Sparse High-Dimensional Models in Economics

Jianqing Fan,[1,2] Jinchi Lv,[3] and Lei Qi[1,2]

[1]Bendheim Center for Finance and [2]Department of
Operations Research and Financial Engineering,
Princeton University, Princeton, New Jersey 08544;
email: jqfan@princeton.edu, lqi@princeton.edu

[3]Information and Operations Management Department,
Marshall School of Business, University of Southern California,
Los Angeles, California 90089; email: jinchilv@marshall.usc.edu

Annu. Rev. Econ. 2011. 3:291–317

First published online as a Review in Advance on March 14, 2011

The *Annual Review of Economics* is online at economics.annualreviews.org

This article's doi:
10.1146/annurev-economics-061109-080451

JEL codes: C13, C51

1941-1383/11/0904-0291$20.00

Keywords

variable selection, independence screening, oracle properties, penalized likelihood, factor models, portfolio selection

Abstract

This article reviews the literature on sparse high-dimensional models and discusses some applications in economics and finance. Recent developments in theory, methods, and implementations in penalized least-squares and penalized likelihood methods are highlighted. These variable selection methods are effective in sparse high-dimensional modeling. The limits of dimensionality that regularization methods can handle, the role of penalty functions, and their statistical properties are detailed. Some recent advances in sparse ultra-high-dimensional modeling are also briefly discussed.

1. INTRODUCTION

1.1. High Dimensionality in Economics and Finance

High-dimensional models recently have gained considerable importance in several areas of economics. For example, the vector autoregressive (VAR) model (Sims 1980, Stock & Watson 2001) is a key technique to analyze the joint evolution of macroeconomic time series and can deliver a great deal of structural information. Because the number of parameters grows quadratically with the size of the model, standard VAR models usually include no more than 10 variables. However, econometricians may observe hundreds of data series. To enrich the model information set, Bernanke et al. (2005) proposed to augment standard VAR models with estimated factors (FAVAR) to measure the effects of monetary policy. Factor analysis also plays an important role in forecasting using large dimensional data sets (for reviews, see Stock & Watson 2006, Bai & Ng 2008).

Another example of high dimensionality is large panels of home-price data. To incorporate cross-sectional effects, one may consider that the price in one county depends on several other counties, most likely its geographic neighbors. Because such correlation is unknown, initially the regression equation may include about 1,000 counties in the United States, which makes direct ordinary least-squares (OLS) estimation impossible. One technique to reduce dimensionality is variable selection. Recently, statisticians and econometricians have developed algorithms to simultaneously select relevant variables and estimate parameters efficiently (see Fan & Lv 2010 for an overview). Variable selection techniques have also been widely used in financial portfolio construction, treatment-effects models, and credit-risk models.

Volatility matrix estimation is a high-dimensional problem in finance. To optimize the performance of a portfolio (Campbell et al. 1997, Cochrane 2005) or to manage the risk of a portfolio, asset managers need to estimate the covariance matrix or its inverse matrix of the returns of assets in the portfolio. Suppose that we have 500 stocks to be managed. There are 125,250 parameters in the covariance matrix. High dimensionality here poses challenges to the estimation of matrix parameters, as small element-wise estimation errors may result in huge errors matrix-wise. In the time domain, high-frequency financial data also provide both opportunities and challenges to high-dimensional modeling in economics and finance. On a finer timescale, the market microstructure noise may no longer be negligible.

1.2. High Dimensionality in Science and Technology

High-dimensional data have commonly emerged in other fields of sciences, engineering, and humanities, thanks to advances in computing technologies. Examples include marketing, e-commerce, and warehouse data in business; genetic, microarray, and proteomics data in genomics and heath sciences; and biomedical imaging, functional magnetic resonance imaging, tomography, signal processing, high-resolution imaging, and functional and longitudinal data. For instance, for drug sales data collected in many geographical regions, cross-sectional correlation makes the dimensionality increase quickly; the consideration of 1,000 neighborhoods requires 1 million parameters. In meteorology and earth sciences, temperatures and other attributes are recorded over time and in many regions. Large panel data over a short time horizon are frequently encountered. In biological sciences, one may want to classify diseases and predict clinical outcomes using microarray gene expression or proteomics data, in which tens of thousands of expression levels are potential covariates, but there are typically only tens or hundreds of subjects. Hundreds of thousands of single-nucleotide polymorphisms are potential predictors in genome-wide association studies. The dimensionality of the feature spaces grows rapidly when interactions of such predictors are considered. Large-scale data analysis is also a common feature of many problems in machine learning, such as text and document classification and computer vision (see Hastie et al. 2009 for more examples).

All the above examples exhibit various levels of high dimensionality. To be more precise, relatively high dimensionality refers to the asymptotic framework in which the dimensionality p is growing but is of a smaller order of the sample size n [i.e., $p = o(n)$], moderately high dimensionality refers to the asymptotic framework in which p grows proportionately to n (i.e., $p \sim cn$ for some $c > 0$), high dimensionality refers to the asymptotic framework in which p can grow polynomially with n [i.e., $p = O(n^{\alpha})$ for some $\alpha > 1$], and ultra-high dimensionality refers to the asymptotic framework in which p can grow nonpolynomially with n [i.e., $\log p = O(n^{\alpha})$ for some $\alpha > 0$], the so-called nonpolynomial (NP) dimensionality. The inference and prediction are based on high-dimensional feature space.

1.3. Challenges of High Dimensionality

High dimensionality poses numerous challenges to statistical theory, methods, and implementations in those problems. For example, in a linear regression model with noise variance σ^2, when the dimensionality p is comparable to or exceeds the sample size n, the OLS estimator is not well behaved. A regression model built on all regressors usually has a prediction error of order $(1 + p/n)^{1/2}\,\sigma$ when $p \leq n$ rather than $(1 + s/n)^{1/2}\,\sigma$ when there are only s intrinsic predictors. This reflects two well-known phenomena in high-dimensional modeling: spurious correlations and the noise accumulation. The spurious correlations among the predictors are an intrinsic difficulty of high-dimensional model selection. There can be high spurious correlation even for independent and identically distributed (i.i.d.) predictors when p is large compared with n (see, e.g., Fan & Lv 2008, Fan et al. 2011b). In fact, conventional intuition might no longer be accurate in high dimensions. Another example is the data-piling problems in high-dimensional space shown by Hall et al. (2005).

Noise accumulation is a common phenomenon in high-dimensional prediction. Although it is well known in regression problems, explicit theoretical quantification of the impact of dimensionality on classification was not well understood until the recent work of Fan & Fan (2008). These authors showed that for the independence classification rule, classification using all features has a misclassification rate determined by a quantity $C_p/\sqrt{p}$, which trades off between the dimensionality p and overall signal strength C_p, the distance between the centroids of two classes. Although C_p is nondecreasing with p, the accompanying penalty on dimensionality $\sqrt{p}$ can significantly deteriorate the performance. They showed indeed that classification using all features can be as bad as random guessing because of the noise accumulation in estimating the population centroids in high dimensions. Hall et al. (2008) considered a similar problem for distance-based classifiers and showed that the misclassification rate converges to zero when $C_p/\sqrt{p} \to \infty$.

As clearly demonstrated above, variable selection is fundamentally important in high-dimensional modeling. Bickel (2008) pointed out that the main goals of high-dimensional modeling are (*a*) to construct as effective a method as possible to predict future observations and (*b*) to gain insight into the relationship between features and response for scientific purposes, as well as hopefully to construct an improved prediction method.

Examples of the former goal include portfolio optimization and text and document classification, and the latter is important in many scientific endeavors such as genomic studies. In addition to noise accumulation, the inclusion of spurious predictors can prevent the appearance of some important predictors due to the spurious correlation between the predictors and response (see, e.g., Fan & Lv 2008, 2010). In such cases, those predictors help predict the noise, which can be a rather serious issue when we need to accurately characterize the contribution from each identified predictor to the response variable.

Sparse modeling has been widely used to deal with high dimensionality. The main assumption is that the p-dimensional parameter vector is sparse, with many components being exactly zero or negligibly small. Such an assumption is crucial in identifiability, especially for the relatively small sample size. Although the notion of sparsity gives rise to biased estimation in general, it has proved to be effective in many applications. In particular, variable selection can increase the estimation accuracy by effectively identifying important predictors and can improve the model interpretability.

