Statistical challenges with high dimensionality: feature selection in knowledge discovery

Jianqing Fan and Runze Li*

Abstract. Technological innovations have revolutionized the process of scientific research and knowledge discovery. The availability of massive data and challenges from frontiers of research and development have reshaped statistical thinking, data analysis and theoretical studies. The challenges of high-dimensionality arise in diverse fields of sciences and the humanities, ranging from computational biology and health studies to financial engineering and risk management. In all of these fields, variable selection and feature extraction are crucial for knowledge discovery. We first give a comprehensive overview of statistical challenges with high dimensionality in these diverse disciplines. We then approach the problem of variable selection and feature extraction using a unified framework: penalized likelihood methods. Issues relevant to the choice of penalty functions are addressed. We demonstrate that for a host of statistical problems, as long as the dimensionality is not excessively large, we can estimate the model parameters as well as if the best model is known in advance. The persistence property in risk minimization is also addressed. The applicability of such a theory and method to diverse statistical problems is demonstrated. Other related problems with high-dimensionality are also discussed.

Mathematics Subject Classification (2000). Primary 62J99; Secondary 62F12.

Keywords. AIC, BIC, LASSO, bioinformatics, financial econometrics, model selection, oracle property, penalized likelihood, persistent, SCAD, statistical learning.

1. Introduction

Technological innovations have had deep impact on society and on scientific research. They allow us to collect massive amount of data with relatively low cost. Observations with curves, images or movies, along with many other variables, are frequently seen in contemporary scientific research and technological development. For example, in biomedical studies, huge numbers of magnetic resonance images (MRI) and functional MRI data are collected for each subject with hundreds of subjects involved. Satellite imagery has been used in natural resource discovery and agriculture, collecting thousands of high resolution images. Examples of these kinds are plentiful in computational biology, climatology, geology, neurology, health science, economics,

*Fan's research was supported partially by NSF grant DMS-0354223, DMS-0532370 and NIH R01-GM072611. Li's research was supported by NSF grant DMS-0348869 and National Institute on Drug Abuse grant P50 DA10075. The authors would like to thank Professors Peter Hall and Michael Korosok for their constructive comments and John Dziak for his assistance. The article was presented by Jianqing Fan.

Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006

and finance among others. Frontiers of science, engineering and the humanities differ in the problems of their concerns, but nevertheless share one common theme: massive and high-throughput data have been collected and new knowledge needs to be discovered using these data. These massive collections of data along with many new scientific problems create golden opportunities and significant challenges for the development of mathematical sciences.

The availability of massive data along with new scientific problems have reshaped statistical thinking and data analysis. Dimensionality reduction and feature extraction play pivotal roles in all high-dimensional mathematical problems. The intensive computation inherent in these problems has altered the course of methodological development. At the same time, high-dimensionality has significantly challenged traditional statistical theory. Many new insights need to be unveiled and many new phenomena need to be discovered. There is little doubt that the high dimensional data analysis will be the most important research topic in statistics in the 21st century [19].

Variable selection and feature extraction are fundamental to knowledge discovery from massive data. Many variable selection criteria have been proposed in the literature. Parsimonious models are always desirable as they provide simple and interpretable relations among scientific variables in addition to reducing forecasting errors. Traditional variable selection such as C_p, AIC and BIC involves a combinatorial optimization problem, which is NP-hard, with computational time increasing exponentially with the dimensionality. The expensive computational cost makes traditional procedures infeasible for high-dimensional data analysis. Clearly, innovative variable selection procedures are needed to cope with high-dimensionality.

Computational challenges from high-dimensional statistical endeavors forge cross-fertilizations among applied and computational mathematics, machine learning, and statistics. For example, Donoho and Elad [20] and Donoho and Huo [21] show that the NP-hard best subset regression can be solved by a penalized L_1 least-squares problem, which can be handled by a linear programming, when the solution is sufficiently sparse. Wavelets are widely used in statistics function estimation and signal processing [1], [14], [17], [23], [24], [64], [65], [71]. Algebraic statistics, the term coined by Pistone, Riccomagno, Wynn [73], uses polynomial algebra and combinatorial algorithms to solve computational problems in experimental design and discrete probability [73], conditional inferences based on Markovian chains [16], parametric inference for biological sequence analysis [72], and phylogenetic tree reconstruction [78].

In high-dimensional data mining, it is helpful to distinguish two types of statistical endeavors. In many machine learning problems such as tumor classifications based on microarray or proteomics data and asset allocations in finance, the interests often center around the classification errors, or returns and risks of selected portfolios rather than the accuracy of estimated parameters. On the other hand, in many other statistical problems, concise relationship among dependent and independent variables are needed. For example, in health studies, we need not only to identify risk factors, but also to assess accurately their risk contributions. These are needed for prognosis and understanding the relative importance of risk factors. Consistency results are

inadequate for assessing the uncertainty in parameter estimation. The distributions of selected and estimated parameters are needed. Yet, despite extensive studies in classical model selection techniques, no satisfactory solutions have yet been produced.

In this article, we address the issues of variable selection and feature extraction using a unified framework: penalized likelihood methods. This framework is applicable to both machine learning and statistical inference problems. In addition, it is applied to both exact and approximate statistical modeling. We outline, in Section 2, some high-dimensional problems from computational biology, biomedical studies, financial engineering, and machine learning, and then provide a unified framework to address the issues of feature selection in Sections 3 and 4. In Sections 5 and 6, the framework is then applied to provide solutions to some problems outlined in Section 2.

2. Challenges from sciences and humanities

We now outline a few problems from various frontiers of research to illustrate the challenges of high-dimensionality. Some solutions to these problems will be provided in Section 6.

2.1. Computational biology. Bioinformatic tools have been widely applied to genomics, proteomics, gene networks, structure prediction, disease diagnosis and drug design. The breakthroughs in biomedical imaging technology allow scientists to monitor large amounts of diverse information on genetic variation, gene and protein functions, interactions in regulatory processes and biochemical pathways. Such technology has also been widely used for studying neuron activities and networks. Genomic sequence analysis permits us to understand the homologies among different species and infer their biological structures and functionalities. Analysis of the network structure of protein can predict the protein biological function. These quantitative biological problems raise many new statistical and computational problems. Let us focus specifically on the analysis of microarray data to illustrate some challenges with dimensionality.

DNA microarrays have been widely used in simultaneously monitoring mRNA expressions of thousands of genes in many areas of biomedical research. There are two popularly-used techniques: c-DNA microarrays [5] and Affymetrix GeneChip arrays [61]. The former measures the abundance of mRNA expressions by mixing mRNAs of treatment and control cells or tissues, hybridizing with cDNA on the chip. The latter uses combined intensity information from 11-20 probes interrogating a part of the DNA sequence of a gene, measuring separately mRNA expressions of treatment and control cells or tissues. Let us focus further on the cDNA microarray data.

The first statistical challenge is to remove systematic biases due to experiment variations such as intensity effect in the scanning process, block effect, dye effect, batch effect, amount of mRNA, DNA concentration on arrays, among others. This is collectively referred to as normalization in the literature. Normalization is critical

for multiple array comparisons. Statistical models are needed for estimation of these systematic biases in presence of high-dimensional nuisance parameters from treatment effects on genes. See, for example, lowess normalization in [26], [83], semiparametric model-based normalization by [36], [37], [50], and robust normalization in [63]. The number of significantly expressed genes is relatively small. Hence, model selection techniques can be used to exploit the sparsity. In Section 6.1, we briefly introduce semiparametric modeling techniques to issues of normalization of cDNA microarray.

Once systematic biases have been removed, the statistical challenge becomes selecting statistically significant genes based on a relatively small sample size of arrays (e.g. $n = 4, 6, 8$). Various testing procedures have been proposed in the literature. See, for example, [30], [37], [50], [83], [84]. In carrying out simultaneous testing of orders of hundreds or thousands of genes, classical methods of controlling the probability of making one falsely discovered gene are no longer relevant. Therefore various innovative methods have been proposed to control the false discovery rates. See, for example, [2], [22], [25], [27], [44], [57], [77]. The fundamental assumption in these developments is that the null distribution of test statistics can be determined accurately. This assumption is usually not granted in practice and new probabilistic challenge is to answer the questions how many simultaneous hypotheses can be tested before the accuracy of approximations of null distributions becomes poor. Large deviation theory [45], [46], [53] is expected to play a critical role in this endeavor. Some progress has been made using maximal inequalities [55].

Tumor classification and clustering based on microarray and proteomics data are another important class of challenging problems in computational biology. Here, hundreds or thousands of gene expressions are potential predictors, and the challenge is to select important genes for effective disease classification and clustering. See, for example, [79], [82], [88] for an overview and references therein.

Similar problems include time-course microarray experiments used to determine the expression pathways over time [79], [80] and genetic networks used for understanding interactions in regulatory processes and biochemical pathways [58]. Challenges of selecting significant genes over time and classifying patterns of gene expressions remain. In addition, understanding genetic network problems requires estimating a huge covariance matrix with some sparsity structure. We introduce a modified Cholesky decomposition technique for estimating large scale covariance matrices in Section 6.1.

2.2. Health studies. Many health studies are longitudinal: each subject is followed over a period of time and many covariates and responses of each subject are collected at different time points. Framingham Heart Study (FHS), initiated in 1948, is one of the most famous classic longitudinal studies. Documentation of its first 50 years can be found at the website of National Heart, Lung and Blood Institute (http://www.nhlbi.nih.gov/about/framingham/). One can learn more details about this study from the website of American Heart Association. In brief, the FHS follows a representative sample of 5,209 adult residents and their offspring aged 28–62 years in

Framingham, Massachusetts. These subjects have been tracked using (a) standardized biennial cardiovascular examination, (b) daily surveillance of hospital admissions, (c) death information and (d) information from physicians and other sources outside the clinic.

In 1971 the study enrolled a second-generation group to participate in similar examinations. It consisted of 5,124 of the original participants' adult children and their spouses. This second study is called the Framingham Offspring Study.

