therefore either $v^{(k-1)} = \sum_{j=0}^{p}\left|\beta_j^{(k-1)}\right|$ or $\tilde{v}_l^{(k-1)} = \sum_{j\neq l}\left|\beta_j^{(k-1)}\right|$ respectively. For the minimization in (3.3.1), we can first consider each fixed coordinate $l \in \{1,\ldots,p\}$ and pick the l with the smallest objective value in (3.3.1) produced by the optimal α and γ. Thus the essential optimization is over α and γ when holding l fixed. This objective function is convex in (α,γ) where any general convex optimization algorithms could be applied. However, we will derive the closed form formulas for the optimal (α,γ) to achieve a better performance.

We will first consider the feasible range for α to be $\mathbb{R}$, because of its simplicity. The constrained version for $\alpha \in [0,1]$ (and others) can be then obtained by projecting the corresponding full range optimal (α,γ) into the constrained set. For a fixed α, as the LASSO solution in a single dimension, the minimizing γ in (3.3.1) is a soft-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

thresholded coefficient from regressing $y - (1-\alpha)f^{(k-1)}$ on $x_{\cdot l}$, see [27, 33]. On the other hand, the minimizing $\bar{\alpha} = 1 - \alpha$ (we work with $\bar{\alpha}$ instead of α since this simplifies the expression) for a fixed γ is also analogous to a LASSO solution in a single dimension. The optimization for $\bar{\alpha}$ and γ together is then a two dimensional extension of the known solution in a single dimension, and we will show that our optimization over $\bar{\alpha}$ and γ can be achieved by a closed form formula like the one obtained in one dimension. We denote the empirical ℓ_2 norm $\|z\|_2^2 = \frac{1}{n}\sum_{i=1}^n z_i^2$ for a vector z of length n and also the inner product $\langle y, z\rangle = \frac{1}{n}\sum_{i=1}^n y_i z_i$. For a general two-dimensional LASSO problem, the closed form minimization will be derived. The rest of the calculation involves applying these formulas to our setting.

First, suppose the minimizer occurs outside the axes (where both coordinates are nonzero), then the minimization is achieved at the point with zero derivatives.

Lemma 3.3.1. *If the minimizer for (β_1, β_2) is nonzero in the both coordinates that minimizes the LASSO criterion*

$$L(\beta_1, \beta_2) = \{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 + \lambda_1|\beta_1| + \lambda_2|\beta_2|\},$$

and x_1 and x_2 are linearly independent, then the solution set $(\tilde{\beta}_1^*, \tilde{\beta}_2^*)$ under these conditions can be written explicitly as

$$\left(\tilde{\beta}_1^*, \tilde{\beta}_2^*\right) = \begin{cases} (C_1 - S_1^+, C_2 - S_2^+) & \text{if } C_1 > S_1^+ \text{ and } C_2 > S_2^+ \\ (C_1 + S_1^+, C_2 + S_2^+) & \text{if } C_1 < -S_1^+ \text{ and } C_2 < -S_2^+ \\ (C_1 + S_1^-, C_2 - S_2^-) & \text{if } C_1 < -S_1^- \text{ and } C_2 > S_2^- \\ (C_1 - S_1^-, C_2 + S_2^-) & \text{if } C_1 > S_1^- \text{ and } C_2 < -S_2^- \end{cases}$$

where we denote $d = \|x_1\|_2^2\|x_2\|_2^2 - (\langle x_1, x_2\rangle)^2$, $C_1 = \left(\|x_2\|_2^2\langle y, x_1\rangle - \langle x_1, x_2\rangle\langle y, x_2\rangle\right)/d$,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$C_1 = \left(\|x_1\|_2^2 \langle y, x_2\rangle - \langle x_1, x_2\rangle \langle y, x_1\rangle\right)/d$ *and* $S_1^{\pm} = \left(\lambda_1 \|x_2\|_2^2 \mp \lambda_2 \langle x_1, x_2\rangle\right)/(2d)$, $S_2^{\pm} = \left(\lambda_2 \|x_1\|_2^2 \mp \lambda_1 \langle x_1, x_2\rangle\right)/(2d)$.

The proof of the lemma involves setting the subgradients of the objective to be zero and solving a two-dimensional linear system for the optimal (β_1^*, β_2^*). It could also be thought to set the derivatives to zero by considering one of the four combinations with β_1 and β_2 each either positive or negative. Our solution uses the exact regression coefficients $(C_1/d, C_2/d)$ for ordinary linear models in two dimensions, but with shrinkage factors from the ℓ_1 penalty. The shrinkage factors $S_1^{\pm}$ and $S_2^{\pm}$ (and analogously the soft-thresholding solution in one dimension) are subtracted from the regression coefficients. The divider d stays positive simply by Cauchy-Schwartz because of linear independence.

When $d = 0$ (or equivalently, that $x_1 = \theta x_2$ for some constant θ), the coefficient for the variable with smaller norm, say x_2 with $\|x_1\|_2 \geq \|x_2\|_2$, will have a negligible coefficient $\beta_2 = 0$. Moreover, fixing either β_1 or β_2 is zero, say $\beta_2 = 0$, the solution for β_1 is then the known soft-thresholded regression coefficient, that is

$$\beta_{1,thresh} = sgn\left(\langle y, x_1\rangle\right) \frac{\max\left(|\langle y, x_1\rangle| - \lambda_1/2, 0\right)}{\|x_1\|_2^2}$$

and $\beta_{2,thresh}$ follows a similar form if $\beta_1 = 0$. Thus we have now handled all special cases in a two-dimensional LASSO problem not covered by Lemma 3.3.1.

The objective $L(\beta_1, \beta_2)$ is convex and continuous except on the axes. Thus, the minimum is achieved at the point either with zero derivative values or on the axes with the smallest objective value. We have the following lemma that shows the solution explicitly for $L(\beta_1, \beta_2)$.

Lemma 3.3.2. *The solution set* (β_1^*, β_2^*) *for the two-dimensional LASSO equals* $(\tilde{\beta}_1^*, \tilde{\beta}_2^*)$ *when the conditions of Lemma reflem:twolasso are met, and otherwise it*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

equals one of the two candidate solutions $(\beta_{1,thresh}, 0)$ *and* $(0, \beta_{2,thresh})$ *that produces the smallest objective value* $L(\beta_1, \beta_2)$.

Now that the minimization in (3.3.1) can be reduced to a two-dimensional LASSO problem with an explicit solution. We can substitute the factors in the general solution with appropriate ones in our specific problem to obtain an explicit minimizer set. Moreover, the choice of setting $\bar{\alpha} = 0$ will not produce a better objective value when $k \geq 2$ since $f^{(1)} = x_{l_1}$ already minimizes the LASSO criterion with a single variable. For a fixed $l \in \{1, \ldots p\}$ and $k \geq 1$, we use the following notation for simplicity:

$$C^{(k)}_{\bar{\alpha},l} = \langle y, f^{(k-1)} \rangle - \langle f^{(k-1)}, x_l \rangle \langle y, x_l \rangle \quad C^{(k)}_{\gamma,l} = \left\| f^{(k-1)} \right\|_2^2 \langle y, x_l \rangle - \langle f^{(k-1)}, x_l \rangle \langle y, f^{(k-1)} \rangle$$
$$S^{(k)\pm}_{\bar{\alpha},l} = \lambda \left(v_f^{(k-1)} \mp \langle f^{(k-1)}, x_l \rangle \right) /2 \qquad S^{(k)\pm}_{\gamma,l} = \lambda \left(\left\| f^{(k-1)} \right\|_2^2 \mp v_f^{(k-1)} \langle f^{(k-1)}, x_l \rangle \right) /2,$$

and $d_l^{(k)} = \left\| f^{(k-1)} \right\|_2^2 - \langle f^{(k-1)}, x_l \rangle^2$. We also define the following two cases:

- case positive: $sgn(C^{(k)}_{\bar{\alpha},l}) = sgn(C^{(k)}_{\gamma,l})$, $\left| C^{(k)}_{\bar{\alpha},l} \right| > S^{(k)+}_{\bar{\alpha},l}$, and $\left| C^{(k)}_{\gamma,l} \right| > S^{(k)+}_{\gamma,l}$;
- case negative: $sgn(C^{(k)}_{\bar{\alpha},l}) = -sgn(C^{(k)}_{\gamma,l})$, $\left| C^{(k)}_{\bar{\alpha},l} \right| > S^{(k)-}_{\bar{\alpha},l}$, and $\left| C^{(k)}_{\gamma,l} \right| > S^{(k)-}_{\gamma,l}$.

