Statistical Science
2012, Vol. 27, No. 4, 447–449
DOI: 10.1214/12-STS409

Introduction to the Special Issue on Sparsity and Regularization Methods

Jon Wellner and Tong Zhang

1. INTRODUCTION

Traditional statistical inference considers relatively small data sets and the corresponding theoretical analysis focuses on the asymptotic behavior of a statistical estimator when the number of samples approaches infinity. However, many data sets encountered in modern applications have dimensionality significantly larger than the number of training data available, and for such problems the classical statistical tools become inadequate. In order to analyze high-dimensional data, new statistical methodology and the corresponding theory have to be developed.

In the past decade, sparse modeling and the corresponding use of sparse regularization methods have emerged as a major technique to handle high-dimensional data. While the data dimensionality is high, the basic assumption in this approach is that the actual estimator is sparse in the sense that only a small number of components are nonzero. On the practical side, the sparsity phenomenon has been ubiquitously observed in applications, including signal recovery, genomics, computer vision, etc. On the theoretical side, this assumption makes it possible to overcome the problem associated with estimating more parameters than the number of observations which is impossible to deal with in the classical setting.

There are a number of challenges, including developing new theories for high-dimensional statistical estimation as well as new formulations and computational procedures. Related problems have received a lot of attention in various research fields, including applied math, signal processing, machine learning, statistics and optimization. Rapid advances have been made in recent years. In view of the growing research activities and their practical importance, we have organized this special issue of *Statistical Science* with the goal of providing overviews of several topics in modern sparsity analysis and associated regularization methods. Our hope is that general readers will get a broad idea of the field as well as current research directions.

Jon Wellner is Professor of Statistics and Biostatistics, Department of Statistics, University of Washington, Seattle, Washington 98112-2801, USA (e-mail: jaw@stat.washington.edu). Tong Zhang is Professor of Statistics, Department of Statistics, Rutgers University, New Jersey, USA (e-mail: tzhang@stat.rutgers.edu).

2. SPARSE MODELING AND REGULARIZATION

One of the central problem in statistics is linear regression, where we consider an $n \times p$ design matrix X and an n-dimensional response vector $Y \in \mathbb{R}^n$ so that

$$Y = X\bar{\beta} + \varepsilon, \tag{1}$$

where $\bar{\beta} \in \mathbb{R}^p$ is the true regression coefficient vector and $\varepsilon \in \mathbb{R}^n$ is a noise vector. In the case of $n < p$, this problem is ill-posed because the number of parameters is more than the number of observations. This ill-posedness can be resolved by imposing a sparsity constraint: that is, by assuming that $\|\bar{\beta}\|_0 \leq s$ for some s, where the ℓ_0-norm of $\bar{\beta}$ is defined as $\|\bar{\beta}\|_0 = |\mathrm{supp}(\bar{\beta})|$, and the support set of $\bar{\beta}$ is defined as $\mathrm{supp}(\bar{\beta}) := \{j : \bar{\beta}_j \neq 0\}$. If $s \ll n$, then the effective number of parameters in (1) is smaller than the number of observations.

The sparsity assumption may be viewed as the classical model selection problem, where models are indexed by the set of nonzero coefficients. The classical model selection criteria such as AIC, BIC or Cp [1, 7, 11] naturally lead to the so-called ℓ_0 regularization estimator:

$$\hat{\beta}^{(\ell_0)} = \arg\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{n} \|X\beta - Y\|_2^2 + \lambda \|\beta\|_0 \right]. \tag{2}$$

The main difference of modern ℓ_0 analysis in high-dimensional statistics and the classical model selection methods is that the choice of λ will be different, and the modern analysis requires choosing a larger λ than that considered in the classical model selection setting because it is necessary to compensate for the effect of considering many models in the high-dimensional setting. The analysis for ℓ_0 regularization in the high-dimensional setting (e.g., [15] in this issue) employs different techniques and the results obtained are also different from the classical literature.

The ℓ_0 regularization formulation leads to a nonconvex optimization problem that is difficult to solve computationally. On the other hand, an important requirement for modern high-dimensional problems is to design computationally efficient and statistically effective algorithms. Therefore, the main focus of the existing literature is on convex relaxation methods that use ℓ_1-regularization (Lasso) to replace sparsity constraints:

$$\hat{\beta}^{(\ell_1)} = \arg\min_{\beta\in\mathbb{R}^p}\left[\frac{1}{n}\|X\beta - Y\|_2^2 + \lambda\|\beta\|_1\right]. \tag{3}$$

This method is referred to as Lasso [12] in the literature and its theoretical properties have been intensively studied. Since the formulation is regarded as an approximation of (2), a key question is how good this approximation is, and how good is the estimator $\hat{\beta}^{(\ell_1)}$ for estimating $\bar{\beta}$.

Many extensions of Lasso have appeared in the literature for more complex problems. One example is group Lasso [14] that assumes that variables are selected in groups. Another extension is the estimation of graphical models, where one can employ Lasso to estimate unknown graphical model structures [3, 8]. A third example is matrix regularization, where the concept of sparsity can be replaced by the concept of low-rankness, and sparsity constraints become low-rank constraints. Of special interest is the so-called matrix completion problem, where we want to recover a matrix from a few observations of the matrix entries. This problem is encountered in recommender system applications (e.g., a person buys a book at amazon.com will be recommended other books purchased by other users with similar interests), and low-rank matrix factorization is one of the main techniques for this problem. Similar to sparsity regularization, using low-rank regularization leads to nonconvex formulations and, thus, it is natural to consider its convex relaxation which is referred to as trace-norm (or Nuclear norm) regularization. The theoretical properties and numerical algorithms for trace-norm regularization methods have received attention.