The main goal of this study is to identify major risk factors associated with heart disease, stroke and other diseases, and to learn the circumstances under which cardiovascular diseases arise, evolve and end fatally in the general population. The findings in this studies created a revolution in preventive medicine, and forever changed the way the medical community and general public view on the genesis of disease. In this study, there are more than 25,000 samples, each consisting of more than 100 variables. Because of the nature of this longitudinal study, some participant cannot be followed up due to their migrations. Thus, the collected data contain many missing values. During the study, cardiovascular diseases may develop for some participants, while other participants may never experience with cardiovascular diseases. This implies that some data are censored because the event of particular interest never occurred. Furthermore, data between individuals may not be independent because data for individuals in a family are clustered and likely positively correlated. Missing, censoring and clustering are common features in health studies. These three issues make data structure complicated and identification of important risk factors more challenging. In Section 6.2, we present a penalized partial likelihood approach to selecting significant risk factors for censored and clustering data. The penalized likelihood approach has been used to analyze a data subset of Framingham study in [9].

High-dimensionality is frequently seen in many other biomedical studies. For example, ecological momentary assessment data have been collected for smoking cessation studies. In such a study, each of a few hundreds participants is provided a hand-held computer, which is designed to randomly prompt the participants five to eight times per day over a period of about 50 days and to provide 50 questions at each prompt. Therefore, the data consist of a few hundreds of subjects and each of them may have more than ten thousand observed values [60]. Such data are termed intensive longitudinal data. Classical longitudinal methods are inadequate for such data. Walls and Schafer [86] presents more examples of intensive longitudinal data and some useful models to analyze this kind of data.

2.3. Financial engineering and risk management. Technological revolution and trade globalization have introduced a new era of financial markets. Over the last three decades, an enormous number of new financial products have been created to meet customers' demands. For example, to reduce the impact of the fluctuations of currency exchange rates on corporate finances, a multinational corporation may decide to buy options on the future of exchange rates; to reduce the risk of price fluctuations of a commodity (e.g. lumbers, corns, soybeans), a farmer may enter

into a future contract of the commodity; to reduce the risk of weather exposures, amusement parks and energy companies may decide to purchase financial derivatives based on the weather. Since the first options exchange opened in Chicago in 1973, the derivative markets have experienced extraordinary growth. Professionals in finance now routinely use sophisticated statistical techniques and modern computing power in portfolio management, securities regulation, proprietary trading, financial consulting, and risk management. For an overview, see [29] and references therein.

Complex financial markets [51] make portfolio allocation, asset pricing and risk management very challenging. For example, the price of a stock depends not only on its past values, but also its bond and derivative prices. In addition, it depends on prices of related companies and their derivatives, and on overall market conditions. Hence, the number of variables that influence asset prices can be huge and the statistical challenge is to select important factors that capture the market risks. Thanks to technological innovations, high-frequency financial data are now available for an array of different financial instruments over a long time period. The amount of financial data available to financial engineers is indeed astronomical.

Let us focus on a specific problem to illustrate the challenge of dimensionality. To optimize the performance of a portfolio [10], [12] or to manage the risk of a portfolio [70], we need to estimate the covariance matrix of the returns of assets in the portfolio. Suppose that we have 200 stocks to be selected for asset allocation. There are 20,200 parameters in the covariance matrix. This is a high-dimensional statistical problem and estimating it accurately poses challenges.

Covariance matrices pervade every facet of financial econometrics, from asset allocation, asset pricing, and risk management, to derivative pricing and proprietary trading. As mentioned earlier, they are also critical for studying genetic networks [58], as well as other statistical applications such as climatology [54]. In Section 6.1, a modified Cholesky decomposition is used to estimate huge covariance matrices using penalized least squares approach proposed in Section 2. We will introduce a factor model for covariance estimation in Section 6.3.

2.4. Machine learning and data mining. Machine learning and data mining extend traditional statistical techniques to handle problems with much higher dimensionality. The size of data can also be astronomical: from grocery sales and financial market trading to biomedical images and natural resource surveys. For an introduction, see the books [47], [48]. Variable selections and feature extraction are vital for such high-dimensional statistical explorations. Because of the size and complexity of the problems, the associated mathematical theory also differs from the traditional approach. The dimensionality of variables is comparable with the sample size and can even be much higher than the sample size. Selecting reliable predictors to minimize risks of prediction is fundamental to machine learning and data mining. On the other hand, as the interest mainly lies in risk minimization, unlike traditional statistics, the model parameters are only of secondary interest. As a result, crude consistency results suffice for understanding the performance of learning theory. This eases considerably

the mathematical challenges of high-dimensionality. For example, in the supervised (classification) or unsupervised (clustering) learning, we do not need to know the distributions of estimated coefficients in the underlying model. We only need to know the variables and their estimated parameters in the model. This differs from high-dimensional statistical problems in health sciences and biomedical studies, where statistical inferences are needed in presence of high-dimensionality. In Sections 4.2 and 6.4, we will address further the challenges in machine learning.

3. Penalized least squares

With the above background, we now consider the variable selection in the least-squares setting to gain further insights. The idea will be extended to the likelihood or pseudo-likelihood setting in the next section. We demonstrate how to directly apply the penalized least squares approach for function estimation or approximation using wavelets or spline basis, based on noisy data in Section 5. The penalized least squares method will be further extended to penalized empirical risk minimization for machine learning in Section 6.4.

Let $\{\boldsymbol{x}_i, y_i\}$, $i = 1, \ldots, n$, be a random sample from the linear regression model

$$y = \boldsymbol{x}^T \boldsymbol{\beta} + \varepsilon, \tag{3.1}$$

where ε is a random error with mean 0 and finite variance σ^2, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ is the vector of regression coefficients. Here, we assume that all important predictors, and their interactions or functions are already in the model so that the full model (3.1) is correct.

Many variable selection criteria or procedures are closely related to minimize the following penalized least squares (PLS)

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{3.2}$$

where d is the dimension of $\boldsymbol{x}$, and $p_{\lambda_j}(\cdot)$ is a penalty function, controlling model complexity. The dependence of the penalty function on j allows us to incorporate prior information. For instance, we may wish to keep certain important predictors in the model and choose not to penalize their coefficients.

The form of $p_{\lambda_j}(\cdot)$ determines the general behavior of the estimator. With the entropy or L_0-penalty, namely, $p_{\lambda_j}(|\beta_j|) = \frac{1}{2}\lambda^2 I(|\beta_j| \neq 0)$, the PLS (3.2) becomes

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \frac{1}{2}\lambda^2 |M|, \tag{3.3}$$

where $|M| = \sum_j I(|\beta_j| \neq 0)$, the size of the candidate model. Among models with m variables, the selected model is the one with the minimum residual sum of squares

(RRS), denoted by RSS_m. A classical statistical method is to choose m by maximizing the adjusted R^2, given by

$$R_{\mathrm{adj},m} = 1 - \frac{n-1}{n-m}\frac{\mathrm{RSS}_m}{\mathrm{RSS}_1},$$

or equivalently by minimizing $\mathrm{RSS}_m/(n-m)$, where RSS_1 is the total sum of squares based on the null model (using the intercept only). Using $\log(1+x) \approx x$ for small x, it follows that

$$\log\{\mathrm{RSS}_m/(n-m)\} \approx (\log\sigma^2 - 1) + \sigma^{-2}\left\{\frac{1}{n}\mathrm{RSS}_m + \frac{1}{n}m\sigma^2\right\}. \tag{3.4}$$

Therefore, maximization of $R_{\mathrm{adj},m}$ is asymptotically equivalent to minimizing the PLS (3.3) with $\lambda = \sigma/\sqrt{n}$. Similarly, generalized cross-validation (GCV) given by

$$\mathrm{GCV}(m) = \mathrm{RSS}_m/\{n(1-m/n)^2\}$$

is asymptotically equivalent to the PLS (3.3) with $\lambda = \sqrt{2}\sigma/\sqrt{n}$ and so is the cross-validation (CV) criterion.

Many popular variable selection criteria can be shown asymptotically equivalent to the PLS (3.3) with appropriate values of λ, though these criteria were motivated from different principles. See [69] and references therein. For instance, RIC [38] corresponds to $\lambda = \sqrt{2\log(d)}(\sigma/\sqrt{n})$. Since the entropy penalty function is discontinuous, minimizing the entropy-penalized least-squares requires exhaustive search, which is not feasible for high-dimensional problem. In addition, the sampling distributions of resulting estimates are hard to derive.

Many researchers have been working on minimizing the PLS (3.2) with L_p-penalty for some $p > 0$. It is well known that the L_2-penalty results in a ridge regression estimator, which regularizes and stabilizes the estimator but introduces biases. However, it does not shrink any coefficients directly to zero.

The L_p-penalty with $0 < p < 2$ yields bridge regression [39], intermediating the best-subset (L_0-penalty) and the ridge regression (L_2-penalty). The non-negative garrote [8] shares the same spirit as that of bridge regression. With the L_1-penalty specifically, the PLS estimator is called LASSO in [81]. In a seminal paper, Donoho and Elad [20] show that penalized L_0-solution can be found by using penalized L_1-method for sparse problem. When $p \leq 1$, the PLS automatically performs variable selection by removing predictors with very small estimated coefficients.

Antoniadis and Fan [1] discussed how to choose a penalty function for wavelets regression. Fan and Li [33] advocated penalty functions with three properties:

a. *Sparsity:* The resulting estimator should automatically set small estimated coefficients to zero to accomplish variable selection.

b. *Unbiasedness:* The resulting estimator should have low bias, especially when the true coefficient β_j is large.

c. *Continuity:* The resulting estimator should be continuous to reduce instability in model prediction.