Applying Lemma 3.3.1, we have the following nonzero solution at iteration k for a fixed l,

$$\begin{pmatrix} \bar{\alpha}_l^{(k)} \\ \gamma_l^{(k)} \end{pmatrix} = \begin{cases} \begin{pmatrix} sgn(C^{(k)}_{\bar{\alpha},l}) \left(\left| C^{(k)}_{\bar{\alpha},l} \right| - S^{(k)+}_{\bar{\alpha},l} \right) / d_l^{(k)} \\ sgn(C^{(k)}_{\gamma,l}) \left(\left| C^{(k)}_{\gamma,l} \right| - S^{(k)+}_{\gamma,l} \right) / d_l^{(k)} \end{pmatrix} & \text{if case positive} \\ \begin{pmatrix} sgn(C^{(k)}_{\bar{\alpha},l}) \left(\left| C^{(k)}_{\bar{\alpha},l} \right| - S^{(k)-}_{\bar{\alpha},l} \right) / d_l^{(k)} \\ sgn(C^{(k)}_{\gamma,l}) \left(\left| C^{(k)}_{\gamma,l} \right| - S^{(k)-}_{\gamma,l} \right) / d_l^{(k)} \end{pmatrix} & \text{if case negative,} \end{cases}$$

and also the soft thresholding ones

$$(\bar{\alpha}^{(k)}_{l,thresh}, 0) = sgn\left(\langle y, f^{(k-1)} \rangle \right) \frac{\max\left(\left| \langle y, f^{(k-1)} \rangle \right| - \lambda v_f^{(k-1)}/2, 0 \right)}{\left\| f^{(k-1)} \right\|_2^2}$$
$$(0, \gamma_{l,thresh}) = sgn\left(\langle y, x_l \rangle \right) \max\left(\left| \langle y, x_l \rangle \right| - \lambda/2, 0 \right).$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Naturally, we let $\alpha_l^{(k)} = 1 - \bar{\alpha}_l^{(k)}$ and likewise $\alpha_{l,thresh}^{(k)} = 1 - \bar{\alpha}_{l,thresh}^{(k)}$. The following proposition gives the closed form formula for our algorithms at each iteration.

Proposition 3.3.3. *The extended approximate LPGP or extended exact LPGP at iteration k solves for an objective of the form $L_f^{(k)}(\alpha, \gamma, l)$ for a fixed coordinate l, and the solution for $k \geq 2$ can be written explicitly as*

$$(\alpha_l^{(k)}, \gamma_l^{(k)}) = \arg \min_{(\alpha_l, \gamma_l) \in \{(\alpha_l^{(k)}, \gamma_l^{(k)}), (\alpha_{l,thresh}^{(k)}, 0)\}} L_f^{(k)}(\alpha, \gamma, l)$$

with

$$l^{(k)} = \arg \min_{l \in \{1, \ldots, p\}} L(\alpha_l^{(k)}, \gamma_l^{(k)}, l).$$

The solution for $k = 1$ is

$$(\gamma^{(1)}, l^{(1)}) = \arg \min_{l \in \{1, \ldots, p\}} L(0, \gamma_{l,thresh}, l)$$

with arbitrary $\alpha_l^{(1)}$.

This proposition is an application of Lemma 3.3.1 replacing the two variables with $f^{(k-1)}$ and x_l for a fixed l. Our iterative fit will never increase the objective value since the objective with the optimal (α, γ, l) will have an objective value not exceeding the one with $(0, 0, l)$, the objective value from the previous iteration. Therefore, we can drop the solution $(0, \gamma_{l,thresh})$ because it will reduce to the fit $f^{(1)}$, which cannot have smaller objective than $f^{(k)}$ when $k \geq 2$.

For the non-extended algorithms that restrict α to be within $[0, 1]$ for each fixed l, we can project the extended solution to the convex set $[0, 1] \times \mathbb{R}$ by convex projection. A suitable simple solution is to take the extended solution shown above if it is indeed within this convex set, and otherwise, to compute the solution along the closest boundary $\alpha = 0$ or $\alpha = 1$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.3.2 Update Rules

For our analysis, we need to compute various inner products and norms to produce the minimizing (α, γ). Among those, $\langle y, x_l \rangle$ for $l \in \{1, \ldots, p\}$ does not change with the iteration step k and we can store these products after they are precomputed. The others are $\langle y, f^{(k-1)} \rangle$, $\left\| f^{(k-1)} \right\|_2^2$, $v_f^{(k-1)}$ and $\langle f^{(k-1)}, x_l \rangle$ for every l. They all depend on $f^{(k-1)}$ and we need to update those quantities when we move to the next iteration.

A naive update rule would be to compute those things on the fly requiring $O(n)$ operations for a fixed l with total computations needed being $O(kpn)$ for k iterations, with one each picking l out of p coordinates. However, notice at iteration $k+1$ the computation of

$$\left\langle f^{(k)}, x_l \right\rangle = (1 - \alpha^{(k)}) \left\langle f^{(k-1)}, x_l \right\rangle + \gamma^{(k)} \left\langle x_{l^{(k)}}, x_l \right\rangle .$$

We can just compute $\langle x_{l^{(k)}}, x_l \rangle$ with the same $O(n)$ operations and then use the iterative updating formula to update $\left\langle f^{(k)}, x_l \right\rangle$ with $\left\langle f^{(k-1)}, x_l \right\rangle$ already computed in the previous iteration. We call this improved rule a covariance update, analogous to the proposal for coordinate descent in [33].

The naive update requires an $O(n)$ update that takes the form $f^{(k)} = (1 - \alpha^{(k)}) f^{(k-1)} + \gamma^{(k)} x_l$ for the n coordinates and another $O(n)$ operations to compute inner product with x_l for a fixed l; the total computation is $O(2n)$. On the other hand, the covariance update will only require $O(n)$ computations for $\langle x_{l^{(k)}}, x_l \rangle$. This can be improved even further by storing those inner products between x_i and x_j in all previous iterations. Thus, no $O(n)$ computation is needed if the required inner product has been already computed and stored. Similar arguments can be applied to other inner products and norms involving $f^{(k-1)}$. Most significantly, we may not

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

need $O(n)$ operations for each l at each iteration. Depending on the problem, the covariance update can have far fewer operations than the $O(kpn)$ of naive updates if we only hit a small set of coordinates all the way to convergence. We will use the covariance update in practice.

The coordinate descent method requires the same order of operations, $O(pn)$, for each iteration and a total of $O(Kpn)$ for K iterations. However, the number of iterations needed is unclear with coordinate descent, and K could be very large. Our algorithm has the same order $O(kpn)$ (or better with numerical improvement) with a specific control over the k iterations needed with the computation accuracy theorem. Small prices (with a constant order) we pay are to optimize for the α, and to track a few inner products and norms, which may not need as much extra computation as in the covariance update rule.

3.3.3 Compressed Storage

To store the computed covariances in the covariance update rule, we use a compressed row storage for simplicity and effectiveness. We have a vector storing the inner products between x_i and x_j and a vector storing the starting index of $\langle x_i, x_j \rangle$ for varying j and fixed i if i^{th} coordinate is picked by our algorithms at a certain step. Notice that in our covariance updating rule, only the inner product $\langle x_{l^{(k)}}, x_l \rangle$ is needed for a fixed $l^{(k)}$ at iteration k, and we would not expect the set of coordinates picked by all previous iterations to be large because we only pick the best coordinate at each iteration. In the situation that $p \gg n$ for example, we will only need our storage to be slightly bigger than $n \times p$ in the worst case with small λ. Another option for storage is to use an existing sparse matrix package, but we think our implementation is very effective in handling these particular cases where the column index always loops through all the coordinates.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.3.4 Path Sweep

Our algorithms compute the coefficient for a single penalty level λ in the LASSO problem. We can extend the algorithms to compute for multiple λ along a grid. A considerable speedup is achieved by solving the multiple lasso problems sequentially from the largest λ to the smallest, and then using the computed coefficients from a slightly larger λ as the initializing $\beta^{(0)}$ for the next computation. This is called warm start in [33], and a more general treatment is in [41].