The rest of the article is organized as follows. In Section 2, we survey some developments of the penalized least-squares (PLS) estimation and its applications to econometrics. Section 3 presents some further applications of sparse models in finance. We provide a review of more general likelihood–based sparse models in Section 4. In Section 5, we review some recent developments of sure screening methods for ultra-high-dimensional sparse inference. Conclusions are given in Section 6.

2. PENALIZED LEAST SQUARES

Assume that the collected data are of the form $(\mathbf{x}_i^T, y_i)_{i=1}^n$, in which y_i is the i-th observation of the response variable, and $\mathbf{x}_i$ is the associated p-dimensional predictors vector. The data are often assumed to be a random sample from the population $(\mathbf{x}^T, y)$, where, conditional on $\mathbf{x}$, the response variable y has a mean depending on $\boldsymbol{\beta}^T\mathbf{x}$ with $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$. In sparse high-dimensional modeling, we assume ideally that most parameters β_j are exactly zero, meaning that only a few of the predictors contribute to the response. The objective of variable selection is to identify all important predictors having nonzero regression coefficients and giving accurate estimates of those parameters.

2.1. Univariate Penalized Least Squares

We start with the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$ is an n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ is an $n \times p$ design matrix, and $\boldsymbol{\varepsilon}$ is an n-dimensional noise vector. Consider the specific case of a canonical linear model with a rescaled orthonormal design matrix, i.e., $\mathbf{X}^T\mathbf{X} = nI_p$. The PLS problem is

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^p} \left\{ \frac{1}{2n} \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \sum_{j=1}^p p_\lambda(|\boldsymbol{\beta}_j|) \right\}, \tag{2}$$

where $\|\cdot\|_2$ denotes the L_2 norm, and $p_\lambda(\cdot)$ is a penalty function indexed by the regularization parameter $\lambda \geq 0$. By regularizing the conventional least-squares estimation, we hope to simultaneously select important variables and estimate their regression coefficients with sparse estimates.

In the above canonical case of $\mathbf{X}^T\mathbf{X} = nI_p$, the PLS problem (Equation 2) can be transformed into the following component-wise minimization problem:

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^p} \left\{ \frac{1}{2n} \| \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \|_2^2 + \frac{1}{2} \| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \|_2^2 + \sum_{j=1}^{p} p_\lambda(| \beta_j |) \right\}, \tag{3}$$

where $\widehat{\boldsymbol{\beta}} = n^{-1}\mathbf{X}^T\mathbf{y}$ is the OLS estimator or, more generally, the marginal regression estimator. Thus we consider the univariate PLS problem

$$\hat{\theta}(z) = \underset{\theta\in\mathbf{R}}{\operatorname{argmin}} \left\{ \frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|) \right\}. \tag{4}$$

For any increasing penalty function $p_\lambda(\cdot)$, we have a corresponding shrinkage rule in the sense that $|\hat{\theta}(z)| \leq |z|$ and $\hat{\theta}(z) = \operatorname{sgn}(z)|\hat{\theta}(z)|$ (Antoniadis & Fan 2001). Antoniadis & Fan (2001) further showed that the PLS estimator $\hat{\theta}(z)$ has the following properties: (*a*) sparsity if $\min_{t\geq 0}\{t + p'_\lambda(t)\} > 0$, in which case the resulting estimator automatically sets small estimated coefficients to zero to accomplish variable selection and reduce model complexity; (*b*) approximate unbiasedness if $p'_\lambda(t) = 0$ for large t, in which case the resulting estimator is nearly unbiased, especially when the true coefficient β_j is large, to reduce model bias; (*c*) and continuity if and only if arg $\min_{t\geq 0}\{t + p'_\lambda(t)\} = 0$, in which case the resulting estimator is continuous in the data to reduce instability in model prediction (see, e.g., the discussion in Breiman 1996). Here $p_\lambda(t)$ is nondecreasing and continuously differentiable on $[0, \infty)$, the function $-t - p'_\lambda(t)$ is strictly unimodal on $(0, \infty)$, and $p'_\lambda(0)$ represents $p'_\lambda(0+)$. Generally speaking, the singularity of the penalty function at the origin, i.e., $p'_\lambda(0+) > 0$, is necessary to generate sparsity for variable selection, and its concavity is needed to reduce the estimation bias when the true parameter is nonzero. In addition, the continuity ensures the stability of the selected models.

There are many commonly used penalty functions such as the L_q penalties $p_\lambda(|\theta|) = \lambda\,|\theta|^q$ for $q > 0$ and $\lambda I(|\theta| \neq 0)$ for $q = 0$. The use of the L_0 penalty $p_\lambda(t) = \frac{\lambda^2}{2}I(t \neq 0)$ and L_1 penalty in Equation 4 gives the hard-thresholding estimator $\hat{\theta}_H(z) = zI(|z| > \lambda)$ and the soft-thresholding estimator $\hat{\theta}_S(z) = \operatorname{sgn}(z)(|z| - \lambda)_+$, respectively. It is easy to see that none of the L_q penalties simultaneously satisfies all three conditions given above. As such, Fan & Li (2001) introduced the smoothly clipped absolute deviation (SCAD) penalty, whose derivative is given by

$$p'_\lambda(t) = \lambda\left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\} \quad \text{for some } a > 2, \tag{5}$$

where $p_\lambda(0) = 0$, and $a = 3.7$ is often used. It satisfies the aforementioned three properties and, in particular, ameliorates the bias problems of convex penalty functions. A closely related minimax concave penalty (MCP) was proposed by Zhang (2010), whose derivative is given by $p'_\lambda(t) = (a\lambda - t)_+/a$. In particular, when $a = 1$, $p_\lambda(t) = \frac{1}{2}[\lambda^2 - (\lambda - t)_+^2]$ is referred to as the hard-thresholding penalty by Fan & Li (2001) and Antoniadis (1996), who showed that the solution of Equation 4 is also the hard-thresholding estimator $\hat{\theta}_H(z)$. Therefore, the MCP produces discontinuous solutions with model instability.

2.2. Multivariate Penalized Least Squares

Consider the multivariate PLS (Equation 2) with general design matrix $\mathbf{X}$. The goal is to estimate the true unknown sparse regression coefficients vector $\boldsymbol{\beta}_0 = (\beta_{0,1}, \cdots, \beta_{0,p})^T$ in the linear model (Equation 1), where the dimensionality p can be comparable to or even greatly exceed the sample size n. The L_0 regularization, which is used in many classical model selection methods, such as the AIC and BIC, has been shown to have nice sampling properties (see, e.g., Barron et al. 1999). However, these best-subset-selection methods require an exhaustive search over all submodels, which is prohibitive even in moderate dimensions. Such computational difficulty motivated various continuous relaxations. For example, the bridge regression (Frank & Friedman 1993) uses the L_q penalty, $0 < q \leq 2$. In particular, the use of the L_2 penalty is called the ridge regression. The nonnegative garrote was introduced by Breiman (1995) for variable selection and shrinkage estimation. The L_1 PLS method was termed Lasso by Tibshirani (1996), which is also collectively referred to as the L_1 penalization methods in other contexts. Other commonly used penalty functions include the SCAD and MCP (see Section 2.1). A family of concave penalties that bridge the L_0 and L_1 penalties was introduced by Lv & Fan (2009) for model selection and sparse recovery. A linear combination of L_1 and L_2 penalties was called an elastic net by Zou & Hastie (2005), with the L_2 component encouraging grouping of variables.

What kind of penalty functions are desirable for variable selection in sparse modeling? Some appealing properties of the regularized estimator were first outlined by Fan & Li (2001). They advocated penalty functions giving estimators with the three properties mentioned in Section 2.1. In particular, they considered penalty functions $p_\lambda(|\theta|)$ that are nondecreasing in $|\theta|$ and provided insights into these properties. As mentioned above, the SCAD penalty satisfies the above three properties, whereas Lasso (the L_1 penalty) suffers from the bias issue.