To gain further insights, let us assume that the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ for model (3.1) is orthogonal and satisfies that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = I_d$. Let $\boldsymbol{z} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ be the least squares estimate of $\boldsymbol{\beta}$. Then (3.2) becomes

$$\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{z}\| + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{\beta}\|^2 + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|).$$

Thus the PLS reduces to a componentwise minimization problem:

$$\min_{\beta_j}\left\{\frac{1}{2}(z_j - \beta_j)^2 + p_{\lambda_j}(|\beta_j|)\right\}, \quad \text{for } j = 1, \ldots, d,$$

where z_j is the j-th component of $\boldsymbol{z}$. Suppress the subscript j and let

$$Q(\beta) = \frac{1}{2}(z - \beta)^2 + p_\lambda(|\beta|). \tag{3.5}$$

Then the first order derivative of $Q(\beta)$ is given by

$$Q'(\beta) = \beta - z + p'_\lambda(|\beta|)\mathrm{sgn}(\beta) = \mathrm{sgn}(\beta)\{|\beta| + p'_\lambda(|\beta|)\} - z.$$

Antoniadis and Fan [1] and Fan and Li [33] derived that the PLS estimator possesses the following properties:

(a) *sparsity* if $\min_\beta\{|\beta| + p'_\lambda(|\beta|)\} > 0$;

(b) *unbiasedness* $p'_\lambda(|\beta|) = 0$ for large $|\beta|$;

(c) *continuity* if and only if $\mathrm{argmin}_\beta\{|\beta| + p'_\lambda(|\beta|)\} = 0$.

The L_p-penalty with $0 \le p < 1$ does not satisfy the continuity condition, the L_1 penalty does not satisfy the unbiasedness condition, and L_p with $p > 1$ does not satisfy the sparsity condition. Therefore, none of the L_p-penalties satisfies the above three conditions simultaneously, and L_1-penalty is the such penalty that is both convex and produces sparse solutions. Of course, the class of penalty functions satisfying the aforementioned three conditions are infinitely many. Fan and Li [33] suggested the use of the smoothly clipped absolute deviation (SCAD) penalty defined as

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & \text{if } 0 \le |\beta| < \lambda; \\ -(|\beta|^2 - 2a\lambda|\beta| + \lambda^2)/\{2(a-1)\}, & \text{if } \lambda \le |\beta| < a\lambda; \\ (a+1)\lambda^2/2, & \text{if } |\beta| \ge a\lambda. \end{cases}$$

They further suggested using $a = 3.7$. This function has similar feature to the penalty function $\lambda|\beta|/(1 + |\beta|)$ advocated in [71]. Figure 1 depicts the SCAD, $L_{0.5}$-penalty,

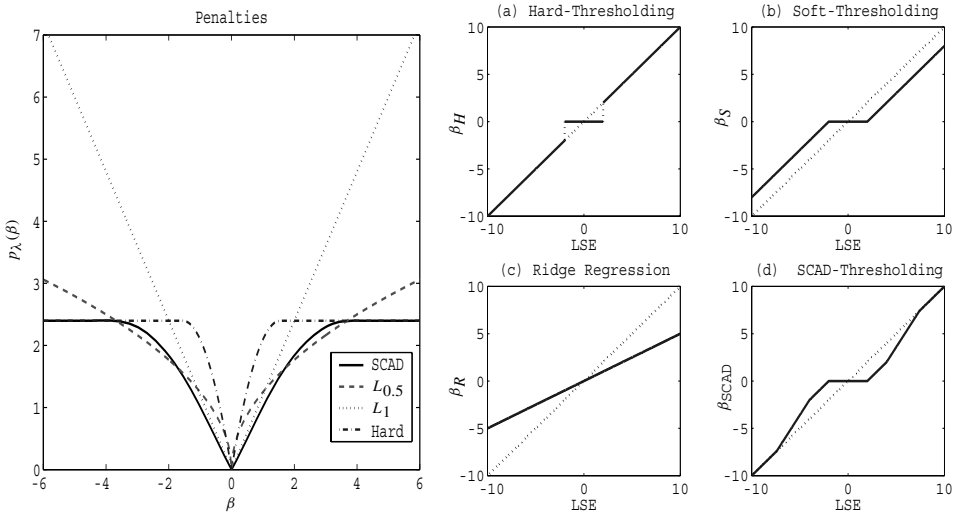

Figure 1. Penalty functions (left panel) and PLS estimators (right panel).

L_1-penalty, and hard thresholding penalty (to be introduced) functions. These four penalty functions are singular at the origin, a necessary condition for sparsity in variable selection. Furthermore, the SCAD, hard-thresholding and $L_{0.5}$ penalties are nonconvex over $(0, +\infty)$ in order to reduce the estimation bias.

Minimizing the PLS (3.5) with the entropy penalty or hard-thresholding penalty $p_\lambda(\beta) = \lambda^2 - (\lambda - |\beta|)_+^2$ (which is smoother) yields the hard-thresholding rule [23] $\hat{\beta}_H = zI(|z| > \lambda)$. With the L_1-penalty, the PLS estimator is $\hat{\beta}_S = \text{sgn}(z)(|z| - \lambda)_+$, the soft-thresholding rule [3], [23]. The L_2-penalty results in the ridge regression $\hat{\beta}_R = (1 + \lambda)^{-1} z$ and the SCAD penalty gives the solution

$$\hat{\beta}_{\text{SCAD}} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \le 2\lambda; \\ \{(a-1)z - \text{sgn}(z)a\lambda\}/(a-2), & \text{when } 2\lambda < |z| \le a\lambda; \\ z, & \text{when } |z| > a\lambda. \end{cases}$$

These functions are also shown in Figure 1. The SCAD is an improvement over the L_0-penalty in two aspects: saving computational cost and resulting in a continuous solution to avoid unnecessary modeling variation. Furthermore, the SCAD improves bridge regression by reducing modeling variation in model prediction. Although similar in spirit to the L_1-penalty, the SCAD also improves the L_1-penalty by avoiding excessive estimation bias since the solution of the L_1-penalty could shrink all regression coefficients by a constant, e.g., the soft thresholding rule.

4. Penalized likelihood

PLS can easily be extended to handle a variety of response variables, including binary response, counts, and continuous response. A popular family of this kind is called generalized linear models. Our approach can also be applied to the case where the likelihood is a quasi-likelihood or other discrepancy functions. This will be demonstrated in Section 6.2 for analysis of survival data, and in Section 6.4 for machine learning.

Suppose that conditioning on $\boldsymbol{x}_i$, y_i has a density $f\{g(\boldsymbol{x}_i^T\boldsymbol{\beta}), y_i\}$, where g is a known inverse link function. Define a penalized likelihood as

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \log f\{g(\boldsymbol{x}_i^T\boldsymbol{\beta}), y_i\} - \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|). \tag{4.1}$$

Maximizing the penalized likelihood results in a penalized likelihood estimator. For certain penalties, such as the SCAD, the selected model based on the nonconcave penalized likelihood satisfies $\beta_j = 0$ for certain β_j's. Therefore, parameter estimation is performed at the same time as the model selection.

Example (*Logistics Regression*). Suppose that given $\boldsymbol{x}_i$, y_i follows a Bernoulli distribution with success probability $P\{y_i = 1|\boldsymbol{x}_i\} = p(\boldsymbol{x}_i)$. Take $g(u) = \exp(u)/(1 + \exp(u))$, i.e. $p(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})/\{1 + \exp(\boldsymbol{x}^T\boldsymbol{\beta})\}$. Then (4.1) becomes

$$\frac{1}{n}\sum_{i=1}^{n}[y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \log\{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\}] - \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|).$$

Thus, variable selection for logistics regression can be achieved by maximizing the above penalized likelihood.

Example (*Poisson Log-linear Regression*). Suppose that given $\boldsymbol{x}_i$, y_i follows a Poisson distribution with mean $\lambda(\boldsymbol{x}_i)$. Take $g(\cdot)$ to be the log-link, i.e. $\lambda(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$. Then (4.1) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\} - \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|)$$

after dropping a constant. Thus, maximizing the above penalized likelihood with certain penalty functions yields a sparse solution for $\boldsymbol{\beta}$.

4.1. Oracle properties. Maximizing a penalized likelihood selects variables and estimates parameters simultaneously. This allows us to establish the sampling properties of the resulting estimators. Under certain regularity conditions, Fan and Li [33] demonstrated how the rates of convergence for the penalized likelihood estimators

depend on the regularization parameter λ_n and established the oracle properties of the penalized likelihood estimators.

In the context of variable selection for high-dimensional modeling, it is natural to allow the number of introduced variables to grow with the sample sizes. Fan and Peng [35] have studied the asymptotic properties of the penalized likelihood estimator for situations in which the number of parameters, denoted by d_n, tends to ∞ as the sample size n increases. Denote $\boldsymbol{\beta}_{n0}$ to be the true value of $\boldsymbol{\beta}$. To emphasize the dependence of λ_j on n, we use notation $\lambda_{n,j}$ for λ_j in this subsection. Define

$$a_n = \max\{p'_{\lambda_{n,j}}(|\beta_{n0j}|) : \beta_{n0j} \neq 0\} \quad \text{and} \quad b_n = \max\{|p''_{\lambda_{n,j}}(|\beta_{n0j}|)| : \beta_{n0j} \neq 0\}. \tag{4.2}$$

Fan and Peng [35] showed that if both a_n and b_n tend to 0 as $n \to \infty$, then under certain regularity conditions, there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{n0}\| = O_P\{\sqrt{d_n}(n^{-1/2} + a_n)\}. \tag{4.3}$$

It is clear from (4.3) that by choosing a proper $\lambda_{n,j}$ such that $a_n = O(n^{-1/2})$, there exists a root-(n/d_n) consistent penalized likelihood estimator. For example, for the SCAD, the penalized likelihood estimator is root-(n/d_n) consistent if all $\lambda_{n,j}$'s tend to 0.