3. ARTICLES IN THIS ISSUE

The eight articles in this issue present general overviews of the state of the art in a number of different topics concerning sparsity analysis and regularization methods. Moreover, many articles go beyond the current state of the art in various ways. Therefore, these articles not only give some high level ideas about the current topics, but will also be valuable for experts working in the field.

- Bach, Jenatton, Mairal and Guillaume (Structured sparsity through convex optimization, [2]) study convex relaxations based on structured norms incorporating further structural prior knowledge. An extension of the standard ℓ_0 sparsity model that has received a lot of attention in recent years is *structured sparsity*. The basic idea is that not all sparsity patterns for $\text{supp}(\bar{\beta})$ are equally likely. A simple example is group sparsity where nonzero coefficients occur together in predefined groups. More complex structured sparsity models have been investigated in recent years. Although the paper by Bach et al. focuses on the convex optimization approach, they also give an extensive survey of recent developments, including the use of sub-modular set functions.
- van de Geer and Müller (Quasi-likelihood and/or robust estimation in high dimensions, [13]) extend ℓ_1 regularization methods to generalized linear models. This involves consideration of loss functions beyond the usual least-squares loss and, in particular, loss functions arising via quasi-likelihoods.
- Huang, Breheny and Ma (A selective review of group selection in high-dimensional regression, [5]) provide a detailed review of the most important special case of structured sparsity, namely, group sparsity. Their review covers both convex relaxation (or group Lasso) and approaches based on nonconvex group penalties.
- Huet, Giraud and Verzelen (High-dimensional regression with unknown variance, [4]) address issues in high-dimensional regression estimation connected with lack of knowledge of the error variance. In the standard Lasso formulation (3), the regularization parameter λ is considered as a tuning parameter that needs to be chosen proportionally to the standard deviation σ of the noise vector. A natural question is whether it is possible to automatically estimate σ instead of leaving λ as a tuning parameter. This problem has received much attention and a number of developments have been made in recent years. This paper reviews and compares several approaches to this problem.
- Lafferty, Liu and Wasserman (Sparse nonparametric graphical models, [6]) discuss another important topic in sparsity analysis, the graphical model estimation problem. While much of the current work assumes that the data come from a multivariate Gaussian distribution, this paper goes beyond the standard practice. The authors outline a number of possible approaches and introduce more flexible

models for the problem. The authors also describe some of their recent work, and describe future research directions.

- Negahban, Ravikumar, Wainwright and Yu (A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, [9]) provide a unified treatment of existing approaches to sparse regularization. The paper extends the standard sparse recovery analysis of ℓ_1 regularized least squares regression problems by introducing a general concept of restricted strong convexity. This allows the authors to study more general formulations with different convex loss functions and a class of "decomposable" regularization conditions.
- Rigollet and Tsybakov (Sparse estimation by exponential weighting, [10]) present a thorough analysis of oracle inequalities in the context of model averaging procedures, a class of methods which has its original in the Bayesian literature. Model averaging is in general more stable than model selection. For example, in the scenario that two models are very similar and only one is correct, model selection forces us to choose one of the models even if we are not certain which model is true. On the other hand, a model averaging procedure does not force us to choose one of the two models, but only to take the average of the two models. This is beneficial when several of the models are similar and we cannot tell which is the correct one. The modern analysis of model averaging procedures leads to oracle inequalities that are sharper than the corresponding oracle inequalities for model selection methods such as Lasso. The authors give an extensive discussion of such oracle inequalities using an exponentially weighted model averaging procedure. Such procedures have advantages over model selection when the underlying models are correlated and when the model class is misspecified.
- Zhang and Zhang (A general theory of concave regularization for high-dimensional sparse estimation problems, [15]) focus on nonconvex penalties and study a variety of issues related to such penalties. Although the natural formulation of a sparsity constraint is ℓ_0 regularization, due to its computational difficulty, most of the recent literature focuses on the simpler ℓ_1 regularization method (Lasso) that approximates ℓ_0 regularization. However, it is also known that ℓ_1 regularization is not a very good approximation to ℓ_0 regularization. This leads to the study of nonconvex penalties. The nonconvex formulations are both harder to analyze statistically and harder to handle computationally. Some fundamental understanding of high-dimensional nonconvex procedures has only started to emerge recently. Nevertheless, some basic questions have remained unanswered: for example, properties of the global solution of nonconvex formulations and whether it is possible to compute the global optimal solution efficiently under suitable conditions. The authors go a considerable distance toward providing a general theory that answers some of these fundamental questions.

REFERENCES

[1] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) 267–281. Akadémiai Kiadó, Budapest. MR0483125

[2] BACH, F., JENATTON, R., MAIRAL, J. and OBOZINSKI, G. (2012). Structured sparsity through convex optimization. *Statist. Sci.* **27** 450–468.

[3] BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

[4] GIRAUD, C., HUET, S. and VERZELEN, N. (2012). High-dimensional regression with unknown variance. *Statist. Sci.* **27** 500–518.

[5] HUANG, J., BREHENY, P. and MA, S. (2012). A selective review of group selection in high dimensional models. *Statist. Sci.* **27** 481–499.

[6] LAFFERTY, J., LIU, H. and WASSERMAN, L. (2012). Sparse nonparametric graphical models. *Statist. Sci.* **27** 519–537.

[7] MALLOWS, C. L. (1973). Some comments on Cp. *Technometrics* **12** 661–675.

[8] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[9] NEGAHBAN, S., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557.

[10] RIGOLLET, P. and TSYBAKOV, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575.

[11] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

[12] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

[13] VAN DE GEER, S. and MULLER, P. (2012). Quasi-likelihood and/or robust estimation in high dimensions. *Statist. Sci.* **27** 469–480.

[14] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

[15] ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593.