Much effort has been devoted to developing algorithms to solve the PLS problem (Equation 2). Fan & Li (2001) proposed an effective local quadratic approximation (LQA) algorithm. This translates the nonconvex minimization problem into a sequence of convex minimization problems. Specifically, for a given initial value $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_p^*)^T$, the penalty function p_λ is locally approximated by a quadratic function as

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^*|) + \frac{1}{2}\frac{p_\lambda'(|\beta_j^*|)}{|\beta_j^*|}[\beta_j^2 - (\beta_j^*)^2] \quad \text{for } \beta_j \approx \beta_j^*. \tag{6}$$

With quadratic approximation (Equation 6), the PLS problem (Equation 2) becomes a convex PLS problem with weighted L_2 penalty and admits a closed-form solution. To avoid numerical instability, it sets the estimated coefficient $\widehat{\beta}_j = 0$ if β_j^* is close to zero, that is, deleting the j-th covariate from the final model. One potential issue of LQA is that the value zero is an absorbing state in the sense that once a coefficient is set to zero, it remains zero in subsequent iterations. Recently, the local linear approximation (LLA)

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^*|) + p_\lambda'(|\beta_j^*|)(|\beta_j| - |\beta_j^*|) \quad \text{for } \beta_j \approx \beta_j^* \tag{7}$$

was introduced by Zou & Li (2008), after the least angle regression (LARS) algorithm (Efron et al. 2004) was proposed to efficiently compute Lasso. Both LLA and LQA are convex majorants of a concave penalty function $p_\lambda(\cdot)$ on $[0, \infty)$, but LLA is a better approximation as it is the minimum (tightest) convex majorant of the concave function on $[0, \infty)$. For both approximations, the resulting sequence of target values is always nonincreasing, which is a specific feature of minorization-maximization algorithms (Hunter & Li 2005). This can easily be seen by the following argument. If at the k-th iteration $L_k(\boldsymbol{\beta})$ is a convex majorant of the target function $Q(\boldsymbol{\beta})$ such that $L_k(\boldsymbol{\beta}_k) = Q(\boldsymbol{\beta}_k)$ and $\boldsymbol{\beta}_{k+1}$ minimizes $L_k\,(\boldsymbol{\beta})$, then

$$Q(\boldsymbol{\beta}_{k+1}) \leq L_k(\boldsymbol{\beta}_{k+1}) \leq L_k(\boldsymbol{\beta}_k) = Q(\boldsymbol{\beta}_k).$$

There are several powerful algorithms for Lasso. Osborne et al. (2000) casted it as quadratic programming. Efron et al. (2004) proposed a fast and efficient LARS algorithm for variable selection, which, with a simple modification, produces the entire Lasso solution path $\{\widehat{\boldsymbol{\beta}}(\lambda) : \lambda > 0\}$. This algorithm uses the fact that the Lasso solution path is piecewise linear in λ (see also Rosset & Zhu 2007 for more discussion). The LARS algorithm starts with a sufficiently large λ, which picks only one predictor that has the largest correlation with the response and decreases the λ value until the second variable is selected, at which time the selected variables have the same absolute correlation with the current working residual as the first one, and so on. By the Karush-Kuhn-Tucker (KKT) conditions, a sign constraint is needed to obtain the Lasso solution path. Zhang (2010) extended the idea of the LARS algorithm and introduced the PLUS algorithm to compute the PLS solution path when the penalty function $p_\lambda(\cdot)$ is a quadratic spline such as the SCAD and MCP.

With the linear approximation (Equation 7), the PLS problem (Equation 2) becomes the adaptively weighted Lasso:

$$\min_{\boldsymbol{\beta} \in \mathbf{R}^p} \left\{ \frac{1}{2n} \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \sum_{j=1}^{p} w_j |\, \beta_j \,| \right\}, \tag{8}$$

where the weights are $w_j = p'_\lambda(|\beta_j^*|)$. Thus algorithms for Lasso can easily be adapted to solve such problems. Different penalty functions give different weighting schemes, and, in particular, Lasso gives a constant weighting scheme. In this sense, the nonconvex PLS can be regarded as an iteratively reweighted Lasso. The weight function is chosen adaptively to reduce the biases due to penalization. The adaptive Lasso proposed by Zou (2006) uses the weighting scheme $w_j = |\beta_j^*|^{-\gamma}$ for some $\gamma > 0$. However, zero is an absorbing state. In contrast, penalty functions such as SCAD and MCP do not have such an undesirable property. In fact, if the initial estimate is zero, then $w_j = \lambda$, and the resulting estimate is the Lasso estimate. Fan & Li (2001), Zou (2006), and Zou & Li (2008) suggested the use of a consistent estimate such as the unpenalized estimator as the initial value, which implicitly assumes that $p \ll n$. When the dimensionality p exceeds n, it is not applicable. Fan & Lv (2008) recommended the use of $\beta_j^* = 0$, which is equivalent to using the Lasso estimate as the initial estimate. The SCAD does not stop here. It further reduces the bias problem of Lasso by assigning an adaptive weighting scheme. Other possible initial values include estimators given by the stepwise addition fit or componentwise regression. They put forward the recommendation that only a few iterations are needed.

Coordinate optimization has also been widely used to solve regularization problems. For example, for the PLS problem (Equation 2), Fu (1998), Daubechies et al. (2004), and Wu & Lange (2008) proposed a coordinate descent algorithm that iteratively optimizes Equation 2 one component at a time. Such an algorithm can also be applied to solve other problems such as in Meier et al. (2008) for the group Lasso (Yuan & Lin 2006), Friedman et al. (2008) for penalized precision matrix estimation, and Fan & Lv (2011) for penalized likelihood (see Section 4.1 for more details).

The tuning parameter λ in the regularization methods can be selected using, e.g., the multifold cross-validation method, generalized cross-validation method (Craven & Wahba 1978), and various information criteria. For example, Wang et al. (2007) proposed that one select λ by minimizing the generalized BIC:

$$\text{BIC}(\lambda) = \log\hat{\sigma}_\lambda^2 + \text{df}(\lambda)(\log n)/n,$$

where $\hat{\sigma}_\lambda^2$ is the mean-squared error, and df (λ) is the degrees of freedom of the regularized estimator. They showed that under fixed dimensions, with probability tending to one, the SCAD-BIC estimate possesses the oracle property. Wang et al. (2009) extended those results to the setting of a diverging number of parameters.

There have been many studies of the theoretical properties of PLS methods. We give only a sketch of the developments owing to space limitations. A more detailed account can be found in, e.g., Fan & Lv (2010). In a seminal paper, Fan & Li (2001) laid down the theoretical framework of the nonconcave penalized likelihood method and introduced the oracle property, which means that the estimator enjoys the same sparsity as the oracle estimator with asymptotic probability one and attains an information bound mimicking that of the oracle estimator. Here the oracle estimator $\widehat{\boldsymbol{\beta}}^O$ is referred to as the statistically infeasible estimator with knowledge of the true subset S ahead of time, namely, the component $\widehat{\boldsymbol{\beta}}_{S^c}^O = 0$, and $\widehat{\boldsymbol{\beta}}_S^O$ is the least-squares estimate using only the variables in S. They showed that for certain penalties, the resulting estimator possesses the oracle property in the classical framework of fixed dimensionality p. In particular, they showed that such conditions can be satisfied by SCAD, but not by Lasso, which suggests that the Lasso estimator generally does not have the oracle property. This has indeed been shown by Zou (2006) in the finite-parameter setting. Fan & Peng (2004) later extended the results of Fan & Li (2001) to the diverging dimensional setting of $p = o(n^{1/5})$ or $o(n^{1/3})$. Recently, extensive efforts have been made to study the properties with NP dimensionality.

Another L_1 regularization method is the Dantzig selector recently proposed by Candes & Tao (2007). It is defined as the solution to

$$\min \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \|n^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \le \lambda, \tag{9}$$

where $\lambda \ge 0$ is a regularization parameter. Under the uniform uncertainty principle on the design matrix $\mathbf{X}$, which is a condition on the bounded condition number for all submatrices of $\mathbf{X}$, they showed that, with large probability, the Dantzig selector $\widehat{\boldsymbol{\beta}}$ mimics the risk of the oracle estimator up to a logarithmic factor log p, specifically

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \le C\sqrt{(2\log p)/n}\left[\sigma^2 + \sum\nolimits_{j\in\text{supp}(\boldsymbol{\beta}_0)} \beta_{0,j}^2 \wedge \sigma^2\right]^{1/2}, \tag{10}$$

where $\boldsymbol{\beta}_0$ is the true regression coefficients vector, $C > 0$ is some constant, and $\lambda \sim \sqrt{(2\log p)/n}$. The uniform uncertainty principle condition can be stringent in high dimensions (see, e.g., Fan & Lv 2008, Fan & Song 2010 for more discussion). The oracle inequality (Equation 10) does not specify the sparsity of the estimate. Bickel et al. (2009) presented a simultaneous theoretical comparison of the Lasso estimator and the Dantzig selector in a general high-dimensional nonparametric regression model:

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}, \tag{11}$$

where $\mathbf{f} = [f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n)]^T$, with f an unknown function of p variates, and $\mathbf{y}$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$, and $\boldsymbol{\varepsilon}$ are the same as in Equation 1. Under a sparsity scenario, Bickel et al. (2009) derived parallel oracle inequalities for the prediction risk for both methods and established the asymptotic equivalence of the Lasso estimator and the Dantzig selector. They also considered the specific case of the linear model (Equation 1), i.e., Equation 11 with true regression function $\mathbf{f} = \mathbf{X}\boldsymbol{\beta}_0$, and gave bounds under the L_q estimation loss for $1 \leq q \leq 2$.