Without loss of generality assume that, unknown to us, the first s_n components of $\boldsymbol{\beta}_{n0}$, denoted by $\boldsymbol{\beta}_{n01}$, are nonzero and do not vanish and the remaining $d_n - s_n$ coefficients, denoted by $\boldsymbol{\beta}_{n02}$, are 0. Denote by

$$\Sigma = \operatorname{diag}\bigl\{p''_{\lambda_{n,1}}(|\beta_{n01}|), \ldots, p''_{\lambda_{n,s_n}}(|\beta_{n0s_n}|)\bigr\}$$

and

$$\boldsymbol{b} = \bigl(p'_{\lambda_{n,1}}(|\beta_{n01}|)\operatorname{sgn}(\beta_{n01}), \ldots, p'_{\lambda_{n,s_n}}(|\beta_{n0s_n}|)\operatorname{sgn}(\beta_{n0s_n})\bigr)^T.$$

Theorem 1. *Assume that as* $n \to \infty$, $\min_{1\le j\le s_n} |\beta_{n0j}|/\lambda_{n,j} \to \infty$ *and that the penalty function* $p_{\lambda_j}(|\beta_j|)$ *satisfies*

$$\liminf_{n\to\infty} \liminf_{\beta_j \to 0+} p'_{\lambda_{n,j}}(\beta_j)/\lambda_{n,j} > 0. \tag{4.4}$$

If $\lambda_{n,j} \to 0$, $\sqrt{n/d_n}\lambda_{n,j} \to \infty$ *and* $d_n^5/n \to 0$ *as* $n \to \infty$, *then with probability tending to* 1, *the root* n/d_n *consistent local maximizers* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{n1}^T, \hat{\boldsymbol{\beta}}_{n2}^T)^T$ *must satisfy:*

(i) *Sparsity:* $\hat{\boldsymbol{\beta}}_{n2} = \mathbf{0}$;

(ii) *Asymptotic normality: for any* $q \times s_n$ *matrix* $\boldsymbol{A}_n$ *such that* $\boldsymbol{A}_n\boldsymbol{A}_n^T \to \boldsymbol{G}$, *a* $q \times q$ *positive definite symmetric matrix,*

$$\sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_1^{-1/2}\{\boldsymbol{I}_1 + \Sigma\}\bigl\{\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10} + (\boldsymbol{I}_1 + \Sigma)^{-1}\boldsymbol{b}\bigr\} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{G})$$

where $\boldsymbol{I}_1 = \boldsymbol{I}_1(\boldsymbol{\beta}_{n10}, \mathbf{0})$, *the Fisher information knowing* $\boldsymbol{\beta}_{n20} = \mathbf{0}$.

The theorem implies that any finite set of elements of $\hat{\boldsymbol{\beta}}_{n1}$ are jointly asymptotically normal. For the SCAD, if all $\lambda_{j,n} \to 0$, $a_n = 0$. Hence, when $\sqrt{n/d_n}\lambda_{n,j} \to \infty$, its corresponding penalized likelihood estimators possess the oracle property, i.e., perform as well as the maximum likelihood estimates for estimating $\boldsymbol{\beta}_{n1}$ knowing $\boldsymbol{\beta}_{n2} = \mathbf{0}$. That is, with probability approaching to 1,

$$\hat{\boldsymbol{\beta}}_{n2} = 0, \quad \text{and} \quad \sqrt{n}\boldsymbol{A}_n \boldsymbol{I}_1^{1/2}(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10}) \to N(\mathbf{0}, \boldsymbol{G}).$$

For the L_1-penalty, $a_n = \max_j \lambda_{j,n}$. Hence, the root-$n/d_n$ consistency requires that $\lambda_{n,j} = O(\sqrt{d_n/n})$. On the other hand, the oracle property in Theorem 2 requires that $\sqrt{n/d_n}\lambda_{n,j} \to \infty$. These two conditions for LASSO cannot be satisfied simultaneously. It has indeed been shown that the oracle property does not hold for the L_1-penalty even in the finite parameter setting [90].

4.2. Risk minimization and persistence. In machine learning such as tumor classifications, the primary interest centers on the misclassification errors or more generally expected losses, not the accuracy of estimated parameters. This kind of properties is called persistence in [42], [43].

Consider predicting the response Y using a class of model $g(\boldsymbol{x}^T\boldsymbol{\beta})$ with a loss function $\ell\{g(\boldsymbol{X}^T\boldsymbol{\beta}), Y)$. Then the risk is

$$L_n(\boldsymbol{\beta}) = E\ell\{g(\boldsymbol{X}^T\boldsymbol{\beta}), Y\},$$

where n is used to stress the dependence of dimensionality d on n. The minimum risk is obtained at $\boldsymbol{\beta}_n^* = \operatorname{argmin}_{\boldsymbol{\beta}} L_n(\boldsymbol{\beta})$. In the likelihood context, $\ell = -\log f$. Suppose that there is an estimator $\hat{\boldsymbol{\beta}}_n$ based on a sample of size n. This can be done by the penalized empirical risk minimization similarly to (4.1):

$$n^{-1}\sum_{i=1}^{n} \ell\{g(\boldsymbol{x}_i^T\boldsymbol{\beta}), y_i\} + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{4.5}$$

based on a set of training data $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$. The persistence requires

$$L_n(\hat{\boldsymbol{\beta}}_n) - L_n(\boldsymbol{\beta}_n^*) \xrightarrow{P} 0, \tag{4.6}$$

but not necessarily the consistency of $\hat{\boldsymbol{\beta}}_n$ to $\boldsymbol{\beta}_n^*$. This is in general a much weaker mathematical requirement. Greenshtein and Ritov [43] show that if the non-sparsity rate $s_n = O\{(n/\log n)^{1/2}\}$ and $d_n = n^\alpha$ for some $\alpha > 1$, LASSO (penalized L_1 least-squares) is persistent under the quadratic loss. Greenshtein [42] extends the results to the case where $s_n = O\{n/\log n\}$ and more general loss functions. Meinshausen [66] considers a case with finite non-sparsity s_n but with $\log d_n = n^\xi$, with $\xi \in (0, 1)$. It is shown there that for the quadratic loss, LASSO is persistent, but the rate to persistency is slower than a relaxed LASSO. This again shows the bias problems in LASSO.

4.3. Issues in practical implementation. In this section, we address practical implementation issues related to the PLS and penalized likelihood.

Local quadratic approximation (LQA). The L_p, $(0 < p < 1)$, and SCAD penalty functions are singular at the origin, and they do not have continuous second order derivatives. Therefore, maximizing the nonconcave penalized likelihood is challenging. Fan and Li [33] propose locally approximating them by a quadratic function as follows. Suppose that we are given an initial value $\boldsymbol{\beta}^0$ that is close to the optimizer of $Q(\boldsymbol{\beta})$. For example, take initial value to be the maximum likelihood estimate (without penalty). Under some regularity conditions, the initial value is a consistent estimate for $\boldsymbol{\beta}$, and therefore it is close to the true value. Thus, we can locally approximate the penalty function by a quadratic function as

$$p_{\lambda_n}(|\beta_j|) \approx p_{\lambda_n}(|\beta_j^0|) + \frac{1}{2}\{p'_{\lambda_n}(|\beta_j^0|)/|\beta_j^0|\}(\beta_j^2 - \beta_j^{02}), \quad \text{for } \beta_j \approx \beta_j^0. \tag{4.7}$$

To avoid numerical instability, we set $\hat{\beta}_j = 0$ if β_j^0 is very close to 0. This corresponds to deleting x_j from the final model. With the aid of the LQA, the optimization of penalized least-squares, penalized likelihood or penalized partial likelihood (see Section 6.2) can be carried out by using the Newton–Raphson algorithm. It is worth noting that the LQA should be updated at each step during the course of iteration of the algorithm. We refer to the modified Newton–Raphson algorithm as the LQA algorithm.

The convergence property of the LQA algorithm was studied in [52], whose authors first showed that the LQA plays the same role as the E-step in the EM algorithm [18]. Therefore the behavior of the LQA algorithm is similar to the EM algorithm. Unlike the original EM algorithm, in which a full iteration for maximization is carried out after every E-step, we update the LQA at each step during the iteration course. This speeds up the convergence of the algorithm. The convergence rate of the LQA algorithm is quadratic which is the same as that of the modified EM algorithm [56].

When the algorithm converges, the estimator satisfies the condition

$$\partial \ell(\hat{\boldsymbol{\beta}})/\partial \beta_j + n p'_{\lambda_j}(|\hat{\beta}_j|)\operatorname{sgn}(\hat{\beta}_j) = 0,$$

the penalized likelihood equation, for non-zero elements of $\hat{\boldsymbol{\beta}}$.

Standard error formula. Following conventional techniques in the likelihood setting, we can estimate the standard error of the resulting estimator by using the sandwich formula. Specifically, the corresponding sandwich formula can be used as an estimator for the covariance of the estimator $\hat{\boldsymbol{\beta}}_1$, the non-vanishing component of $\hat{\boldsymbol{\beta}}$. That is,

$$\widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}}_1) = \{\nabla^2 \ell(\hat{\boldsymbol{\beta}}_1) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_1)\}^{-1} \widehat{\operatorname{cov}}\{\nabla \ell(\hat{\boldsymbol{\beta}}_1)\}\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}_1) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_1)\}^{-1}, \tag{4.8}$$

where $\widehat{\operatorname{cov}}\{\nabla \ell(\hat{\boldsymbol{\beta}}_1)\}$ is the usual empirically estimated covariance matrix and

$$\Sigma_\lambda(\hat{\boldsymbol{\beta}}_1) = \operatorname{diag}\{p'_{\lambda_1}(|\hat{\beta}_1|)/|\hat{\beta}_1|, \ldots, p'_{\lambda_{s_n}}(|\hat{\beta}_{s_n}|)/|\hat{\beta}_{s_n}|\}$$

and s_n the dimension of $\hat{\boldsymbol{\beta}}_1$. Fan and Peng [35] demonstrated the consistency of the sandwich formula:

Theorem 2. *Under the conditions of Theorem* 1*, we have*

$$\boldsymbol{A}_n \widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_1)\boldsymbol{A}_n^T - \boldsymbol{A}_n \Sigma_n \boldsymbol{A}_n^T \xrightarrow{P} 0 \quad \text{as } n \to \infty$$

for any matrix $\boldsymbol{A}_n$ *such that* $\boldsymbol{A}_n \boldsymbol{A}_n^T = \boldsymbol{G}$*, where* $\Sigma_n = (\boldsymbol{I}_1 + \Sigma)^{-1}\boldsymbol{I}_1^{-1}(\boldsymbol{I}_1 + \Sigma)^{-1}$.