A large λ tends to have smaller coefficients as a solution compared to a small λ, but our algorithms, especially the extended algorithms, have opportunity in the first step to boost up the all the initializing coefficients by multiplying all coordinates with a α fraction to be bigger than 1 if it is numerically best. The coordinate descent algorithms, on the other hand, have to update each coordinate one at a time.

3.3.5 Hybrid

Computation to a high precision, though not really needed for statistical computation as discussed earlier with respect to adaptive ℓ_1 penalized least squares, can be very expensive. Once we use our LPGP algorithms to identify the right set of coefficients, we can allow a coordinate descent in the inner loop to refine a higher-precision solution. Our algorithms accommodate coordinate descent very easily, which corresponds to setting $\alpha = 0$ and looping through coordinates instead of picking the best coordinate. This hybrid does improve the computation speed in practical settings, as will be shown in the next section.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.4 Comparisons

We compare the numerical performance of our algorithms with other leading algorithms. Our main numerical work is coded in C and linked to the R interface. We compare with the R package "glmnet", implementing coordinate descent, and the least angle implementation of package "lars". The same data set is used across all algorithms and we use the same grid values of λ and convergence precision for "glmnet" and ours. "lars" solves for the whole solution path for varying λ, and "glm" mimics the path by obtaining the coefficients for a great number of λ values in a grid. Ours are like "glm" to solve for the same grid values of λ that it uses. We will report only a small number of variants of our algorithms that have particularly good computation speed.

The computation does not need to obtain high-precision solutions as because of the computational accuracy and statistical risk tradeoff in adaptive ℓ_1 penalized least squares discussed earlier. For a fair comparison, we run our algorithms long enough to obtain smaller objective values than produced by "glm" and "lars".

3.4.1 Simulation

We generate Gaussian data for n observations and p predictors. Each pair of predictors X_j and $X_{j'}$ has a correlation of ρ. The response is generated by

$$Y = \sum_j \beta_j X_j + \eta Z$$

where $\beta_j = (-1)^j exp(-2(j-1)/20)$, $Z \sim N(0,1)$ and η is chosen so that the signal to noise ratio is 3. We compare the average running time of ours with others, varying n, p and ρ. As shown in the attached table, our algorithms are faster in almost all

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

situations. "glm" has a slight advantage when the design is near orthogonal but ours have comparable performance in that situation as well. The hybrid variant has an advantage when $p > n$ but may suffer from high correlations as does the coordinate descent method in the same situation.

3.5 Summary

We provide fast computational algorithms for the ℓ_1 penalized least squares problem. We consider numerous modifications of the original LPGP algorithm to achieve better computation compared with other competitors. The same computation guarantee of LPGP applies to our algorithms, and our new proposals enjoy significant improvement in computation speed. It is still an open question as to whether improved theoretical results could also be established for our proposals.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

	$n = 1000, p = 100$			
ρ	0	0.1	0.5	0.9
LPGP	0.027	**0.024**	**0.032**	**0.113**
LPGP-cord	0.028	0.049	0.635	8.058
lars	0.577	0.583	0.488	0.531
glm.na	0.103	0.352	6.083	62.259
glm	**0.023**	0.043	0.508	4.997

	$n = 5000, p = 100$			
ρ	0	0.1	0.5	0.9
LPGP	0.079	0.081	**0.115**	**0.235**
LPGP-cord	0.111	0.137	0.831	6.67
lars	2.331	2.236	2.551	2.483
glm.na	0.389	1.786	36.992	320.014
glm	**0.053**	**0.072**	0.608	4.753

	$n = 100, p = 1000$			
ρ	0	0.1	0.5	0.9
LPGP	**0.023**	**0.028**	0.54	**0.023**
LPGP-cord	0.029	0.037	**0.3**	0.075
lars	0.803	0.715	0.684	0.668
glm.na	0.115	0.131	0.433	0.416
glm	0.131	0.131	0.391	0.232

	$n = 100, p = 5000$			
ρ	0	0.1	0.5	0.9
LPGP	0.604	0.196	**0.195**	0.849
LPGP-cord	**0.285**	**0.147**	0.324	**0.571**
lars	2.854	2.72	2.893	2.899
glm.na	0.316	0.265	0.74	1.767
glm	0.564	0.443	1.251	2.057

Table 3.1: Average running time in seconds. LPGP-cord is an hybrid of our LPGP and coordinate descent; glm.na is also an variant of glm.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.6 Appendix: Another Variant and Proof

A variant of LPGP is to pick

$$l^{(k)} = \arg\max \left\langle x_l, y - f^{(k-1)} \right\rangle$$

instead of minimizing the original criterion along with α and γ. Then, we optimize over α and γ to minimize

$$L^{(k)}(\alpha, \gamma) = \left\| y - \alpha f^{(k-1)} - \gamma x_{l^{(k)}} \right\|_2^2 + \lambda \left[\alpha v_f^{(k-1)} + |\gamma| \right].$$

We will show a slightly better constant in computational accuracy than the ones in [42] in some situations.

Theorem 3.6.1. *The LPGP algorithm variant has a similar computational accuracy bound as the other LPGP algorithms, that is*

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j^{(k)} \right)^2 + \lambda v^{(k)} \leq \inf_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda V_\beta + \frac{4 b_f}{k+1} \right\}$$

where $b_f = 3V_f^2 + 2V_f \|y - f\|_2 + \|y\|_2^2 - \|y - f\|_2^2$.

Proof. Denote the difference in the objective value between the k^{th} fit $f^{(k)}$ and a arbitrary fixed reference f by

$$e^{(k)} = \left\| y - f^{(k)} \right\|_2^2 - \|y - f\|_2^2 + \lambda \left[v_f^{(k)} - V_f \right].$$

Substitute $f^{(k)} = \alpha^{(k)} f^{(k-1)} + \gamma^{(k)} x_{l^{(k)}}$ and the bound $v_f^{(k)} \leq \alpha^{(k)} v_f^{(k-1)} + \left| \gamma^{(k)} \right|$, then e_k is not smaller when its minimizer $(\alpha^{(k)}, \gamma^{(k)})$ is replaced by a customary choice $(\alpha_k, \alpha_k V_f)$ for a $l^{(k)}$ picked first and with a fixed α_k to be specified later. Rearrange

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

terms to show that

$$
\begin{aligned}
e^{(k)} \leq\ & \alpha_k e^{(k-1)} + (1-\alpha_k)^2 V_f^2 && (3.6.1)\\
& -\alpha_k(1-\alpha_k)\left\|f^{(k-1)}\right\|_2^2 - (1-\alpha_k)\left\|f\right\|_2^2 + 2(1-\alpha_k)\left\langle f, y\right\rangle && (3.6.2)\\
& +2\alpha_k(1-\alpha_k)V_f\left\langle x_{l^{(k)}}, f^{(k-1)} - y\right\rangle - 2(1-\alpha_k)^2 V_f\left\langle x_{l^{(k)}}, y\right\rangle. && (3.6.3)
\end{aligned}
$$