For variable selection, we are concerned with the model selection consistency of regularization methods in addition to the estimation consistency under some loss. Zhao & Yu (2006) characterized the model selection consistency of Lasso by studying a stronger but technically more convenient property of sign consistency: $P\left[\operatorname{sgn}(\widehat{\boldsymbol{\beta}}) = \operatorname{sgn}(\boldsymbol{\beta}_0)\right] \to 1$ as $n \to \infty$. They showed that the weak irrepresentable condition

$$\| \mathbf{X}_2^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \operatorname{sgn}(\boldsymbol{\beta}_1) \|_\infty < 1, \tag{12}$$

where we assume covariates have been standardized, is a necessary condition for the sign consistency of Lasso, and the strong irrepresentable condition stating that the left-hand side of Equation 12 is uniformly bounded by a constant $0 < C < 1$ is a sufficient condition for the sign consistency of Lasso, where $\boldsymbol{\beta}_1$ is the subvector of $\boldsymbol{\beta}_0$ on its support $\operatorname{supp}(\boldsymbol{\beta}_0)$, and $\mathbf{X}_1$ and $\mathbf{X}_2$ denote the submatrices of the $n \times p$ design matrix $\mathbf{X}$ formed by columns in $\operatorname{supp}(\boldsymbol{\beta}_0)$ and its complement, respectively. However, the irrepresentable condition is restrictive in high dimensions. It requires that the L_1 norm of all regression coefficients of all variables in $\mathbf{X}_2$ regressed on $\mathbf{X}_1$ be bounded by 1 (see, e.g., Lv & Fan 2009 and Fan & Song 2010 for a simple illustrative example). This demonstrates that in high dimensions, the Lasso estimator can easily select an inconsistent model, which explains why Lasso tends to include many false positive variables in the selected model. The latter is also related to the bias problem in Lasso, which requires a small penalization λ, whereas the sparsity requires choosing a large λ.

Three questions of interest naturally arise for regularization methods. What limits of the dimensionality can PLS methods handle? What is the role of penalty functions? What are the statistical properties of PLS methods when the penalty function p_λ is no longer convex? As mentioned above, Fan & Li (2001) and Fan & Peng (2004) provided answers via the framework of an oracle property for fixed or relatively slowly growing dimensionality p. Recently, Lv & Fan (2009) introduced the weak oracle property, which means that the estimator enjoys the same sparsity as the oracle estimator with asymptotic probability one and has consistency, and established regularity conditions under which the PLS estimator given by folded-concave penalties has a nonasymptotic weak oracle property when the dimensionality p can grow nonpolynomially with the sample size n. They considered a wide class of folded-concave penalties including SCAD, and the L_1 penalty at its boundary. In particular, their results show that concave penalties can be more advantageous than convex penalties in high-dimensional variable selection. Later, Fan & Lv (2011) extended the results of Lv & Fan (2009) to folded-concave penalized likelihood in generalized linear models with ultra-high dimensionality. Fan & Lv (2009) also characterized the global optimality of the regularized estimator [see, e.g., Kim et al. (2008) and Kim & Kwon (2009), who showed that the SCAD estimator equals the oracle estimator with probability tending to one]. Other work on PLS methods includes Donoho et al. (2006), Meinshausen & Bühlmann (2006), Wainwright (2006), Huang et al. (2008), Koltchinskii (2008), Belloni & Chernozhukov (2009), and Zhang (2010). When the error distribution is heavy tailed, one may want to use a loss other than the quadratic one to improve the estimation efficiency. For example, quantile regression techniques have been widely used in econometrics problems (see Belloni & Chernozhukov 2011 and Bradic et al. 2011 for the properties of quantile regression with the L_1 penalty in high dimensions).

2.3. Multivariate Time Series

High dimensionality arises easily from VAR models. A p-dimensional time series with d lags gives dp^2 autoregressive parameters. As an illustration, we focus on an application of PLS to home-price estimation and forecasting. Analysis on the housing market based on state-level panel data can capture state-specific dynamics and variations. Calomiris et al. (2008) performed panel VAR regression to reveal the strong effect of foreclosure on home prices. Stock & Watson (2010) used a dynamic factor model with stochastic volatility to examine the link between housing construction and the decline in macro volatility since the mid-1980s. Rapach & Strass (2007) considered combinations of individual VAR forecasts, with each equation consisting of only one macroeconomic variable, in forecasting home-price growth in several states. Ng & Moench (2011) performed a hierarchical factor model consisting of regions and states to draw a linkage between housing and consumption.

If the primary focus is to forecast home price on local levels, however, factor models have difficulties in explicitly capturing the cross-sectional correlation among local levels. For example, at the state level, the home price in Nevada may have statistical correlation with California and Arizona; at the county or zip-code level, prices in suburbs may respond sensitively to price changes in the city center, but the response in the reverse direction might be insignificant. For this reason, we work on regressions that include all lag variables of all county-level home-price appreciations (HPAs) as regressors.

The addition of neighborhood variables to the regression equation results in a high-dimensional problem, and standard regression techniques often fail to estimate. If we let y_t^i be the HPA in county i, an s-period-ahead county-level forecast model is written as

$$y_{t+s}^i = \sum_{j=1}^{p} b_{ij} y_t^j + \mathbf{X}_t \boldsymbol{\beta}_i + \varepsilon_{t+s}^i, \quad i = 1, \cdots, p,$$

where $\mathbf{X}_t$ are observable factors, y_t^j are the HPAs of other counties, and b_{ij} and $\boldsymbol{\beta}_i$ are regression coefficients. On the one hand, because p is large (several hundred counties in the United States), such model cannot be estimated by OLS simply because there is not a long-enough time series (for 10-year monthly data, $n = 120$). On the other hand, only a handful of county-level lag HPAs should be useful for prediction conditioning on national factors, and the regression result should be sparse. PLS can be used to estimate b_{ij} and obtain sparse solutions (and hence neighborhood selection) at the same time. A simple solution is to minimize for each given target region i the following object:

$$\min_{\{b_{ij}, j=1,\cdots,N, \boldsymbol{\beta}_i\}} \sum_{t=1}^{T-s} \Big(y_{t+s}^i - \mathbf{X}_t^i \boldsymbol{\beta}_i - \sum_{j=1}^{N} b_{ij} y_t^j\Big)^2 + \lambda \sum_{j=1}^{N} w_{ij} p_\lambda(|b_{ij}|),$$

where the weights w_{ij} are chosen according to the geographical distances between counties i and j, and $p(\lambda)$ is the SCAD penalty. Counties far away from the target county receive a larger penalty. This choice of penalty reflects the intuition that if two counties are far apart, their correlation is more naturally explained by national factors, which are already included in **X**.

To fit the model, we use monthly HPA data for the 352 largest counties in the United States in terms of monthly repeated sales from January 2000 to December 2009. The measurements of HPAs are more reliable for those counties. As an illustration, the market factor is chosen to be the national HPA. Therefore, it is a reduced-form forecasting model of county-level HPA, taking the national HPA forecast as an input. **Figure 1** (see color insert) shows how cross-county correlation is captured by a sparse VAR model. **Figure 1*a*** shows heavy spatial correlation. Whereas the spatial correlation is reduced significantly in the residuals in **Figure 1*b***, the national factor cannot fully capture the local dependence. The residual correlations in **Figure 1*c*** look essentially like white noise, indicating that the national HPA along with the neighborhood selection capture the cross-dependence of regional HPAs. The model achieves both parsimony and in-sample estimation accuracy.

Sparse cross-sectional modeling translates into more forecasting power. This is illustrated by an out-of-sample test. Periods 2000.1–2005.12 are now used as a training sample, and 2006.1–2009.12 are testing periods. We use the discounted aggregated squared error as a measure of overall performance for each county:

$$\text{Forecast Error}_i = \sum_{s=1}^{\tau} \rho^s (\hat{y}^i_{T+s} - y^i_{T+s})^2, \quad \rho = 0.95.$$

The results show that, over 352 counties, the sparse VAR method with neighborhood information performs on average 30% better in terms of prediction error than the model without neighborhood information. Details of improvements can be seen in **Figure 2*a*** (see color insert). **Figure 2*b*** compares backtest forecasts using OLS with only the national HPA and PLS with additional neighborhood information for the largest counties with the historical HPAs.