Selection of regularization parameters. To implement the methods described in previous sections, it is desirable to have an automatic method for selecting the thresholding parameter λ in $p_\lambda(\cdot)$ based on data. Here, we estimate λ via minimizing an approximate generalized cross-validation (GCV) statistic in [11]. By some straightforward calculation, the effective number of parameters for $Q(\boldsymbol{\beta})$ in the last step of the Newton–Raphson algorithm iteration is

$$e(\boldsymbol{\lambda}) \equiv e(\lambda_1, \ldots, \lambda_d) = \mathrm{tr}[\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}\nabla^2 \ell(\hat{\boldsymbol{\beta}})].$$

Therefore the generalized cross-validation statistic is defined by

$$\mathrm{GCV}(\boldsymbol{\lambda}) = -\ell(\hat{\boldsymbol{\beta}})/[n\{1 - e(\boldsymbol{\lambda})/n\}^2]$$

and $\hat{\boldsymbol{\lambda}} = \mathrm{argmin}_{\boldsymbol{\lambda}}\{\mathrm{GCV}(\boldsymbol{\lambda})\}$ is selected.

To find an optimal $\boldsymbol{\lambda}$, we need to minimize the GCV over a d_n-dimensional space. This is an unduly onerous task. Intuitively, it is expected that the magnitude of λ_j should be proportional to the standard error of the maximum likelihood estimate of β_j. Thus, we set $\boldsymbol{\lambda} = \lambda \mathrm{se}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})$ in practice, where $\mathrm{se}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})$ denotes the standard error of the MLE. Therefore, we minimize the GCV score over the one-dimensional space, which will save a great deal of computational cost. The behavior of such a method has been investigated recently.

5. Applications to function estimation

Let us begin with one-dimensional function estimation. Suppose that we have noisy data at possibly irregular design points $\{x_1, \ldots, x_n\}$:

$$y_i = m(x_i) + \varepsilon_i,$$

where m is an unknown regression and ε_i's are iid random error following $N(0, \sigma^2)$. Local modeling techniques [31] have been widely used to estimate $m(\cdot)$. Here we focus on global function approximation methods.

Wavelet transforms are a device for representing functions in a way that is local in both time and frequency domains [13], [14], [64], [65]. During the last decade, they have received a great deal of attention in applied mathematics, image analysis, signal

compression, and many other fields of engineering. Daubechies [17] and Meyer [68] are good introductory references to this subject. Wavelet-based methods have many exciting statistical properties [23]. Earlier papers on wavelets assume the regular design points, i.e, $x_i = \frac{i}{n}$ (usually $n = 2^k$ for some integer k) so that fast computation algorithms can be implemented. See [24] and references therein. For an overview of wavelets in statistics, see [87].

Antoniadis and Fan [1] discussed how to apply wavelet methods for function estimation with irregular design points using penalized least squares. Without loss of generality, assume that $m(x)$ is defined on [0, 1]. By moving nondyadic points to dyadic points, we assume $x_i = n_i/2^J$ for some n_i and some fine resolution J that is determined by users. To make this approximation errors negligible, we take J large enough such that $2^J \geq n$. Let $\boldsymbol{W}$ be a given wavelet transform at all dyadic points $\{i/2^J :, i = 1, \ldots, 2^J - 1\}$. Let $N = 2^J$ and $\boldsymbol{a}_i$ be the n_i-th column of $\boldsymbol{W}$, an $N \times N$ matrix, and $\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{m}$ be the wavelet transform of the function m at dyadic points. Then it is easy to see that $m(x_i) = \boldsymbol{a}_i^T \boldsymbol{\beta}$. This yields an overparameterized linear model

$$y_i = \boldsymbol{a}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{5.1}$$

which aims at reducing modeling biases. However, one cannot find a reasonable estimate of $\boldsymbol{\beta}$ by using the ordinary least squares method since $N \geq n$. Directly applying penalized least squares, we have

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{a}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{N} p_{\lambda_j}(|\beta_j|). \tag{5.2}$$

If the sampling points are equally spaced and $n = 2^J$, the corresponding design matrix of linear model (5.1) becomes a square orthogonal matrix. From the discussion in Section 3, minimizing the PLS (5.2) with the entropy penalty or the hard-thresholding penalty results in a hard-thresholding rule. With the L_1 penalty, the PLS estimator is the soft-thresholding rule. Assume that $p_\lambda(\cdot)$ is nonnegative, nondecreasing, and differentiable over $(0, \infty)$ and that function $-\beta - p'_\lambda(\beta)$ is strictly unimodal on $(0, \infty)$, $p'_\lambda(\cdot)$ is nonincreasing and $p'_\lambda(0+) > 0$. Then Antoniadis and Fan [1] showed that the resulting penalized least-squares estimator that minimizes (5.2) is adaptively minimax within a factor of logarithmic order as follows. Define the Besov space ball $B^r_{p,q}(C)$ to be

$$B^r_{p,q}(C) = \left\{ m \in L_p : \textstyle\sum_j (2^{j(r+1/2-1/p)} \|\boldsymbol{\theta}_{j\cdot}\|_p)^q < C \right\},$$

where $\boldsymbol{\theta}_{j\cdot}$ is the vector of wavelet coefficients of function m at the resolution level j. Here r indicates the degree of smoothness of the regression functions m.

Theorem 3. *Suppose that the regression function $m(\cdot)$ is in a Besov ball with $r + 1/2 - 1/p > 0$. Then the maximum risk of the PLS estimator $\hat{m}(\cdot)$ over $B^r_{p,q}(C)$ is of rate $O(n^{-2r/(2r+1)} \log(n))$ when the universal thresholding $\sqrt{2\log(n)/n}$ is used. It also achieves the rate of convergence $O\{n^{-2r/(2r+1)} \log(n)\}$ when the minimax thresholding $p_n/\sqrt{n}$ is used, where p_n is given in* [1].

We next consider multivariate regression function estimation. Suppose that $\{\boldsymbol{x}_i, y_i\}$ is a random sample from the regression model

$$y = m(\boldsymbol{x}) + \varepsilon,$$

where, without loss of generality, it is assumed that $\boldsymbol{x} \in [0, 1]^d$. Radial basis and neural-network are also popular for approximating multi-dimensional functions. In the literature of spline smoothing, it is typically assumed that the mean function $m(\boldsymbol{x})$ has a low-dimensional structure. For example,

$$m(\boldsymbol{x}) = \mu_0 + \sum_j m_j(x_j) + \sum_{k<l} m_{kl}(x_k, x_l).$$

For given knots, a set of spline basis functions can be constructed. The two most popular spline bases are the truncated power spline basis $1, x, x^2, x^3, (x - t_j)_+^3$, $(j = 1, \ldots, J)$, where t_j's are knots, and the B-spline basis (see [6] for definition). The B-spline basis is numerically more stable since the multiple correlation among the basis functions is smaller, but the power truncated spline basis has the advantage that deleting a basis function is the same as deleting a knot.

For a given set of 1-dimensional spline bases, we can further construct a multivariate spline basis using tensor products. Let $\{B_1, \ldots, B_J\}$ be a set of spline basis functions on $[0, 1]^d$. Approximate the regression function $m(\boldsymbol{x})$ by a linear combination of the basis functions, $\sum \beta_j B_j(\boldsymbol{x})$, say. To avoid a large approximation bias, we take a large J. This yields an overparameterized linear model, and the fitted curve of the least squares estimate is typically undersmooth. Smoothing spline suggested penalizing the roughness of the resulting estimate. This is equivalent to the penalized least squares with a quadratic penalty. In a series of work by Stone and his collaborators (see [76]), they advocate using regression splines and modifying traditional variable selection approaches to select useful spline subbases. Ruppert *et al.* [75] advocated penalized splines in statistical modeling, in which power truncated splines are used with the L_2 penalty. Another kind of penalized splines method proposed by [28] shares the same spirit of [75].

6. Some solutions to the challenges

In this section, we provide some solutions to problems raised in Section 2.

6.1. Computational biology. As discussed in Section 2.1, the first statistical challenge in computational biology is how to remove systematic biases due to experiment variations. Thus, let us first discuss the issue of normalization of cDNA-microarrays. Let Y_g be the log-ratio of the intensity of gene g of the treatment sample over that of the control sample. Denote by X_g the average of the log-intensities of gene g at the treatment and control samples. Set r_g and c_g be the row and column of the block

where the cDNA of gene g resides. Fan *et al.* [37] use the following model to estimate the intensity and block effect:

$$Y_g = \alpha_g + \beta_{r_g} + \gamma_{c_g} + f(X_g) + \varepsilon_g, \quad g = 1, \ldots, N \tag{6.1}$$

where α_g is the treatment effect on gene g, β_{r_g} and γ_{c_g} are block effects that are decomposed into the column and row effect, $f(X_g)$ represents the intensity effect and N is the total number of genes. Based on J arrays, an aim of microarray data analysis is to find genes g with α_g statistically significantly different from 0. However, before carrying multiple array comparisons, the block and treatment effects should first be estimated and removed. For this normalization purpose, parameters α_g are nuisance and high-dimensional (recall N is in the order of tens of thousands). On the other hand, the number of significantly expressed genes is relatively small, yielding the sparsity structure of α_g.

Model (6.1) is not identifiable. Fan *et al.* [37] use within-array replicates to infer about the block and treatment effects. Suppose that we have I replications for G genes, which could be a small fraction of N. For example, in [37], only 111 genes were repeated at random blocks ($G = 111$, $I = 2$), whereas in [63], all genes were repeated three times, i.e. $I = 3$ and $N = 3G$, though both have about $N \approx 20,000$ genes printed on an array. Using I replicated data on G genes, model (6.1) becomes

$$Y_{gi} = \alpha_g + \beta_{r_{gi}} + \gamma_{c_{gi}} + f(X_{gi}) + \varepsilon_{gi}, \quad g = 1, \ldots, G, \ i = 1, \ldots, I. \tag{6.2}$$

With estimated coefficients $\hat{\beta}$ and $\hat{\gamma}$ and the function $\hat{f}$, model (6.1) implies that the normalized data are $Y_g^* = Y_g - \hat{\beta}_{r_g} - \hat{\gamma}_{c_g} - \hat{f}(X_g)$ even for non-repeated genes.