The absolute value of the inner product in the last term is $|\langle x_{l^{(k)}}, y\rangle| \leq |\langle x_{l^{(k)}}, y - f\rangle| + |\langle x_{l^{(k)}}, f\rangle| \leq \|y - f\|_2 + V_f$ (by the standardization), and the last term can be bounded by a function not of $l^{(k)}$. The next to the last term is minimized with our choice of $l^{(k)}$ and thus is bounded by the average of all possible coordinates. In particular, for $f = \sum_l \beta_l x_l$ we choose the probability of picking a particular l to be β_l / V_f, and then the average over random x_l is f/V_f. We then write the bound in the square form that

$$
\begin{aligned}
e^{(k)} \leq\ & \alpha_k e^{(k-1)} + (1-\alpha_k)^2\left[3V_f^2 + 2V_f\left\|y - f\right\|_2 + \left\|y\right\|_2^2 - \left\|y - f\right\|_2^2\right] && (3.6.4)\\
& -\alpha_k(1-\alpha_k)\left\|f^{(k-1)} - f\right\|_2^2. && (3.6.5)
\end{aligned}
$$

We choose $\alpha_k = (k-1)/(k+1)$ and then the last negative squares can be disregarded for the purpose of upper bounding. The iterative bound is therefore

$$
e^{(k)} \leq \frac{k-1}{k+1} e^{(k-1)} + \frac{4}{(k+1)^2} b_f
$$

where we write $b_f = 3V_f^2 + 2V_f\left\|y - f\right\|_2 + \left\|y\right\|_2^2 - \left\|y - f\right\|_2^2$. Likewise we check that $e^{(1)} \leq b_f$. Assume $e^{(k-1)} \leq 4b_f/k$ and by an induction argument we show that

$$
e^{(k)} \leq \frac{4b_f}{k+1}.
$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

□

The sampling argument over the choice of x_l is due to [45] and is used in [42] to show similar bounds with different b_f, which could be larger for constant multipliers than our b_f proved here in some situations where $y \neq f$. Consider a trivial example in regression with $\lambda = 0$ where $x_1 = (\sqrt{2}, 0)^T$, $y = (1, 1)^T$ and the fitted $f = (1, 0)^T = x_1/\sqrt{2}$. The two b_f given by them is either $[V + \|y\|]^2$ or $[2V + \|y - f\|]^2$, and equals $3 + \sqrt{5}$ and 8 correspondingly in this explicit example, as compared to ours with a smaller numerical value 4.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 4

Gaussian Graphical Models

4.1 Introduction

Large covariance matrix estimation is an important topic, particularly when the sample size n is much smaller than the number of variables p. There are many reasons for its importance, both in theory and applications. Principle examples include principle component analysis, linear discriminant analysis, and graphical models. There is also a wide range of applications, for example genetic association, brain imaging, climate data and many others.

The usual sample covariance matrix is an unstable estimator for the population covariance matrix, see [44] for a review. There is an upsurge in the literature on the improvements of such problems with different focuses. One may be interested in the asymptotic behavior in various matrix norms under special models, for example [29, 13] and references within, and also a recent minimax result in [19]. A different goal that we share is to identify the sparse entries of the inverse covariance matrix. The latter is especially important for constructing graphical models, representing the dependence structure of multiple variables.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The ℓ_1 penalty approach has been widely adopted for sparse graphical models, see [61, 43, 52, 64, 1, 35, 49], and one popular model used in these results is multivariate Gaussian. We will generalize our log-density estimation result to study this problem. Computational results will be reported here. The corresponding risk bound may be obtained by extending the computational result as we did in chapter 1, and is still a topic for further research efforts.

4.2 Model

Consider the data $X_1, X_2, \ldots, X_n$ i.i.d. from zero mean multivariate Gaussian data with an unknown covariance matrix $\Sigma_{p\times p}$, and we write the data matrix as $X = (X_1, X_2, \ldots, X_n)^T$ of size $n \times p$. The twice per-sample ℓ_1 penalized negative loglikelihood criterion seeks a matrix estimator M for Σ^{-1} to minimize

$$L(M) = tr\,[MS] - \log\det(M) + \rho\,\|M\|_1 \tag{4.2.1}$$

where $\|M\|_1$ is the sum of absolute values of all entries, $S = x^T x/n$ is the sample covariance matrix and $\rho \geq 0$ is the penalization parameter.

A slightly more general ℓ_1 penalty replaces the scalar parameter ρ with various levels such that the entry $M(i,j)$ is associated with the corresponding penalty parameter ρ_{ij}. This permits, for example, penalizing only the off-diagonal entries by assigning $\rho(i,i) = 0$ for $1 \leq i \leq p$, as one may prefer non-penalized diagonal entries. For simplicity, we will stick to the setting with a single parameter ρ. The result for general ρ is analogous.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4.3 Algorithm

One initializes with a diagonal matrix M_0 with a sensible choice of diagonal entries, for example one possible choice is $M_0(i,i) = 1/(S(i,i) + \rho)$ for $1 \leq i \leq p$. At the $(k+1)^{th}$ iteration given M_k at the previous iteration, we select the entry $(i,j) \in \{1,\ldots,p\}^2$, $\alpha \in [0,1]$ and $\beta \in \mathbb{R}$ to minimize

$$tr\left[\left(\left(1-\alpha\right)M_k + \beta\delta_{ij}\right)S\right] - \log\det\left[\left(1-\alpha\right)M_k + \beta\delta_{ij}\right] + \rho\left[\left(1-\alpha\right)v_k + |\beta|\right] \quad (4.3.1)$$

where $v_k = \|M_k\|_1$ and δ_{ij} is a zero matrix except that it takes the value $1/2$ at the entries (i,j) and (j,i) if $i \neq j$; if $i = j$ the matrix δ_{ij} is non-zero only at the diagonal entry (where it take the value 1). The $(k+1)^{th}$ fit is taken to be $M_{k+1} = (1-\alpha_k)M_k + \beta_k\delta_{i_kj_k}$ where (i_k,j_k), α_k and β_k are optimal. We repeat until achieving the desired accuracy. We call this algorithm greedy likelihood pursuit (GLP).

It is easy to see that the optimal update $(1-\alpha)M_k + \beta\delta_{ij}$ from the GLP algorithm will have the objective value not bigger than the one M_k would give. The GLP algorithm, as a down-hill strategy, for the convex objective $L(M)$ will eventually converge to the global minimum for any given data when the number of iterations is big enough. More importantly, the computational accuracy bound controlling the number of iterations needed can be established as discussed in the next section.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4.4 Main Result

We denote the level set of matrices with smaller objective value than a positive symmetric M_0 by

$$\Omega(M_0) = \{\text{positive, symmetric } M : L(M) \leq L(M_0)\}.$$

In order to establish a computation theorem, we will study the eigenvalue characteristics of these matrices. Indeed, by Lemma 4.5.2, there exists an eigenvalue bound depending on the initializing M_0 for all $M \in \Omega(M_0)$, that is

$$\lambda(M) \geq e^{-(L(M_0)-p)-1} \det(S_\rho)/(\lambda_m + \rho)$$

where λ_m is the maximum eigenvalue of S and $S_\rho = S + \rho I$ with I being the identity matrix. For simplicity, we denote the quantity in the lower bound by ϵ. For a special choice of $M_0 = S_\rho^{-1}$, it simplifies to $\epsilon = exp\{-\rho\|S_\rho^{-1}\|_{1,off} - 1\}/(\lambda_m + \rho)$, where the off diagonal ℓ_1 norm $\|S\|_{1,off} = \sum_{i \neq j} |S_{ij}|$. All the matrices produced by our GLP algorithm will be within this set $\Omega(M_0)$, as well as any target M we wish to compute accurately by the criterion.

The proof is an extension of the log-density result to the case of unbounded functions (but with moment control) in $\mathcal{H}$. A difference is that we will consider the schedule of α to be a fixed small number in the proof instead of varying α as before, and the result holds for some large m iterations where m is roughly $1/\alpha$. A modification allowing decaying α is also briefly discussed near the end.