2.4. Benchmark of Prediction Errors and Spurious Correlation

How good is a prediction method? The ideal prediction is to use the true model, and the residual variance σ^2 provides a benchmark measure of prediction errors. However, in high-dimensional econometrics problems, the spurious correlations among realized random variables are high, and some predictors can easily be selected to predict the realized noise vector. Therefore, the residual variance can substantially underestimate σ^2, as the realized noises can be predicted well by these predictors. Specifically, let $\widehat{\mathcal{S}}$ and $\mathcal{S}_0$ be the sets of selected and true variables, respectively. Fan et al. (2011b) argued that the variables in $\widehat{\mathcal{S}} \cap \mathcal{S}_0^c$ are used to predict the realized noise when $\widehat{\mathcal{S}} \supset \mathcal{S}_0$. As a result, in the linear model (Equation 1), the residual sum of squares substantially underestimates the error variance. Thus it is important and necessary to screen variables that are not truly related to the response and reduce their influence.

One effective way of handling spurious correlations and their influence is to use the refitted cross-validation (RCV) method proposed by Fan et al. (2011b). The sample is randomly split into two equal halves, and a variable selection procedure is applied to both subsamples. For each subsample, a variance estimate is obtained by regressing the response on the set of predictors selected based on the other subsample. The average of those two estimates gives a new variance estimate. Specifically, let $\hat{S}_1$ be the selected variables using the first half of the data, and then refit the regression coefficients of variables in $\hat{S}_1$ using the second half of the data and compute the resulting residual variance $\hat{\sigma}_1^2$. A similar estimate $\hat{\sigma}_2^2$ can be obtained, and the final estimate is simply $\hat{\sigma}^2 = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2$ or its weighted version using the degrees of freedom in the computation of two residual variances. Fan et al. (2011b) showed that when the variable selection procedure has the sure screening property, $S_0 \subset \hat{S}_1 \cap \hat{S}_2$ (see Section 5.1 for more discussion), the resulting estimator can perform as well as the oracle variance estimator, which knows S_0 in advance. They also explain the robustness of the RCV method to the sure screening property. The idea of RCV can be applied to variance estimation and variable selection in more general sparse high-dimensional models.

As an illustration, following Fan et al. (2011b), we consider the benchmark one-step forecasting errors σ^2 in San Francisco and Los Angeles, using HPA data from January 1998 to December 2005 (96 months). **Figure 3** shows the estimates as a function of the selected model size s. Clearly, the naive estimates of directly computing residual variances decrease steadily with the selected model size s due to spurious correlation, whereas the RCV method gives reasonably stable estimates for a range of selected models. The benchmark for both regions is approximately 0.53%, whereas the standard deviations of month-over-month variations of HPAs are 1.08% and 1.69% in the San Francisco and Los Angeles areas, respectively. To see how the PLS method works in comparison with the benchmark, we compute rolling one-step prediction errors over 12 months in 2006. The prediction errors are 0.67% and 0.86% for the San Francisco and Los Angeles areas, respectively. They are clearly larger than the benchmark, as expected, but are considerably smaller than the standard deviations, which use no variables to forecast.

3. SPARSE MODELS IN FINANCE

3.1. Estimation of High-Dimensional Volatility Matrices

Covariance matrix estimation is fundamental in many areas of multivariate analysis. For example, an estimate of the covariance matrix $\mathbf{\Sigma}$ is required in financial risk assessment and longitudinal studies, whereas an inverse of the covariance matrix, called the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, is needed in optimal portfolio selection, linear discriminant analysis, and graphical models. In particular, estimating a $p \times p$ covariance or precision matrix is challenging when the number of variables p is large. The sample covariance matrix is unbiased and is invertible when p is no larger than n. It is a natural candidate when p is small, but it no longer performs well for moderate or large dimensionality (Johnstone 2001). Additional challenges arise when estimating the precision matrix when $n < p$.

To deal with high dimensionality, researchers have taken two main directions in the literature. One is to remedy the sample covariance matrix estimator using approaches such as the eigen method and shrinkage (see, e.g., Stein 1975, Ledoit & Wolf 2004). The other one is to impose some structure such as the sparsity, factor model, or autoregressive model on the data to reduce the dimensionality (see, for example, Wu & Pourahmadi 2003, Huang et al. 2006, Yuan & Lin 2007, Bai & Ng 2008, Bickel & Levina 2008a,b, Fan et al. 2008, Levina et al. 2008, Rothman et al. 2008, Lam & Fan 2009, Cai et al. 2010).

The PLS and penalized likelihood method (see Section 4.1) can also be used to estimate large-scale covariances effectively and parsimoniously (see, e.g., Huang et al. 2006). Assuming that the covariance matrix has some sparse parameterization, the idea of variable selection can be used to select nonzero matrix parameters. Lam & Fan (2009) gave a

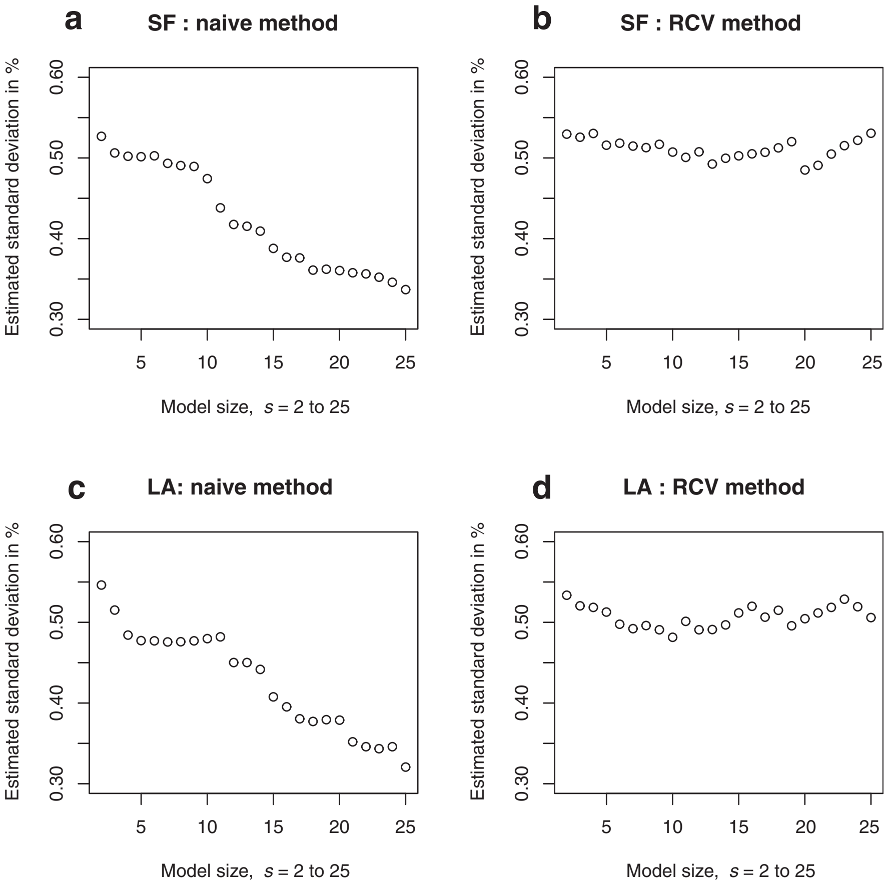

Figure 3

Estimated standard deviation as a function of selected model size s in San Francisco (*top panels*) and Los Angeles (*bottom panels*) using the naive (*left column*) and refitted cross-validation (RCV) (*right column*) methods.

comprehensive treatment on the sparse covariance matrix, sparse precision matrix, and sparse Cholesky decomposition.