Model (6.2) can be used to remove the intensity effect array by array, though the number of nuisance parameters is very large, a fraction of total sample size in (6.2). To improve the efficiency of estimation, Fan *et al.* [36] aggregate the information from other microarrays (total J arrays):

$$Y_{gij} = \alpha_g + \beta_{r_{gi},j} + \gamma_{c_{gi},j} + f_j(X_{gij}) + \varepsilon_{gi}, \quad j = 1, \ldots, J, \tag{6.3}$$

where the subscript j denotes the array effect.

The parameters in (6.2) can be estimated by the profile least-squares method using the Gauss–Seidel type of algorithm. See [36] for details. To state the results, let us write model (6.2) as

$$Y_{gi} = \alpha_g + \boldsymbol{Z}_{gi}^T \boldsymbol{\beta} + f(X_{gi}) + \varepsilon_{gi}, \tag{6.4}$$

by appropriately introducing the dummy variable $\boldsymbol{Z}$. Fan *et al.* [36] obtained the following results.

Theorem 4. *Under some regularity conditions, as $n = IG \to \infty$, the profile least-squares estimator of model* (6.4) *has*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N\Big(0, \frac{I}{I-1}\sigma^2 \Sigma^{-1}\Big),$$

where $\Sigma = E\{\mathrm{Var}(\mathbf{Z}|X)\}$ *and* $\sigma^2 = \mathrm{Var}(\varepsilon)$. *In addition,* $\hat{f}(x) - f(x) = O_P(n^{-2/5})$.

Theorem 5. *Under some regularity conditions, as* $n = IG \to \infty$, *when* X *and* $\mathbf{Z}$ *are independent, the profile least-squares estimator based on* (6.3) *possesses*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) \xrightarrow{\mathcal{D}} N\Big(0, \frac{I(J-1)+1}{J(I-1)}\,\sigma^2\Sigma^{-1}\Big).$$

The above theorems show that the block effect can be estimated at rate $O_P(n^{-1/2})$ and intensity effect f can be estimated at rate $O_P(n^{-2/5})$. This rate can be improved to $O_P(n^{-1/2} + N^{-2/5})$ when data in (6.1) are all used. The techniques have also been adapted for the normalization of Affymetrix arrays [30]. Once the arrays have been normalized, the problem becomes selecting significantly expressed genes using the normalized data

$$Y^*_{gj} = \alpha_g + \varepsilon_{gj}, \quad g = 1, \ldots, N,\ j = 1, \ldots, J, \tag{6.5}$$

where Y^*_{gj} is the normalized expression of gene g in array j. This is again a high-dimensional statistical inference problem. The issues of computing P-values and false discovery are given in Section 2.1.

Estimation of high-dimensional covariance matrices is critical in studying genetic networks. PLS and penalized likelihood can be used to estimate large scale covariance matrices effectively and parsimoniously [49], [59]. Let $\boldsymbol{w} = (W_1, \ldots, W_d)^T$ be a d-dimensional random vector with mean zero and covariance Σ. Using the modified Cholesky decomposition, we have $\boldsymbol{L}\Sigma\boldsymbol{L}^T = \boldsymbol{D}$, where $\boldsymbol{L}$ is a lower triangular matrix having ones on its diagonal and typical element $-\phi_{tj}$ in the (t, j)th position for $1 \le j < t \le d$, and $\boldsymbol{D} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_d^2)^T$ is a diagonal matrix. Denote $\boldsymbol{e} = \boldsymbol{L}\boldsymbol{w} = (e_1, \ldots, e_d)^T$. Since $\boldsymbol{D}$ is diagonal, $e_1, \ldots, e_d$ are uncorrelated. Thus, for $2 \le t \le d$

$$W_t = \sum_{j=1}^{t-1} \phi_{tj} W_j + e_t. \tag{6.6}$$

That is, the W_t is an autoregressive (AR) series, which gives an interpretation for elements of L and D, and allows us to use PLS for covariance selection. We first estimate σ_t^2 using the mean squared errors of model (6.6). Suppose that $\boldsymbol{w}_i$, $i = 1, \ldots, n$, is a random sample from $\boldsymbol{w}$. For $t = 2, \ldots, d$, covariance selection can be achieved by minimizing the following PLS functions:

$$\frac{1}{2n}\sum_{i=1}^{n}\Big(W_{it} - \sum_{j=1}^{t-1}\phi_{tj}W_{ij}\Big)^2 + \sum_{j=1}^{t-1} p_{\lambda_{t,j}}(|\phi_{tj}|). \tag{6.7}$$

This reduces the non-sparse elements in the lower triangle matrix $\boldsymbol{L}$. With estimated $\boldsymbol{L}$, the diagonal elements can be estimated by the sample variance of the components in $\hat{L}\boldsymbol{w}_i$. The approach can easily be adapted to estimate the sparse precision matrix Σ^{-1}. See [67] for a similar approach and a thorough study.

6.2. Health studies. Survival data analysis has been a very active research topic because survival data are frequently collected from reliability analysis, medical studies, and credit risks. In practice, many covariates are often available as potential risk factors. Selecting significant variables plays a crucial role in model building for survival data but is challenging due to the complicated data structure. Fan and Li [34] derived the nonconcave penalized partial likelihood for Cox's model and Cox's frailty model, the most commonly used semiparametric models in survival analysis. Cai *et al.* [9] proposed a penalized pseudo partial likelihood for marginal Cox's model with multivariate survival data and applied the proposed methodology for a subset data in the Framingham study, introduced in Section 2.2.

Let T, C and $\boldsymbol{x}$ be respectively the survival time, the censoring time and their associated covariates. Correspondingly, let $Z = \min\{T, C\}$ be the observed time and $\delta = I(T \leq C)$ be the censoring indicator. It is assumed that T and C are conditionally independent given $\boldsymbol{x}$, that the censoring mechanism is noninformative, and that the observed data $\{(\boldsymbol{x}_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ is an independently and identically distributed random sample from a certain population $(\boldsymbol{x}, Z, \delta)$. The Cox model assumes the conditional hazard function of T given $\boldsymbol{x}$

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta}), \tag{6.8}$$

where $h_0(t)$ is an unspecified baseline hazard function. Let $t_1^0 < \cdots < t_N^0$ denote the ordered observed failure times. Let (j) provide the label for the item failing at t_j^0 so that the covariates associated with the N failures are $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(N)}$. Let R_j denote the risk set right before the time $t_j^0 : R_j = \{i : Z_i \geq t_j^0\}$. Fan and Li [34] proposed the penalized partial likelihood

$$Q(\boldsymbol{\beta}) = \sum_{j=1}^{N}\Big[\boldsymbol{x}_{(j)}^T\boldsymbol{\beta} - \log\Big\{\sum_{i\in R_j}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\Big\}\Big] - n\sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|). \tag{6.9}$$

The penalized likelihood estimate of $\boldsymbol{\beta}$ is to maximize (6.9) with respect to $\boldsymbol{\beta}$.

For finite parameter settings, Fan and Li [34] showed that under certain regularity conditions, if both a_n and b_n tend to 0, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of the penalized partial likelihood function in (6.9) such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$. They further demonstrated the following oracle property.

Theorem 6. *Assume that the penalty function $p_{\lambda_n}(|\beta|)$ satisfies condition* (4.4). *If $\lambda_{n,j} \to 0$, $\sqrt{n}\lambda_{n,j} \to \infty$ and $a_n = O(n^{-1/2})$, then under some mild regularity conditions, with probability tending to* 1, *the root n consistent local maximizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ of $Q(\boldsymbol{\beta})$ defined in* (6.9) *must satisfy*

$$\hat{\boldsymbol{\beta}}_2 = \mathbf{0}, \quad \textit{and} \quad \sqrt{n}(I_1 + \Sigma)\big\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1 + \Sigma)^{-1}\boldsymbol{b}\big\} \xrightarrow{\mathcal{D}} N\big\{\mathbf{0}, I_1(\boldsymbol{\beta}_{10})\big\},$$

where I_1 is the first $s \times s$ submatrix of the Fisher information matrix $I(\boldsymbol{\beta}_0)$ of the partial likelihood.

Cai *et al.* [9] investigated the sampling properties of penalized partial likelihood estimate with a diverging number of predictors and clustered survival data. They showed that the oracle property is still valid for penalized partial likelihood estimation for the Cox marginal models with multivariate survival data.

6.3. Financial engineering and risk management. There are many outstanding challenges of dimensionality in diverse fields of financial engineering and risk management. To be concise, we focus only on the issue of covariance matrix estimation using a factor model.

Let Y_i be the excess return of the i-th asset over the risk-free asset. Let $f_1, \ldots, f_K$ be the factors that influence the returns of the market. For example, in the Fama–French 3-factor model, f_1, f_2 and f_3 are respectively the excessive returns of the market portfolio, which is the value-weighted return on all NYSE, AMEX and NASDAQ stocks over the one-month Treasury bill rate, a portfolio constructed based on the market capitalization, and a portfolio constructed based on the book-to-market ratio. Of course, constructing factors that influence the market itself is a high-dimensional model selection problem with massive amount of trading data. The K-factor model [15], [74] assumes

$$Y_i = b_{i1} f_1 + \cdots + b_{iK} f_K + \varepsilon_i, \quad i = 1, \ldots, d, \tag{6.10}$$

where $\{\varepsilon_i\}$ are idiosyncratic noises, uncorrelated with the factors, and d is the number of assets under consideration. This an extension of the famous Capital Asset Pricing Model derived by Sharpe and Lintner (See [10], [12]). Putting it into the matrix form, we have $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{\varepsilon}$ so that

$$\Sigma = \mathrm{Var}(\boldsymbol{B}\boldsymbol{f}) + \mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{B}\,\mathrm{Var}(\boldsymbol{f})\boldsymbol{B}^T + \Sigma_0, \tag{6.11}$$

where $\Sigma = \mathrm{Var}(\boldsymbol{y})$ and $\Sigma_0 = \mathrm{Var}(\boldsymbol{\varepsilon})$ is assumed to be diagonal.

Suppose that we have observed the returns of d stocks over n periods (e.g., 3 years daily data). Then, applying the least-squares estimate separately to each stock in (6.10), we obtain the estimates of coefficients in $\boldsymbol{B}$ and Σ_0. Now, estimating $\mathrm{Var}(\boldsymbol{f})$ by its sample variance, we obtain a substitution estimator $\widehat{\Sigma}$ using (6.11). On the other hand, we can also use the sample covariance matrix, denoted by $\widehat{\Sigma}_{\mathrm{sam}}$, as an estimator.