Theorem 4.4.1. *The matrix M_m produced by the greedy likelihood pursuit after m iterations achieves a criterion value compared to what is achieved at any target M*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

in $\Omega(M_0)$, *satisfying*

$$tr\left[M_m S\right] - \log\det\left(M_m\right) + \rho v_m \leq tr\left[MS\right] - \log\det\left(M\right) + \rho\|M\|_1 + A_m(M)$$

where

$$A_m(M) = \frac{\log m}{m} B + \frac{1}{m}\max(e_0 - \frac{\log m}{m} B, 0)$$

for $B = [3 + 16\xi(2\alpha V/\epsilon)]\left(\frac{V}{\epsilon}\right)^2$, *provided that* m *is large enough that* $\alpha = (\log m)/m \leq \min(\epsilon/(2V), 1)$ *for* $V = \|M\|_1$. *The function* $\xi(\cdot)$ *and* e_0 *are as given in the proof. Here* ϵ *is the lower bound on the eigenvalues of matrices in* $\Omega(M_0)$.

Remark 4.4.2. *In particular if* M^{opt} *is the choice exactly minimizing the criterion*

$$tr\left[MS\right] - \log\det(M) + \rho\|M\|_1,$$

then applying the bound at $M = M^{opt}$ *shows that after* m *steps our solution* M_m *is within order* $(\log m)/m$ *of the minimum.*

Remark 4.4.3. *The bound could be improved by using a smaller set* $\Omega(M_k)$ *after some initial burn-in* k *iterations, and the constants in the bound that is a function of* $1/\epsilon$ *will be smaller. The bound will be also based on additional steps* $m - k$ *needed.*

Proof. Denote the difference in the objective value of our estimate matrix M_{k+1} and a reference matrix M by e_{k+1}, and we write explicitly the determinants in the integral form, that is

$$\begin{aligned} e_{k+1} = & tr\left[\left(\left(1-\alpha_k\right)M_k + \beta_k \delta_{i_k j_k}\right) S\right] - tr\left[MS\right] \\ & + 2\log\frac{\int e^{-\frac{1}{2}x^T\left[\left(1-\alpha_k\right)M_k + \beta_k\delta_{i_k j_k}\right]x}}{\int e^{-\frac{1}{2}x^T M x}} + \rho\left[\left(1-\alpha_k\right)v_k + \left|\beta_k\right| - V\right]. \end{aligned}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Since α_k, β_k and (i_k, j_k) are optimal, then e_{k+1} is not smaller when replacing with some fixed $\alpha < \min(\epsilon/(2V), 1)$, $\beta = \alpha V$ for all (i,j) pairs. As with the bounded dictionary case, add and subtract $2\log\int exp\left\{-\frac{1}{2}x^T\left[(1-\alpha)M_k + \alpha M\right]x\right\}$ to relate to the log normalizing constants for e_k, and by Jensen's the following inequality holds

$$\log\frac{\int e^{-\frac{1}{2}x^T[(1-\alpha)M_k+\alpha M]x}}{\int e^{-\frac{1}{2}x^T Mx}} \;\leq\; (1-\alpha)\log\frac{\int e^{-\frac{1}{2}x^T M_k x}}{\int e^{-\frac{1}{2}x^T Mx}}$$

where the log ratio in the bound is the difference in normalizing constants in e_k. The trace operator is linear, so the α and β terms separate. Substitute the bound above and rearrange to show

$$e_{k+1} \;\leq\; (1-\alpha)\,e_k + \alpha tr\left[(V\delta_{ij} - M)\,S\right] + 2\log\int p_\alpha(x)e^{-\alpha\frac{1}{2}x^T[V\delta_{ij}-M]x} \quad (4.4.1)$$

where $p_\alpha(x)$ is a multivariate Gaussian proportional to $exp\left\{-x^T\left[(1-\alpha)M_k + \alpha M\right]x\right\}$. By the decreasing property of $L(M_k)$, M_k is within $\Omega(M_0)$ for $k = 0, 1, \ldots$, and by assumption so is the target matrix M. Consequently, the inverse covariance matrix (precision matrix) $(1-\alpha)M_k + \alpha M$ of p_α is also in $\Omega(M_0)$.

This is further upper bounded by replacing the minimizing (i,j) pair with random draws of (i,j). In particular, we consider picking the pair (i,j) with probability $w_{ij} = |M_{ij}|/V$, then the expectation of $V\delta_{ij}$ for random (i,j) is M. The trace term vanishes in the average, and it remains to bound the average logarithmic term.

Define a cumulant generating function averaged over (i,j) as

$$g(\gamma) = \sum_{i,j} w_{ij}\log\int p_\alpha(x)e^{-\gamma\frac{1}{2}x^T[V\delta_{ij}-M]x},$$

the remaining term to bound is twice $g(\alpha)$. We bound $g(\alpha)$ by a Taylor expansion around 0 of the second order. It is obvious that $g(0) = 0$ since p_α is a probability

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

distribution, and that the first order term is zero on average because it is linear in $V\delta_{ij} - M$. The non-vanishing term in the Taylor series for $g(\alpha)$ is of the second order $O(\alpha^2)$ and its coefficient is half a second derivative. By tossing nonpositive terms,

$$g''(\nu) \leq \sum_{i,j} w_{ij} \frac{\int \left[\frac{1}{2} x^T (V\delta_{ij} - M)x\right]^2 e^{-\nu \frac{1}{2} x^T (V\delta_{ij} - M)x} p_\alpha(x)}{\int e^{-\nu \frac{1}{2} x^T (V\delta_{ij} - M)x} p_\alpha(x)}.$$

We seek a bound for it that holds for all $0 \leq \nu \leq \alpha$. We seek upper and lower bounds of the numerator and denominator, respectively, and their ratio then is the bound we obtain.

We consider first obtaining a positive lower bounding the denominator uniformly for (i,j). The exponential $exp\left\{-\nu x^T (V\delta_{ij} - M)x/2\right\}$ in the denominator is not smaller than the exponential of its minus absolute value with ν replaced by a larger α for all (i,j). We can further lower bound the exponential by its Taylor expansion truncated to the first order, that is for all (i,j),

$$\begin{aligned}
\int e^{-\nu \frac{1}{2} x^T (V\delta_{ij} - M)x} p_\alpha(x) &\geq \int e^{-\alpha \frac{1}{2} \left|x^T (V\delta_{ij} - M)x\right|} p_\alpha(x) \\
&\geq \int \left[1 - \frac{\alpha}{2} \left|x^T (V\delta_{ij} - M)x\right|\right] p_\alpha(x) \\
&\geq \int \left[1 - \frac{\alpha}{2} \left|x^T V\delta_{ij} x\right| - \frac{\alpha}{2} \left|x^T M x\right|\right] p_\alpha(x) \\
&\geq 1 - \frac{\alpha V}{2} \int \left[\left|x^T \delta_{ij} x\right| + \sum_{i'j'} w_{i'j'} \left|x^T \delta_{i'j'} x\right|\right] p_\alpha(x)
\end{aligned}$$

where the last inequality is due to the fact that the expectation of the absolute value is not smaller than the absolute value of the expectation, and random (i', j') is an independent copy of (i,j). By Cauchy-Schwartz, each term as the product of x_i and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

x_j for every (i,j) is bounded by

$$\int \left|x^T \delta_{ij} x\right| p_\alpha(x) \leq \left[\int x_i^2 p_\alpha(x_i)\right]^{\frac{1}{2}} \left[\int x_j^2 p_\alpha(x_j)\right]^{\frac{1}{2}}.$$

Suppose there is a bound σ_{max}^2 on the maximum of the variances of Gaussians with parameters in $\Omega(M_0)$, then the denominator is lower bounded by $1 - \alpha V \sigma_{max}^2 \geq 1/2$ for $\alpha \leq 1/(2V\sigma_{max}^2)$. One such bound on the variances of $p_\alpha(x)$ arise from the eigenvalue bound $1/\epsilon$ on the covariance matrix, which give rise to the more stringent condition $\alpha \leq \epsilon/(2V)$ for the lower bound of $1/2$ to hold for the denominator. See remark 4.4.4 for discussion of possible improvements.