The negative Gaussian pseudo-likelihood is

$$\mathrm{tr}(\mathbf{S}\boldsymbol{\Omega}) - \log|\boldsymbol{\Omega}|, \tag{13}$$

where $\mathbf{S}$ is the sample covariance matrix. Therefore, the sparsity of the precision matrix can be explored by the penalized pseudo-likelihood

$$\mathrm{tr}(\mathbf{S}\boldsymbol{\Omega}) - \log|\boldsymbol{\Omega}| + \sum_{i \neq j} p_\lambda(|\omega_{i,j}|), \tag{14}$$

penalizing only the off-diagonal elements $\omega_{i,j}$ of the precision matrix $\boldsymbol{\Omega}$, as the diagonal elements are nonsparse. This allows us to estimate the precision matrix even when $p > n$. Similarly, the sparsity of the covariance matrix can be explored by minimizing

$$\mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) + \log|\boldsymbol{\Sigma}| + \sum_{i \neq j} p_\lambda(|\sigma_{i,j}|), \tag{15}$$

again penalizing only the off-diagonal elements $\sigma_{i,j}$ of the covariance matrix $\mathbf{\Sigma}$. Various algorithms have been developed to optimize Equations 14 and 15 (see, for example, Friedman et al. 2008, Fan et al. 2009a). A comprehensive theoretical study of the properties of these approaches is provided by Bickel & Levina (2008a) and Lam & Fan (2009). They showed that the rates of convergence for these problems under the Frobenius norm are of order $(s \log p/n)^{1/2}$, where s is the number of nonzero elements. This demonstrates that the impact of dimensionality p enters through a logarithmic factor. They also studied the sparsistency of the estimates, which is a property that all zero parameters are estimated as zero with asymptotic probability one, and showed that the L_1 penalty is restrictive in that the number of nonzero off-diagonal elements s' is equal to $O(p)$, whereas for fold-concave penalties such as SCAD and the hard-thresholding penalty, there is no such restriction.

Sparse Cholesky decomposition can be explored similarly. Let $\mathbf{w} = (W_1, \cdots, W_p)^T$ be a p-dimensional random vector with mean zero and covariance matrix $\mathbf{\Sigma}$. Using the modified Cholesky decomposition, we have $\mathbf{L\Sigma L}^T = \mathbf{D}$, where $\mathbf{L}$ is a lower triangular matrix having diagonal elements 1 and off-diagonal elements $-\phi_{t,j}$ in the (t, j) entry for $1 \leq j < t \leq p$, and $\mathbf{D} = \mathrm{diag}\{\sigma_1^2, \cdots, \sigma_p^2\}$ is a diagonal matrix. Denote $\mathbf{e} = \mathbf{Lw} = (e_1, \cdots, e_p)^T$. Because $\mathbf{D}$ is diagonal, $e_1, \cdots, e_p$ are uncorrelated. Thus, for each $2 \leq t \leq p$,

$$W_t = \sum_{j=1}^{t-1} \phi_{tj} W_j + e_t. \tag{16}$$

This shows that W_t is an autoregressive series, which gives an interpretation for elements of matrices $\mathbf{L}$ and $\mathbf{D}$ and enables us to use the PLS for covariance selection. Suppose that $\mathbf{w}_i = (W_{i1}, \cdots, W_{ip})^T$, $i = 1, \cdots, n$, is a random sample from $\mathbf{w}$. For $t = 2, \cdots, p$, covariance selection can be accomplished by solving the following PLS problem:

$$\min_{\phi_{tj}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (W_{it} - \sum_{j=1}^{t-1} \phi_{tj} W_{ij})^2 + \sum_{j=1}^{t-1} p_{\lambda_t}(|\phi_{tj}|) \right\}. \tag{17}$$

With estimated sparse $\mathbf{L}$, the diagonal elements can be estimated by the sample variance of the components of $\widehat{\mathbf{L}}\mathbf{w}_i$. Hence the sparsity in Equation 16 is explored.

When the covariance matrix $\mathbf{\Sigma}$ admits sparsity structure, other simple methods can be exploited. Bickel & Levina (2008b) and El Karoui (2008) considered the approach of directly applying entrywise hard thresholding on the sample covariance matrix. The thresholded estimator has been shown to be consistent under the operator norm, in which the former considered the case of $(\log p) / n \to 0$ and the latter considered the case of $p \sim cn$. The optimal rates of convergence of such covariance matrix estimation were derived by Cai et al. (2010). Bickel & Levina (2008a) studied the methods of banding the sample covariance matrix and banding the inverse of the covariance via the Cholesky decomposition of the inverse for the estimation of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$, respectively. These estimates have been shown to be consistent under the operator norm for $(\log p)^2 / n \to 0$, and explicit rates of convergence were obtained. Meinshausen & Bühlmann (2006) proved that Lasso is consistent in neighborhood selection in high-dimensional Gaussian graphical models, in which the sparsity in the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ amounts to conditional independence between the variables.

3.2. Portfolio Selection

Markowitz (1952, 1959) laid down the seminal framework of mean-variance analysis. In practice, a simple implementation is to construct the mean-variance efficient portfolio using sample means and sample covariance matrix. However, owing to the accumulation of estimation errors, the theoretical optimal allocation vector can be very different from the estimated one, especially when the number of assets under consideration is large. As a result, such portfolios often suffer poor out-of-sample performance, although they are optimal in sample. Recently, a number of works have focused on improving the performance of the Markowitz portfolio using various regularization and stabilization techniques. Jagannathan & Ma (2003) considered the minimum variance portfolio with no short-sale constraints. They showed that such a constrained minimum variance portfolio outperforms the global minimum variance portfolio in practice when unknown quantities are estimated. To bridge the no short-sale constraints, on one extreme, and no constraints on short sales, on the other extreme, Fan et al. (2011c) introduced a gross-exposure parameter c and examined the impact of c on the performance of the minimum portfolio. They showed that with the gross-exposure constraint, the empirically selected optimal portfolios based on estimated covariance matrices have a similar performance to the theoretical optimal portfolios, and there is little error accumulation effect from the estimation of vast covariance matrices when c is modest.

The portfolio optimization problem is

$$\max_{\mathbf{w}} \quad \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}, \quad s.t.\mathbf{w}^T\mathbf{1} = 1, \quad \|\mathbf{w}\|_1 \le c, \quad \mathbf{A}\mathbf{w} = \mathbf{a},$$

where $\boldsymbol{\Sigma}$ is the true covariance matrix. The side constraints $\mathbf{A}\mathbf{w} = \mathbf{a}$ can be on the expected returns of the portfolio, as in the Markowitz (1952, 1959) formulation. They can also be the constraints on the allocations on sectors or industries, or the constraints on the risk exposures to certain known risk factors. They make the portfolio even more stable. Therefore, they can be removed from theoretical studies. Let $R(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}$ and $R_n(\mathbf{w}) = \mathbf{w}^T\widehat{\boldsymbol{\Sigma}}\mathbf{w}$ be the theoretical and empirical portfolio risk with allocation $\mathbf{w}$, where $\widehat{\boldsymbol{\Sigma}}$ is an estimator of covariance matrix with sample size n. Let

$$\mathbf{w}_{opt} = \mathrm{argmin}_{\mathbf{w}^T\mathbf{1}=1, \|\mathbf{w}\|_{1\le c}} R(\mathbf{w}) \text{ and } \widehat{\mathbf{w}}_{opt} = \mathrm{argmin}_{\mathbf{w}^T\mathbf{1}=1, \|\mathbf{w}\|_{1\le c}} R_n(\mathbf{w}).$$

The following theorem shows that the theoretical minimum risk $R(\mathbf{w}_{opt})$ (also called the oracle risk), the actual risk $R(\widehat{\mathbf{w}}_{opt})$, and empirical risk $R_n(\widehat{\mathbf{w}}_{opt})$ are approximately the same for a moderate c and a reasonable covariance matrix estimator.

Theorem 1: Let $a_n = \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty$. Then, we have

$$|R(\widehat{\mathbf{w}}_{opt}) - R(\mathbf{w}_{opt})| \le 2a_n c^2,$$

$$|R(\widehat{\mathbf{w}}_{opt}) - R_n(\widehat{\mathbf{w}}_{opt})| \le a_n c^2,$$

$$|R(\mathbf{w}_{opt}) - R_n(\widehat{\mathbf{w}}_{opt})| \le a_n c^2.$$

Theorem 1, due to Fan et al. (2011c), gives the upper bounds on the approximation errors of risks. The following result further controls a_n.