In the risk management or portfolio allocation, the number of stocks d can be comparable with the sample size n so it is better modeled as d_n. Fan *et al.* [32] investigated thoroughly when the estimate $\widehat{\Sigma}$ outperforms $\widehat{\Sigma}_{\mathrm{sam}}$ via both asymptotic and simulation studies. Let us quote some of their results.

Theorem 7. *Let $\lambda_k(\Sigma)$ be the k-th largest eigenvalue of Σ. Then, under some regularity conditions, we have*

$$\max_{1\le k\le d_n} |\lambda_k(\widehat{\Sigma}) - \lambda_k(\Sigma)| = o_P\{(\log n\, d_n^2/n)^{1/2}\} = \max_{1\le k\le d_n} |\lambda_k(\widehat{\Sigma}_{\mathrm{sam}}) - \lambda_k(\Sigma)|.$$

For a selected portfolio weight $\boldsymbol{\xi}_n$ with $\mathbf{1}^T\boldsymbol{\xi}_n = 1$, we have

$$\left|\boldsymbol{\xi}_n^T\widehat{\Sigma}\boldsymbol{\xi}_n - \boldsymbol{\xi}_n^T\Sigma\boldsymbol{\xi}_n\right| = o_P\{(\log n\, d_n^4/n)^{1/2}\} = \left|\boldsymbol{\xi}_n^T\widehat{\Sigma}_{\text{sam}}\boldsymbol{\xi}_n - \boldsymbol{\xi}_n^T\Sigma_n\boldsymbol{\xi}_n\right|.$$

If, in addition, the all elements in $\boldsymbol{\xi}_n$ are positive, then the latter rate can be replaced by $o_P\{(\log n\, d_n^2/n)^{1/2}\}$.

The above result shows that for risk management where the portfolio risk is $\boldsymbol{\xi}_n^T\Sigma\boldsymbol{\xi}_n$, no substantial gain can be realized by using the factor model. Indeed, there is no substantial gain for estimating the covariance matrix even if the factor model is correct. These have also convincingly been demonstrated in [32] using simulation studies. Fan *et al.* [32] also gives the order d_n under which the covariance matrix can be consistently estimated.

The substantial gain can be realized if Σ^{-1} is estimated. Hence, the factor model can be used to improve the construction of the optimal mean-variance portfolio, which involves the inverse of the covariance matrix. Let us quote one theorem of [32]. See other results therein for optimal portfolio allocation.

Theorem 8. *Under some regularity conditions, if $d_n = n^\alpha$, then for $0 \le \alpha < 2$,*

$$d_n^{-1}\text{tr}(\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I_{d_n})^2 = O_P(n^{-2\beta})$$

with $\beta = \min(1/2, 1-\alpha/2)$, whereas for $\alpha < 1$, $d_n^{-1}\text{tr}(\Sigma^{-1/2}\widehat{\Sigma}_{\text{sam}}\Sigma^{-1/2} - I_{d_n})^2 = O_P(d_n/n)$. In addition, under the Frobenius norm

$$d_n^2\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\|^2 = o(d_n^4\log n/n) = \|\widehat{\Sigma}_{\text{sam}}^{-1} - \Sigma^{-1}\|^2.$$

6.4. Machine learning and data mining. In machine learning, our goal is to build a model with the capability of good prediction of future observations. Prediction error depends on the loss function, which is also referred to as a divergence measure. Many loss functions are used in the literature. To address the versatility of loss functions, let us use the device introduced by [7]. For a concave function $q(\cdot)$, define a q-class of loss function $\ell(\cdot,\cdot)$ to be

$$\ell(y,\hat{m}) = q(\hat{m}) - q(y) - q'(\hat{m})(\hat{m} - y) \tag{6.12}$$

where $\hat{m} \equiv \hat{m}(\boldsymbol{x})$, an estimate of the regression function $m(\boldsymbol{x}) = E(y|\boldsymbol{x})$. Due to the concavity of q, $\ell(\cdot,\cdot)$ is non-negative.

Here are some notable examples of ℓ-loss constructed from the q-function. For binary classification, $y \in \{-1, 1\}$. Letting $q(m) = 0.5\min\{1-m, 1+m\}$ yields the misclassification loss, $\ell_1(y,\hat{m}) = I\{y \ne I(\hat{m} > 0)\}$. Furthermore, $\ell_2(y,\hat{m}) = [1 - y\text{sgn}(\hat{m})]_+$ is the hinge loss if $q(m) = \frac{1}{4}\min\{1-m, 1+m\}$. The function $q_3(m) = \sqrt{1-m^2}$ results in $\ell_3(y,\hat{m}) = \exp\{-0.5y\log\{(1+\hat{m})/(1-\hat{m})\}$, the exponential loss function in AdaBoost [40]. Taking $q(m) = cm - m^2$ for some constant c results in the quadratic loss $\ell_4(y,\hat{m}) = (y-\hat{m})^2$.

For a given loss function, we may extend the PLS to a penalized empirical risk minimization (4.5). The dimensionality d of the feature vectors can be much larger than n and hence the penalty is needed to select important feature vectors. See, for example, [4] for an important study in this direction.

We next make a connection between the penalized loss function and the popularly used support vector machines (SVMs), which have been successfully applied to various classification problems. In binary classification problems, the response y takes values either 1 or -1, the class labels. A classification rule $\delta(\boldsymbol{x})$ is a mapping from the feature vector $\boldsymbol{x}$ to $\{1, -1\}$. Under the 0–1 loss, the misclassification error of δ is $P\{y \neq \delta(\boldsymbol{x})\}$. The smallest classification error is the Bayes error achieved by $\text{argmin}_{c\in\{1,-1\}} P(y = c|\boldsymbol{x})$. Let $\{\boldsymbol{x}_i, y_i\}$, $i = 1, \ldots, n$ be a set of training data, where $\boldsymbol{x}_i$ is a vector with d features, and the output $y_i \in \{1, -1\}$ denotes the class label. The 2-norm SVM is to find a hyperplane $\boldsymbol{x}^T\boldsymbol{\beta}$, in which $x_{i1} = 1$ is an intercept and $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_{(2)}^T)^T$, that creates the biggest margin between the training points from class 1 and -1 [85]:

$$\max_{\boldsymbol{\beta}} \frac{1}{\|\boldsymbol{\beta}_{(2)}\|^2} \quad \text{subject to } y_i(\boldsymbol{\beta}^T\boldsymbol{x}_i) \geq 1 - \xi_i, \text{ for all } i, \xi_i \geq 0, \sum \xi_i \leq B, \quad (6.13)$$

where ξ_i are slack variables, and B is a pre-specified positive number that controls the overlap between the two classes. Due to its elegant margin interpretation and highly competitive performance in practice, the 2-norm SVM has become popular and has been applied for a number of classification problems. It is known that the linear SVM has an equivalent hinge loss formulation [48]

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} [1 - y_i(\boldsymbol{x}_i^T\boldsymbol{\beta})]_+ + \lambda \sum_{j=2}^{d} \beta_j^2.$$

Lin [62] shows that the SVM directly approximates the Bayes rule without estimating the conditional class probability because of the unique property of the hinge loss. As in the ridge regression, the L_2-penalty helps control the model complexity to prevent over-fitting.

Feature selection in the SVM has received increasing attention in the literature of machine learning. For example, the last issue of volume 3 (2002-2003) of *Journal of Machine Learning Research* is a special issue on feature selection and extraction for SVMs. We may consider a general penalized SVM

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} [1 - y_i(\boldsymbol{x}_i^T\boldsymbol{\beta})]_+ + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|).$$

The 1-norm (or LASSO-like) SVM has been used to accomplish the goal of automatic feature selection in the SVM ([89]). Friedman *et al.* [41] shows that the 1-norm SVM is preferred if the underlying true model is sparse, while the 2-norm SVM performs better if most of the predictors contribute to the response. With the SCAD penalty, the penalized SVM may improve the bias properties of the 1-norm SVM.

References

[1] Antoniadis, A., and Fan, J., Regularization of wavelets approximations (with discussions). *J. Amer. Statist. Assoc.* **96** (2001), 939–967.

[2] Benjamini, Y., and Yekutieli, D., The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** (2001), 1165–1188.

[3] Bickel, P. J., Minimax estimation of a normal mean subject to doing well at a point. In *Recent Advances in Statistics* (ed. by M. H. Rizvi, J. S. Rustagi, and D. Siegmund), Academic Press, New York 1983, 511–528.

[4] Bickel, P. J., and Levina, E., Some theory for Fisher's linear discriminant, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** (2004), 989–1010.

[5] Brown, P. O., and Botstein, D., Exploring the new world of the genome with microarrays. *Nat. Genet.* **21** (suppl. 1) (1999), 33–37.

[6] de Boor, C., *A Practical Guide to Splines*. Appl. Math. Sci. 27, Springer-Verlag, New York 1978.

[7] Bregman, L. M., A relaxation method of finding a common points of convex sets and its application to the solulation of problems in convex programming. *U.S.S.R. Comput. Math. Math. Phys.* **7** (1967), 620–631.

[8] Breiman, L., Better subset regression using the nonnegative garrote. *Technometrics* **37** (1995), 373–384.

[9] Cai, J., Fan, J., Li, R., and Zhou, H., Variable selection for multivariate failure time data. *Biometrika* **92** (2005), 303–316.

[10] Campbell, J. Y., Lo, A., and MacKinlay, A. C., *The Econometrics of Financial Markets*. Princeton University Press, Princeton, NJ, 1997.

[11] Craven, P., and Wahba, G., Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** (1979), 377–403.

[12] Cochrane, J. H., *Asset Pricing*. Princeton University Press, Princeton, NJ, 2001.

[13] Coifman, R. R., and Saito, N., Constructions of local orthonormal bases for classification and regression. *C. R. Acad. Sci. Paris Sér. I Math.* **319** (1994), 191–196.