The bound on the numerator is as follows. First, fix x and consider the pointwise bound of the exponential replacing ν with α

$$e^{-\nu\frac{1}{2}x^T(V\delta_{ij}-M)x} \leq e^{-\alpha\frac{1}{2}x^T(V\delta_{ij}-M)x} + e^{\alpha\frac{1}{2}x^T(V\delta_{ij}-M)x},$$

we then expand the point wise upper bound above using an infinite series representation of the exponential. Obviously the odd order terms cancel out. Consequently, the integrand in the numerator is bounded by

$$\begin{aligned}\left[\frac{1}{2}x^T(V\delta_{ij}-M)x\right]^2 e^{-\nu\frac{1}{2}x^T(V\delta_{ij}-M)x} &\leq \frac{1}{2}\left[x^T(V\delta_{ij}-M)x\right]^2 \\ &+ \frac{V^2}{2}\sum_{l=1}^{\infty}\frac{(\alpha V)^{2l}}{2^{2l}(2l)!}\left[x^T\left(\delta_{ij}-\frac{M}{V}\right)x\right]^{2l+2}.\end{aligned}$$

We single out the $l=0$ term so that, averaging over w_{ij}, it will give some savings on the leading constant, and the remaining series can be bounded uniformly in (i,j).

The first term (integrating over p_α and averaging over w_{ij}), changing the order

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

of integration and averaging, is bounded by

$$\frac{1}{2}V^2 \int p_\alpha(x) \sum w_{ij} \left[x^T(\delta_{ij} - \frac{M}{V})x\right]^2 \leq \frac{1}{2}V^2 \int p_\alpha(x) \sum_{ij} w_{ij} \left[x^T \delta_{ij} x\right]^2 \leq \frac{3}{2}\left(V\sigma_{max}^2\right)^2.$$

The first inequality uses the fact that the average of δ_{ij} gives a mean M/V. Changing the order of integration and averaging again, the last inequality uses Cauchy-Schwartz with measure p_α and the fact that the fourth moment of a mean zero univariate Gaussian equals 3 times the variance squared.

Similarly, the integral of the polynomial $\left[x^T\left(\delta_{ij} - \frac{M}{V}\right)x\right]^{2l+2}$ with respect to $p_\alpha(x)$ is bounded uniformly for all (i,j) as

$$\int p_\alpha(x) \left[x^T\left(\delta_{ij} - \frac{M}{V}\right)x\right]^{2l+2} \leq 2^{2l+2} \max_{i,j} \int p_\alpha(x) \left[x^T \delta_{ij} x\right]^{2l+2} \leq 2^{2l+2}(4l+3)!! \left(\sigma_{max}^2\right)^{2l+2} \tag{4.4.2}$$

where the constant $(4l+3)!! = (4l+3)(4l+1)\cdots 3\cdot 1$ appears because of the multiplier in the $2(2l+2)^{th}$ moment of univariate Gaussian. The constant $(4l+3)!!$ is bounded by $2^{2l+2}(2l+2)! = (4l+4)!!$ as easily seen with term-by-term comparison. The infinite summation starting from $l = 1$ after integrating out p_α and w_{ij} is therefore bounded by

$$8\left(V\sigma_{max}^2\right)^2 \sum_{l=1}^{\infty} (2\alpha V \sigma_{max}^2)^{2l}(2l+2)(2l+1).$$

The infinite summation is finite if $\alpha < 1/(2V\sigma_{max}^2)$, and indeed is satisfied by our requirement on α. We denote the infinite summation by $\xi(2\alpha V/\epsilon)$, which can be written in the closed form, checked with the help of Mathematica,

$$\xi(\tau) = \frac{2(6\tau^2 - 2\tau^4 + \tau^6)}{(1-\tau^2)^3}. \tag{4.4.3}$$

Finally, merging all the bounds together, we have our final upper bound at each

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

iteration k to be

$$\begin{aligned} e_{k+1} &\leq (1-\alpha)e_k + \alpha^2 \frac{\frac{3}{2} + 8\xi(2\alpha V\sigma_{max}^2)}{1-\alpha V\sigma_{max}^2} \left(V\sigma_{max}^2\right)^2 \\ &\leq (1-\alpha)e_k + \alpha^2 \left[3 + 16\xi(2\alpha V\sigma_{max}^2)\right] \left(V\sigma_{max}^2\right)^2 \end{aligned}$$

where the last inequality is because $\alpha V\sigma_{max}^2 \leq (1/2)$. By Lemma 4.5.4 on the iterative formula, and choosing m large so that $\alpha = \log(m)/m < \min(\epsilon/(2V\sigma_{max}^2), 1)$, we have that the m step difference is bounded by

$$e_m \leq \frac{\log m}{m} B + \frac{1}{m} \max(e_0 - \frac{\log m}{m} B, 0).$$

□

Remark 4.4.4. *It would be natural to impose a constraint directly on the maximum variances $\int x_i^2 p_M(x) \leq \sigma_{max}^2$ and thereby avoid reference to the eigenvalue bounds. We would need that if this property holds for Gaussians with inverse covariance matrix M in $\Omega(M_0)$, then it also holds when the inverse covariance is $(1-\alpha)M + \alpha\delta_{ij}$ for small α.*

Remark 4.4.5. *In the proof, in order to have a decaying schedule of α like $2/(k+1)$ and a resulting bound analogous to the results in chapter 1, one may consider a big constant K such that the choice $\alpha_k = 2/(k+K)$ at step k is not bigger than $\min(\epsilon/(2V), 1)$ the maximum allowed α in the proof above. An analogous induction step will follow, though with different constants.*

Corollary 4.4.6. *The following computation bound holds for GLP on Graphical models under the same condition*

$$tr[M_k S] - \log\det(M_k) + \rho v_k \leq tr[MS] - \log\det(M) + \rho V + \frac{\tilde{B}}{K+k}.$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where $\tilde{B} \geq \max(Ke_0/4, B')$, $K = \max(\lceil 8V/\epsilon \rceil, 2)$, *and* $B' = 115(V/\epsilon)^2$.

Proof. Consider the difference e_k for $k \geq 1$, we choose $\alpha = \alpha_k = 2/(k+K) \leq 2/(K+1)$ and for the chosen K, $2\alpha V/\epsilon \leq 1/2$ satisfies the requirement for all k. We use the crude constant 7 to bound $\xi(2\alpha V/\epsilon) = \xi(\tau)$ for $0 \leq \tau \leq 1/2$, and then $B' \geq B$ for all iterations. By the same argument, we show

$$e_k \leq (1-\alpha)e_{k-1} + \alpha^2 B'.$$

We first check the $k = 1$ inequality. Indeed,

$$\begin{aligned} e_1 &\leq (1 - \frac{2}{K+1})e_0 + \frac{4B'}{(K+1)^2} \\ &\leq (1 - \frac{2}{K+1})\frac{4\tilde{B}}{K} + \frac{4\tilde{B}}{(K+1)^2} \\ &\leq \frac{4\tilde{B}}{K+1}. \end{aligned}$$

The induction step for $k \geq 1$ is similar to the log-density case. □

4.5 Appendix

Lemma 4.5.1. *If the eigenvalues of real symmetric matrix A and B of the same size are bounded between $[a_l, a_u]$ and $[b_l, b_u]$ respectively, and A and B are positive semidefinite, i.e. $a_l \geq 0$ and $b_l \geq 0$, the eigenvalues of the product AB (and equivalently $B^{\frac{1}{2}}AB^{\frac{1}{2}}$) is then bounded between $[a_l b_l, a_u b_u]$.*

Proof. Matrix B is obviously positive semidefinite and can be decomposed as $B = B^{\frac{1}{2}}B^{\frac{1}{2}}$ where $B^{\frac{1}{2}}$ is the square root matrix of B. It is clear that matrix AB and $B^{\frac{1}{2}}AB^{\frac{1}{2}}$ have the same eigenvalues, therefore it is equivalent to bound the eigenvalues

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

of $B^{\frac{1}{2}}AB^{\frac{1}{2}}$. The lower eigenvalue bound for this matrix is shown here and the upper bound argument is analogous. For any real symmetric matrix C of size p, the eigenvalues of C is bounded below by c if and only if

$$u^T C u \geq c u^T u$$

for all vector $u \in \mathbb{R}^p$. Using this fact twice on A and then on B, we can show the following argument for all $u \in \mathbb{R}^p$ that

$$\left(u^T B^{\frac{1}{2}}\right) A \left(B^{\frac{1}{2}} u\right) \geq a_l u^T B^{\frac{1}{2}} B^{\frac{1}{2}} u = a_l u^T B u \geq a_l b_l.$$

□

Lemma 4.5.2. *For the objective $L(M)$ defined in (4.2.1), the following eigenvalue lower bound holds for all matrices M that has $L(M) \leq L(M_0)$ for a fixed arbitrary positive definite matrix M_0,*

$$\lambda(M) \geq e^{-(L(M_0)-p)-1} \det(S_\rho)/(\lambda_m + \rho)$$

where λ_m is the maximum eigenvalue of S in the objective L and $S_\rho = S + \rho I$ with I being the identity matrix.