Theorem 2: Let σ_{ij} and $\hat{\sigma}_{ij}$ be the (i, j)-th element of the matrices $\mathbf{\Sigma}$ and $\widehat{\mathbf{\Sigma}}$, respectively. If for a sufficiently large x,

$$\max_{i,j} P\{\sqrt{n}|\sigma_{ij} - \hat{\sigma}_{ij}| > x\} < \exp(-Cx^{1/a})$$

for two positive constants a and C, then

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_\infty = O_P\left[\frac{(\log p)^a}{\sqrt{n}}\right]. \tag{18}$$

Fan et al. (2011c) gave further elementary conditions under which Theorem 2 holds. The connection between portfolio minimization with the gross-exposure constraint and the L_1 constrained regression problem enables fast statistical algorithms. The paper uses least-angle regression, or the LARS-Lasso algorithm, to solve for the optimal portfolio under various gross exposure limits c. When $c = 1$, it is equivalent to the no-short-sale constraint; as c increases, the constraint becomes less stringent, and it becomes the global minimum variance problem when $c = \infty$. Empirical studies find that when $c \sim 2$, the portfolio achieves the best out-of-sample performance in terms of variance and Sharpe ratio, when low-frequency daily data are used.

The gross-exposure constraint yields sparse portfolio selection. This feature is also noted by Brodie et al. (2009). DeMiguel et al. (2009) considered other norms to constrain the portfolio.

3.3. Factor Models

Section 3.1 above discusses large covariance matrix estimation via penalization methods. We now introduce an approach that uses a factor model, which provides another effective way of sparse modeling. Consider the multifactor model

$$Y_i = b_{i1}f_1 + \cdots + b_{iK}f_K + \varepsilon, i = 1, \cdots, p, \tag{19}$$

where Y_i is the excess return of the i-th asset over the risk-free asset, $f_1, \cdots, f_K$ are the excess returns of K factors that influence the returns of the market, the b_{ij}'s are unknown factor loadings, and $\varepsilon_1, \cdots, \varepsilon_p$ are idiosyncratic noises. The factor models have been widely applied and studied in economics and finance (see, e.g., Engle & Watson 1981, Chamberlain 1983, Chamberlain & Rothschild 1983, Bai 2003, Stock & Watson 2005). Famous examples include the Fama-French three-factor and five-factor models (Fama & French 1992, 1993). Yet the use of factor models on volatility matrix estimation for portfolio allocation was poorly understood until the work of Fan et al. (2008).

Thanks to the multifactor model (Equation 19), if a few factors can completely capture the cross-sectional risks, the number of parameters in covariance matrix estimation can be significantly reduced. For example, using the Fama-French three-factor model, we find that there are $4p$ instead of $p(p + 1)/2$ parameters. Despite the popularity of factor models, the impact of dimensionality on the estimation errors of covariance matrices and its applications to optimal portfolio allocation and portfolio risk assessment were not well studied until recently. As is now common in many applications, p can be large compared to the size n of the available sample. It is also necessary to study the situation in which the number of factors K diverges, which makes the K-factor model (Equation 19) better approximate the true underlying model as K grows. Thus it is important to study the factor model (Equation 19) in the asymptotic framework of $p \to \infty$ and $K \to \infty$.

Rewrite the factor model (Equation 19) in matrix form

$$\mathbf{y} = \mathbf{B}\mathbf{f} + \boldsymbol{\varepsilon}, \tag{20}$$

where $\mathbf{y} = (Y_1, \cdots, Y_p)^T$, $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_p)^T$ with $\mathbf{b}_i = (b_{i1}, \cdots, b_{iK})^T$, $\mathbf{f} = (f_1, \cdots, f_K)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_p)^T$. Denote $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{y})$, $\mathbf{X} = (\mathbf{f}_1, \cdots, \mathbf{f}_n)$, and $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$, where $(\mathbf{f}_1, \mathbf{y}_1), \cdots, (\mathbf{f}_n, \mathbf{y}_n)$ are n i.i.d. samples of $(\mathbf{f}, \mathbf{y})$. Fan et al. (2008) proposed a substitution estimator for $\boldsymbol{\Sigma}$,

$$\widehat{\Sigma} = \widehat{\mathbf{B}}\widehat{\mathrm{cov}}(\mathbf{f})\widehat{\mathbf{B}}^T + \widehat{\Sigma}_0, \tag{21}$$

where $\widehat{\mathbf{B}} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$ is the matrix of estimated regression coefficients, $\widehat{\mathrm{cov}}(\mathbf{f})$ is the sample covariance matrix of the factors $\mathbf{f}$, and $\widehat{\boldsymbol{\Sigma}}_0 = \mathrm{diag}(n^{-1}\widehat{\mathbf{E}}\widehat{\mathbf{E}}^T)$ is the diagonal matrix of $n^{-1}\widehat{\mathbf{E}}\widehat{\mathbf{E}}^T$ with $\widehat{\mathbf{E}} = \mathbf{Y} - \widehat{\mathbf{B}}\mathbf{X}$ the matrix of residuals. They derived the rates of convergence of the factor-model-based covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}$ and the sample covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_{\mathrm{sam}}$ simultaneously under the Frobenius norm $\|\cdot\|$ and a new norm $\|\cdot\|_{\boldsymbol{\Sigma}}$, where $\|\mathbf{A}\|_{\boldsymbol{\Sigma}} = p^{-1/2}\|\boldsymbol{\Sigma}^{-1/2}\mathbf{A}\boldsymbol{\Sigma}^{-1/2}\|$ for any $p \times p$ matrix $\mathbf{A}$. This new norm was introduced to better understand the factor structure. In particular, they showed that $\widehat{\boldsymbol{\Sigma}}$ has a faster convergence rate than $\widehat{\boldsymbol{\Sigma}}_{\mathrm{sam}}$ under the new norm. The inverses of covariance matrices play an important role in many applications such as optimal portfolio allocation. Fan et al. (2008) also compared the convergence rates of $\widehat{\boldsymbol{\Sigma}}^{-1}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{sam}}^{-1}$, illustrating the advantage of using the factor model. Furthermore, they investigated the impacts of covariance matrix estimation on some applications such as optimal portfolio allocation and portfolio risk assessment. They identified how large p and K can be such that the error in the estimated covariance is negligible in those applications. Explicit convergence rates of various portfolio variances were established.

In many applications, the factors are usually unknown to us. So it is important to study the factor models with unknown factors for the purpose of covariance matrix estimation. Constructing factors that influence the market itself is a high-dimensional variable selection problem. One can apply, e.g., sparse principal component analysis (see Johnstone & Lu 2004, Zou et al. 2006) to construct the unobservable factors. It is also practically important to consider dynamic factor models in which the factor loadings as well as the distributions of the factors evolve over time. The heterogeneity of the observations is another important aspect that needs to be addressed.

4. LIKELIHOOD BASED SPARSE MODELS

4.1. Penalized Likelihood

The ideas of the AIC and BIC suggest that one should choose a parameter vector $\boldsymbol{\beta}$ maximizing the penalized likelihood

$$\ell_n(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_0, \tag{22}$$

where $\ell_n(\boldsymbol{\beta})$ is the log-likelihood function and $\lambda \geq 0$ is a regularization parameter. The computational difficulty of the combinatorial optimization in Equation 22 stimulated many continuous relaxations, leading to a general penalized likelihood

$$n^{-1}\ell_n(\boldsymbol{\beta}) - \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{23}$$

where $p_\lambda(\cdot)$ is a penalty function indexed by the regularization parameter $\lambda \geq 0$ as in PLS (Equation 2).

It is nontrivial to maximize Equation 23 when p_λ is folded concave. In such cases, it is also generally difficult to study the global maximizer without the concavity of the objective function. As is common in the literature, the main attention of theory and implementations has been on local optimizers that have nice statistical properties. Many efficient algorithms have been proposed to optimize nonconcave penalized likelihoods. Fan & Li (2001) introduced the LQA algorithm by using the Newton-Raphson method and a quadratic approximation in Equation 6, and Zou & Li (2008) proposed the LLA algorithm with a linear approximation in Equation 7. With the trivial zero initial value for LLA, SCAD gives exactly the Lasso estimate. In this sense, the SCAD or more generally folded-concave regularization is an iteratively reweighted Lasso.

Coordinate optimization to implement regularization methods is fast when the univariate optimization problem has an analytic solution, which is the case for many commonly used penalty functions, such as Lasso, SCAD, and MCP. For example, Fan & Lv (2011) introduced the iterative coordinate ascent algorithm, a path-following coordinate optimization algorithm, to maximize the penalized likelihood (Equation 23) including PLS (Equation 2). It maximizes one coordinate at a time with successive displacements for Equation 23 with λ in decreasing order. More specifically, for each coordinate within each iteration, it uses the second-order approximation of $\ell_n(\boldsymbol{\beta})$ at the current p vector along that coordinate and maximizes directly the univariate penalized quadratic approximation. It updates each coordinate if the maximizer of that coordinate makes Equation 23 strictly increasing. Thus the iterative coordinate ascent algorithm enjoys the ascent property that the resulting sequence of values of the penalized likelihood (Equation 23) is increasing. Fan & Lv (2011) demonstrated that coordinate optimization works well and efficiently to produce the entire solution paths for concave penalties.