[14] Coifman, R. R., and Wickerhauser, M. V., Entropy-based algorithms for best-basis selection. *IEEE Trans. Inform. Theory* **38** (1992), 713–718.

[15] Chamberlain, G., and Rothschild, M., Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** (1983), 1281–1304.

[16] Diaconis, P., and Sturmfels, B., Albegraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** (1998), 363–397.

[17] Daubechies, I., *Ten Lectures on Wavelets*. SIAM, Philadelphia 1992.

[18] Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. B* **39** (1977), 1–38

[19] Donoho, D. L., High-dimensional data analysis: the curses and blessings of dimensionality. *Aide-Memoire of the lecture in AMS conference "Math challenges of 21st Centrury* (2000). Available at http://www-stat.stanford.edu/~donoho/Lectures.

[20] Donoho, D. L., and Elad, E., Maximal sparsity representation via l_1 Minimization. *Proc. Nat. Aca. Sci.* **100** (2003), 2197–2202.

[21] Donoho, D. L., and Huo, X., Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** (2001), 2845–2862.

[22] Donoho, D. L., and Jin, J., Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** (2004), 962–994.

[23] Donoho, D. L., and Johnstone, I. M., Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** (1994), 425–455.

[24] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D., Wavelet shrinkage: asymptopia? *J. Royal Statist. Soc. B* **57** (1995), 301–369.

[25] Dudoit, S., Shaffer, J. P., and Boldrick, J. C., Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18** (2003), 71–103.

[26] Dudoit, Y., Yang, Y. H., Callow, M. J., and Speed, T. P., Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica* **12** (2002), 111–139.

[27] Efron, B., Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** (2004), 96–104.

[28] Eilers, P. H. C., and Marx, B. D., Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** (1996), 89–121.

[29] Fan, J., A selective overview of nonparametric methods in financial econometrics (with discussion). *Statist. Sci.* **20** (2005), 316–354.

[30] Fan, J., Chen, Y., Chan, H. M., Tam, P., and Ren, Y., Removing intensity effects and identifying significant genes for Affymetrix arrays in MIF-suppressed neuroblastoma cells. *Proc. Natl Acad. Sci. USA* **103** (2005), 17751–17756.

[31] Fan, J., and Gijbels, I. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London 1996.

[32] Fan, J., Fan, Y., and Lv, J., Large dimensional covariance matrix estimation using a factor model. Manuscript, 2005.

[33] Fan, J., and Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** (2001), 1348–1360.

[34] Fan, J., and Li, R., Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** (2002), 74–99.

[35] Fan, J., and Peng, H., Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** (2004), 928–961.

[36] Fan, J., Peng, H., and Huang, T., Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency (with discussion). *J. Amer. Statist. Assoc.* **100** (2005), 781–813.

[37] Fan, J., Tam, P., Vande Woude, G., and Ren, Y., Normalization and analysis of cDNA micro-arrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Natl Acad. Sci. USA* **101** (2004), 1135–1140.

[38] Foster, D. P., and George, E. I., The risk inflation criterion for multiple regression. *Ann. Statist.* **22** (1994), 1947–1975.

[39] Frank, I. E., and Friedman, J. H., A statistical view of some chemometrics regression tools. *Technometrics* **35** (1993), 109–148.

[40] Freund, Y., Schapire, R. E., A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Systems Sci.* **55** (1997), 119–139.

[41] Friedman, J., Hastie, T. Rosset, S., Tibshirani, R. and Zhu, J., Discussion of boosting papers. *Ann. Statist.* **32** (2004), 102–107.

[42] Greenshtein, E., Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1-constraint. *Ann. Statist.* **34** (5) (2006), to appear.

[43] Greenshtein, E., and Ritov, Y., Persistence in high-dimensional predictor selection and the virtue of overparametrization. *Bernoulli* **10** (2004), 971–988.

[44] Genovese, C., and Wasserman, L., A stochastic process approach to false discovery control. *Ann. Statist.* **32** (2004), 1035–1061.

[45] Hall, P., Edgeworth expansion for Student's t statistic under minimal moment conditions. *Ann. Probab.* **15** (1987), 920–931.

[46] Hall, P., Some contemporary problems in statistical sciences. *The Madrid Intelligencer* (2006), to appear.

[47] Hand, D. J., Mannila, H., and Smyth, P., *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001.

[48] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer Ser. Statist., Springer-Verlag, New York 2001.

[49] Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L., Covariance selection and estimation via penalised normal likelihood. *Biometrika* (2006), 85–98.

[50] Huang, J., Wang, D., and Zhang, C., A two-way semi-linear model for normalization and significant analysis of cDNA microarray data. *J. Amer. Statist. Assoc.* **100** (2005), 814–829.

[51] Hull, J. *Options, Futures, and Other Derivatives*. 5th ed., Prentice Hall, Upper Saddle Review, NJ, 2003.

[52] Hunter, D. R., and Li, R., Variable selection using MM algorithms. *Ann. Statist.* **33** (2005), 1617–1642.

[53] Jing, B. Y., Shao, Q.-M., and Wang, Q. Y., Self-normalized Cramér type large deviations for independent random variables. *Ann. Probab.* **31** (2003), 2167–2215.

[54] Johnstone, I. M., On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** (2001), 295–327.

[55] Korosok, M. R., and Ma, S., Marginal asymptotics for the "large p, small n" paradigm: With applications to micorarray data. *Ann. Statist.*, to appear.

[56] Lange, K., A gradient algorithm locally equivalent to the EM algorithm. *J. Royal Statist. Soc. B* **57** (1995), 425–437.

[57] Lehmann, E. L., Romano, J. P., and Shaffer, J. P., On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.* **33** (2005), 1084–1108.

[58] Li, H., and Gui, J., Gradient directed regularization for sparse Gaussian concentration graphs, with applications of inference of genetic networks. *Biostatistics* (2006), to appear.

[59] Li, R., Dziak, J., and Ma, H. Y., Nonconvex penalized least squares: characterizations, algorithm and application. Manuscript, 2006.

[60] Li, R., Root, T., and Shiffman, S., A local linear estimation procedure for functional multilevel modeling. In *Models for Intensive Longitudinal Data* (ed. by T. Walls and J. Schafer), Oxford University Press, New York 2006, 63–83.

[61] Lipschutz, R. J., Fodor, S., Gingeras, T., Lockhart, D. J., High density synthetic oloigonucleotide arrays. *Nat. Genet.* **21** (1999), 20–24.

[62] Lin, Y., Support vector machine and the Bayes rule in classification. *Data Mining and Knowledge Discovery* **6** (2002), 259–275.

[63] Ma, S, Kosorok, M. R., Huang, J., Xie, H., Manzella, L., and Soares, M. B., Robust semiparametric cDNA microarray normalization and significance analysis. *Biometrics* (2006), to appear.

[64] Mallat, S. G., Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Trans. Amer. Math. Soc.* **315** (1989a), 69–87.

[65] Mallat, S. G., A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.* **11** (1989b), 674–693.

[66] Meinshausen, N., Lasso with relaxation. Manuscript, 2005.

[67] Meinshausen, N., and Bühlmann, P., High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** (3) (2006), to appear.

[68] Meyer, Y., *Ondelettes*. Hermann, Paris 1990.

[69] Miller, A. J., *Subset Selection in Regression*. Chapman&Hall/CRC, London 2002.

[70] Moffatt, H. K., *Risk Management: Value at Risk and Beyond*. Cambridge University Press, New York 2003.

[71] Nikolova, M., Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math.* **61** (2000), 633–658.

[72] Pachter, L., and Sturmfels, B., Parametric inference for biological sequence analysis. *Proc. Natl. Acad. Sci. USA* **101** (2004), 16138–16143.

[73] Pistone, G., Riccomagno, E., and Wynn, H. P., *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman & Hall / CRC, London 2000.

[74] Ross, S., The arbitrage theory of capital asset pricing. *J. Economic Theory* **13** (1976), 341–360.

[75] Ruppert, D., Wand, M. P., Carroll, R. J., *Semiparametric regression*. Cambridge University Press, Cambridge 2003.

[76] Stone, C. J., Hansen, M., Kooperberg, C., and Truong, Y. K., Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25** (1997), 1371–1470.

[77] Storey J. D., Taylor J. E., and Siegmund D., Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Royal Statist. Soc. B* **66** (2004), 187–205.

[78] Sturmfels, B., and Sullivant, S., Toric ideals of phylogenetic invariants. *J. Comput. Biol.* **12** (2005), 204–228

[79] Svrakic, N. M., Nesic, O., Dasu, M. R. K., Herndon, D., and Perez-Polo, J. R., Statistical approach to DNA chip analysis. *Recent Prog. Hormone Res.* **58** (2003), 75–93.

[80] Tai, Y. C., and Speed, T. P., A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Statist.* **34** (5) (2006), to appear.

[81] Tibshirani, R., Regression shrinkage and selection via the LASSO. *J. Royal Statist. Soc. B* **58** (1996), 267–288.

[82] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G., Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.* **18** (2003), 104–117.

[83] Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., and Wong, W. H., Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29** (2001), 2549–2557.

[84] Tusher, V. G., Tibshirani, R., and Chu, G., Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98** (2001), 5116–5121.

[85] Vapnik, V., *The Nature of Statistical Learning*. Springer-Verlag, New York 1995.

[86] Walls, T., and Schater, J., *Models for Intensive Longitudinal Data*, Oxford University Press, New York 2006.

[87] Wang, Y., Selective review on wavelets in Statistics. In *Frontiers of Statistics* (ed. by J. Fan and H. Koul), Imperial College Press, 2006.

[88] Zhang, H. P., Yu, C. Y., and Singer, B., Cell and tumor classification using gene expression aata: Construction of forests. *Proc. Natl. Acad. Sci. USA* **100** (2003), 4168–4172.

[89] Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R., 1-norm support vector machines. *Neural Information Processing Systems* **16** (2003).

[90] Zou, H., The adaptive Lasso and its oracle properties. Manuscript, 2005.

Department of Operation Research and Financial Engineering, and Bendheim Center for Finance, Princeton University, Princeton, NJ 08544, U.S.A.
E-mail: jqfan@Princeton.edu

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111, U.S.A.
E-mail: rli@stat.psu.edu