Proof. Denote the symmetrically scaled matrix of M by $\tilde{M} = S_\rho^{\frac{1}{2}} M S_\rho^{\frac{1}{2}}$. Clearly the eigenvalues $0 \leq \rho \leq \lambda(S_\rho) \leq \lambda_m + \rho$ and $\lambda(S_\rho^{-1}) \geq (\lambda_m + \rho)^{-1}$. The objective $L(M)$ equals the following modified objective

$$\tilde{L}(\tilde{M}) = tr\left[\tilde{M}\right] - \log\det(\tilde{M}) + \log\det(S_\rho) + \rho \|S_\rho^{-\frac{1}{2}} \tilde{M} S_\rho^{-\frac{1}{2}}\|_{1,off}$$

where the off diagonal ℓ_1 norm $\|M\|_{1,off} = \sum_{i \neq j} |M_{ij}|$. Rearrange terms in the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

inequality $\tilde{L}(\tilde{M}) = L(M) \leq L(M_0)$, and upper bound by tossing the off diagonal norm term on the left hand side, then

$$tr\left[\tilde{M}\right] - \log\det(\tilde{M}) \leq L(M_0) - \log\det(S_\rho).$$

Write the eigenvalues of $\tilde{M}$ by $\tilde{\lambda}_j$ for $1 \leq j \leq p$, and replace the trace and determinant with their eigenvalue representation to show that

$$\sum_{j=1}^{p}[\tilde{\lambda}_j - \log\det\tilde{\lambda}_j] \leq L(M_0) - \log\det(S_\rho).$$

Equivalently, it is handy to subtract p from both sides, that is

$$\sum_{j=1}^{p}[\tilde{\lambda}_j - 1 - \log\det\tilde{\lambda}_j] \leq L(M_0) - p - \log\det(S_\rho).$$

Each positive summand of the left hand side is then bounded above by the right hand side. Observe the fact that the function $x - 1 - \log x$ for positive x has the left end growth controlled by $\log x$ for $x \leq 1$, then it is easy to see that $x - 1 - \log x \leq b$ implies $x \geq e^{-b-1}$. Using this lower bound control on each summand, we can show for all j that

$$\tilde{\lambda}_j \geq e^{-(L(M_0)-p)-1}\det(S_\rho).$$

By Lemma 4.5.1, the eigenvalues of $M = S_\rho^{-\frac{1}{2}}\tilde{M}S_\rho^{-\frac{1}{2}}$ is not smaller than

$$e^{-(L(M_0)-p)-1}\det(S_\rho)/(\lambda_m + \rho).$$

□

Lemma 4.5.3. *If $X \in \mathbb{R}^p$ follows a multivariate Gaussian distribution $N(0, \Sigma)$,*

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and the covariance matrix Σ is ϵ-well conditioned (that is, the eigenvalues of Σ are bounded above by $1/\epsilon$), then the maximum marginal variance among all coordinates is bounded by

$$\max_{1\le i\le p} EX_i^2 \le \frac{1}{\epsilon} \tag{4.5.1}$$

where the expectation is taken with respect to the Gaussian distribution aforementioned.

Proof. The diagonal entry of Σ is the marginal variance of X under the Gaussian distribution. Thus we need only to bound the diagonal elements, for which we use the bound on eigenvalues of a real symmetric matrix Σ that

$$\max_{u^Tu=1} u^T\Sigma u \le \frac{1}{\epsilon} \tag{4.5.2}$$

for $u \in \mathbb{R}^p$. The diagonal entries are not bigger than the same bound by considering particular choices of u such that u_i is a zero vector except 1 at the i^{th} coordinate for all $1 \le i \le p$. □

Lemma 4.5.4. *If A_0, A_1, ..., A_K and follow a iterative formula $A_k = (1-\theta)A_{k-1} + \eta C$ for $1 \le k \le K$, then A_K can be expressed explicitly by*

$$A_K = (1-\theta)^K A_0 + \frac{1-(1-\theta)^K}{\theta}\eta C.$$

In particular, when $\eta = \theta^2 < 1$, $A_k \ge 0$, $C \ge 0$ and with $\le$ in place of equality in the iterative formula, a convenient bound for A_k with $\theta = \log K/K$ is that

$$A_K \le \frac{\log K}{K}C + \frac{1}{K}\max\left(A_0 - \frac{\log K}{K}C, 0\right).$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Proof. It is obvious that $A_K = (1-\theta)^K A_0 + \eta C \sum_{k=0}^{K-1}(1-\theta)^k$ and the last summation can be calculated as

$$\sum_{k=0}^{K-1}(1-\theta)^k = \frac{1-(1-\theta)^K}{\theta}.$$

For the choice of θ, bound $(1-\theta)^K \leq 1/K$ and the rest would follow. □

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Acknowledgments

I would like to thank the many people who made this dissertation possible.

It is impossible to overstate my deep and sincere gratitude to my adviser, Professor Andrew R. Barron. His enthusiasm, inspiration, well rounded knowledge and penetrating thoughts have motivated me to explore interesting topics in Statistics, and strengthened my desire to research other unknowns in future.

I am greatly in debt to the many people who taught me various topics in statistics and delivered inspiring discussions on research: Joseph Chang, Lisha Chen, John Emerson, John Hartigan, Hannes Leeb, Mokshay Madiman, David Pollard, and Harrison Zhou. I would like to dedicate my special thanks to Professor John Hartigan who introduced such an exciting field to me that I will pursue for my whole life.

I warmly thank my collaborators outside the department for other interesting work we have done together as well as enhancing my understanding in statistics in other fields. I am especially grateful to Professor Chiang-shan Ray Li at Department of Psychiatry for highly productive collaboration, and I wish to thank Professor Arturo Bris, Professor William Goetzmann, and Professor Shyam Sunder at Yale School of Management.

I wish to thank all my student colleagues for providing an stimulating environment, and the administrative staff for keeping the department running smoothly. Daniel Campbell, Joann DelVecchio, Wei Dou, Chandra Erdman, John Ferguson,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Adityanand Guntuboyina, Summer Han, Xing Hu, Cong Huang, Antony Joseph, Michael Kane, Yang Liu, Xiaoxian Luo, Amy Mulholland, Wei Qiu, Stephan Winkler, Patrica Wooding, and Peisi Yan, deserve special mention.

I am grateful for the financial support from Yale University Fellowship, Yale Bateman Fellowship, Annie G. K. Garland Fellowship, and Yale Dissertation Fellowship. I am also thankful for the partial support during the final semester from NIH through the joint work with Professor Chiang-shan Ray Li.

Lastly, and most importantly, I wish to thank my parents, Jianming Luo and Guifeng Kang. They bore me, raised me, educated me, supported me and loved me. To them I dedicate this dissertation.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Bibliography

[1] BANERJEE, O., GHAOUI, L., AND D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning 9*, 485–516.

[2] BARRON, A. R. (1990). Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics*, G. Roussas, Ed. Kluwer Academic Publishers, 561–576.

[3] BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory 39*, 930–945.

[4] BARRON, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning 14*, 113–143.

[5] BARRON, A. R., BIRGE, L., AND MASSART, P. (1999). Risk bounds for model selection by penalization. *Probab. Th. Re. Fields 113*, 301–413.

[6] BARRON, A. R., COHEN, A., DAHMEN, W., AND DEVORE, R. (2008). Approximation and learning by greedy algorithms. *Annals of Statistics* **36**, 1, 64–94.

[7] BARRON, A. R. AND COVER, T. M. (1991). Minimum complexity density-estimation. *IEEE Trans. on Information Theory* **37**, 4, 1034–1054.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[8] BARRON, A. R., HUANG, C., LI, J. Q., AND LUO, X. (2008a). MDL, penalized likelihood, and statistical risk. In *Proceedings IEEE Information Theory Workshop.* Porto, Portugal.

[9] BARRON, A. R., HUANG, C., LI, J. Q., AND LUO, X. (2008b). MDL principle, penalized likelihood, and statistical risk. In *Feschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, P. Grünwald, P. Myllymäki, I. Tabus, M. Weinberger, and B. Yu, Eds. Tampere International Center for Signal Processing, 33–62.

[10] BARRON, A. R. AND LUO, X. (2008). MDL procedures with ℓ_1 penalty and their statistical risk. In *Proceedings Workshop on Information Theoretic Methods in Science and Engineering.* Tampere University of Technology, Tampere, Finland.

[11] BARRON, A. R., RISSANEN, J., AND YU, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions On Information Theory* **44**, 6, 2743–2760.

[12] BHATTACHARYYA, A. (1943). On a measure of divergence between two statistical populations defined by probability distributions. *Bull. Calcutta Math. Soc. 35*, 99–109.

[13] BICKEL, P. J. AND LEVINS, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics* **36**, 1, 199–227.

[14] BICKEL, P. J., RITOV, Y., AND TSYBAKOV, A. Simultaneous analysis of lasso and dantzig selector. To appear in Annals of Statistics.

[15] BOYD, S. AND VANDENBERGHE, L. (2004). *Convex Optimization.* Cambridge University Press.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[16] BUNEA, F., TSYBAKOV, A., AND WEGKAMP, M. (2007a). Aggregation for gaussian regression. *Annals of Statistics* **35**, 4, 1674–1697.

[17] BUNEA, F., TSYBAKOV, A., AND WEGKAMP, M. (2007b). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics 1*, 169–194.

[18] CAI, T., XU, G., AND ZHANG, J. On recovery of sparse signals via ℓ_1 minimization. To appear in IEEE Transactions on Information Theory.

[19] CAI, T. T., ZHANG, C.-H., AND ZHOU, H. H. Optimal rates of covergence for covariance matrix estimation. Submitted to Annals of Statistics.

[20] CANDES, E. AND PLAN, Y. Near-ideal model selection by ℓ_1 minimization. To appear in Annals of Statistics.

[21] CANDES, E. AND TAO, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35**, 6, 2313–2351.

[22] CHEN, S., DONOHO, D., AND SAUNDERS, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 33–61.

[23] CRAMER, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

[24] DAHL, J., VANDENBERGHE, L., AND ROYCHOWDHURY, V. (2008). Covariance selection for non-chordal graphs via chordal embedding. *Optimization Methods & Software* **23**, 4, 501–520.

[25] DAUBECHIES, I., DEFRISE, M., AND DE MOL, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math. 57*, 1413–1457.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[26] DONOHO, D., ELAD, M., AND TEMLYAKOV, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Information Theory* **52**, 1, 6–18.

[27] DONOHO, D. AND JOHNSTONE, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika 81*, 425–455.

[28] EFRON, B., HASTIE, T., JOHNSTONE, I., AND TIBSHIRANI, R. (2004). Least angle regression. *Annals of Statistics* **32**, 2, 407–499.

[29] FAN, J., FAN, Y., AND LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147**, 1, 186 – 197.

[30] FAN, J. AND LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*, 1348–1360.

[31] FAN, J. AND PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* **32**, 3, 928–961.

[32] FRIEDLANDER, M. AND SAUNDERS, M. (2007). Discussion of "dantzig selector"" by E. Candes and T. Tao. *Annals of Statistics 35*, 2385–2391.

[33] FRIEDMAN, J., HASTIE, T., HOFLING, H., AND TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1**, 2, 302–332.

[34] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. Preprint.

[35] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 3, 432–441.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[36] FU, W. AND KNIGHT, K. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

[37] FU, W. J. AND KNIGHT, K. (1998). Penalized regression: The bridge versus the lasso. *J. Comput. Graph. Statist. 7*, 397–416.

[38] GREENSHTEIN, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint. *Annals of Statistics* **34**, 5, 2367–2386.

[39] GREENSHTEIN, E. AND RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virture of overparametrization. *Bernoulli* **10**, 6, 971–988.

[40] GRUNWALD, P. (2007). *The Minimal Description Length Principle.* MIT Press, Cambridge, MA.

[41] HASTIE, T., ROSSET, S., TIBSHIRANI, R., AND ZHU, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research 5*, 1391–1415.

[42] HUANG, C., CHEANG, G., AND BARRON, A. (2008). Risk of penalized least squares, greedy selection and l1 pnealization for flexible function libraries. Submitted to Annals of Statistics.

[43] HUANG, J., LIU, N., POURAHMADI, M., AND LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika 93*, 85–98.

[44] JOHNSTONE, I. (2001). On the distribution of the largest eigenvalue in principal componets analysis. *Annals of Statistics* **29**, 2, 295–327.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[45] JONES, L. K. (1992). A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics* **20**, 1, 608–613.

[46] KOLACZYK, E. D. AND NOWAK, R. D. (2004). Multiscale likelihood analysis and complexity penalized estimation. *Annals of Statistics 32*, 500–527.

[47] KOLACZYK, E. D. AND NOWAK, R. D. (2005). Multiscale generalized linear models for nonparametric function estimation. *Biometrika* **92**, 1, 119–133.

[48] LEE, W. S., BARTLETT, P. L., AND WILLIAMSON, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions of Information Theory* **42**, 6, 2118–2132.

[49] LEVINA, E., ROTHMAN, A., AND ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Annals of Applied Statistics* **2**, 1, 245–263.

[50] LI, J. Q. (2000). Estimation of mixture models. Ph.D. thesis, Department of Statistics, Yale University.

[51] LUO, Z. Q. AND TSENG, P. (1992). On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications* **72**, 1, 7–35.

[52] MEINSHAUSEN, N. AND BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 3, 1436–1462.

[53] OSBORNE, M. R., PRESNELL, B., AND TURLACH, B. A. (2000a). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20**, 3, 389–403.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[54] OSBORNE, M. R., PRESNELL, B., AND TURLACH, B. A. (2000b). On the lasso and its dual. *Journal of Computional and Graphical Statistics* **9**, 2, 319–337.

[55] RÉNYI, A. (1960). On measures of entropy and information. In Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability. *Proc.*.

[56] SHANNON, C. (1948). The mathematical theory of communication. *Bell System Technical Journal 27*, 379–423, 623–656.

[57] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 1, 267–288.

[58] TSENG, P. (2001). Convergence of block coordinate descent method for non-differentiable maximation. *J. Opt. Theory Appl. 109*, 474–494.

[59] VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics 36*, 614–645.

[60] WEISBERG, S. (1980). *Applied Linear Regression.* Wiley, New York.

[61] WU, W. B. AND POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 4, 831–844.

[62] YANG, Y. AND BARRON, A. R. (1998). An asymptopic property of model selection criteria. *IEEE Transactions On Information Theory 44*, 117–133.

[63] YUAN, M. AND LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B 68*, 49–67.

[64] YUAN, M. AND LIN, Y. (2007). Model selection and estimation in the gaussian graphical model. *Boimetrika* **94**, 1, 19–35.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[65] ZHANG, C. AND HUANG, J. (2006). Model-selection consistency of the lasso in high-dimensional linear regression. Tech. rep., Dept. Statistics, Rutgers Univ.

[66] ZHANG, T. (2003). Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions On Information Theory* **49**, 3, 682–691.

[67] ZHANG, T. (2007). Some sharp performance bounds for least squares regression with ℓ_1 regularization. Tech. rep., Rutgers University.

[68] ZHAO, P. AND YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res. 7*, 2541–2563.

[69] ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.