A natural question is, what are the sampling properties of penalized likelihood estimation (Equation 23) when the penalty function p_λ is not necessarily convex? Fan & Li (2001) studied the oracle properties of folded-concave penalized likelihood estimators in the finite-dimensional setting, and Fan & Peng (2004) generalized their results to the relatively high-dimensional setting of $p = o(n^{1/5})$ or $o(n^{1/3})$. Let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ be the true regression coefficients vector, with $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ the subvectors of nonsparse and sparse elements, respectively, and $s = \|\boldsymbol{\beta}_0\|_0$. Denote $\boldsymbol{\Sigma} = \mathrm{diag}\{p''_\lambda(|\boldsymbol{\beta}_1|)\}$ and $\bar{p}_\lambda(\boldsymbol{\beta}_1) = \mathrm{sgn}(\boldsymbol{\beta}_1) \circ p'_\lambda(|\boldsymbol{\beta}_1|)$, where $\circ$ denotes the Hadamard (componentwise) product. Under some regularity conditions, they showed that with probability tending to one as $n \to \infty$, there exists an $(n/p)^{\frac{1}{2}}$ consistent local maximizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ of Equation 23 satisfying (*a*) sparsity, $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$, and (*b*) asymptotic normality. For any unit vector $\mathbf{a}_n$ in R^s,

$$\sqrt{n}\mathbf{a}_n^T \mathbf{I}_1^{-1/2}(\mathbf{I}_1 + \boldsymbol{\Sigma})[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 + (\mathbf{I}_1 + \boldsymbol{\Sigma})^{-1}\bar{p}_\lambda(\boldsymbol{\beta}_1)] \xrightarrow{\mathcal{D}} N(0,1), \tag{24}$$

where $\mathbf{I}_1 = \mathbf{I}(\boldsymbol{\beta}_1)$ is the Fisher information matrix knowing the true model supp($\boldsymbol{\beta}_0$), and $\widehat{\boldsymbol{\beta}}_1$ is a subvector of $\widehat{\boldsymbol{\beta}}$ formed by components in supp($\boldsymbol{\beta}_0$). In particular, the SCAD estimator performs as well as the oracle estimator knowing the true model in advance, whereas the Lasso estimator generally does not.

A long-standing question in the literature is whether the penalized likelihood methods possess the oracle property in ultra-high dimensions. Fan & Lv (2011) addressed this problem in the context of generalized linear models with NP dimensionality: $\log p = O(n^a)$ for some $a > 0$. They proved that under some regularity conditions, there exists a local maximizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ of the penalized likelihood method (Equation 23) such that $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$ with probability tending to one and $\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \|_2 = O_P(\sqrt{s}n^{-1/2})$, where $s = \| \boldsymbol{\beta}_0 \|_0$. They also established asymptotic normality and thus the oracle property. Their studies demonstrate that the technical conditions are less restrictive for folded-concave penalties such as SCAD. The important question of when the folded-concave penalized likelihood estimator is a global maximizer of penalized likelihood (Equation 23) naturally arises. Fan & Lv (2011) characterized such a property from two perspectives: global optimality (for $p \leq n$) and restricted global optimality (for $p > n$). In addition, they showed that the SCAD penalized likelihood estimator can meet the oracle estimator under some regularity conditions. Other work on the topic includes Meier et al. (2008) and van de Geer (2008).

4.2. Penalized Partial Likelihood

Credit risk is a topic that has been extensively studied in the finance and economics literature. Various models have been proposed for pricing and hedging credit risky securities (see Jarrow 2009 for a review of credit-risk models). Cox (1972) introduced the famous Cox's proportional hazards model

$$h(t \mid \mathbf{x}) = h_0(t)e^{\mathbf{x}^T \boldsymbol{\beta}} \tag{25}$$

to accommodate the effect of covariates, in which $h(t|\mathbf{x})$ is the conditional hazard rate at time t, and $h_0(t)$ is the baseline hazard function. This model has been widely used in survival analysis to model time-to-event data. Such a model can naturally be applied to model credit default. Lando (1998) first addressed the issue of default correlation for pricing credit derivatives on baskets, e.g., collateralized debt obligation, by using the Cox processes. The default correlations are induced via common state variables that drive the default intensities (see Jarrow 2009, section 4; Duffie et al. 2009; and Duan et al. 2010 for more detailed discussions of the Cox model for credit default analysis).

Identifying important risk factors and quantifying their risk contributions are crucial aims of survival analysis. It is natural to extend the regularization methods to the Cox model. Tibshirani (1997) introduced the Lasso method (L_1 penalization method) to this model. To overcome the bias issue of convex penalties, Fan & Li (2002) employed the folded-concave penalty and partial likelihood methods. Let $t_1^0 < t_2^0 < \cdots < t_N^0$ be N ordered observed failure times (assuming no common failure times for simplicity). Denote by $\mathbf{x}_{(k)}$ the covariate vector of the subject with failure time t_k^0 and $R_k = \{i : y_i \geq t_k^0\}$ the risk set right before time t_k^0. Fan & Li (2002) considered the penalized partial likelihood

$$n^{-1} \sum_{k=1}^{N} \left[\mathbf{x}_{(k)}^T \boldsymbol{\beta} - \log \Big\{ \sum_{i \in R_k} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \Big\} \right] - \sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{26}$$

They proved the oracle properties for the folded-concave penalized partial likelihood estimator. Later, Zhang & Lu (2007) introduced the adaptive Lasso method for Cox's proportional hazards model, and Zou (2008) proposed a path-based variable selection method by using penalization with adaptive shrinkage. Both papers have shown the asymptotic efficiency of the methods.

5. SURE SCREENING METHODS

5.1. Sure Independence Screening

A natural idea for ultra-high-dimensional modeling is to apply a fast, reliable, and efficient method to reduce the dimensionality p from a large or huge scale [e.g., $\log p = O(n^a)$ for some $a > 0$] to a relatively large scale d [e.g., $O(n^b)$ for some $b > 0$] so that well-developed variable selection techniques can be applied to the reduced feature space. This powerful tool enables us to approach the problem of variable selection in sparse ultra-high-dimensional modeling. The issues of computational cost, statistical accuracy, and model interpretability will be addressed when the variable screening procedures retain all the important variables with asymptotic probability one, the so-called sure screening property introduced by Fan & Lv (2008).

Fan & Lv (2008) recently proposed the sure independence screening (SIS) methodology to reduce computation in sparse ultra-high-dimensional modeling. It also reduces the correlation requirements among predictors. The SIS method ranks features by their marginal correlations with the response. It is a specific case of independence screening, which ranks features with marginal utility; i.e., each feature is treated as an independent predictor to measure its effectiveness for prediction.

Assume that the $n \times p$ design matrix $\mathbf{X}$ has been standardized to have mean zero and variance one for each column and let $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_p)^T = \mathbf{X}^T\mathbf{y}$ be a p-dimensional vector of the componentwise regression estimator. For each d_n, Fan & Lv (2008) defined the submodel consisting of selected predictors as

$$\widehat{\mathcal{M}}_d = \{1 \le j \le p : |\omega_j| \text{ is among the first } d_n \text{ largest of all}\}. \tag{27}$$

This reduces the dimensionality of the feature space from $p \gg n$ to a (much) smaller scale d_n, which can be below n. This correlation learning screens variables that have weak marginal correlations with the response. It reduces the selection of features by two-sample t-test statistics in classification problems with class labels $Y = \pm 1$ (see Fan & Fan 2008). It is easy to see that SIS has computational complexity $O(np)$ and thus is fast to implement.

Denote by $\mathcal{M}_* = \{1 \le j \le p : \beta_j \neq 0\}$ the true underlying sparse model and $s = |\mathcal{M}_*|$ the nonsparsity size. Fan & Lv (2008) studied the ultra-high-dimensional setting of $p \gg n$ with $\log p = O(n^a)$ for some $a \in (0, 1 - 2\kappa)$ (see below for the definition of κ) and Gaussian noise $\varepsilon \sim N(0, \sigma^2)$. They assumed that $\mathrm{var}(Y) < \infty$ and that $\lambda_{\max}(\boldsymbol{\Sigma}) = O(n^\tau)$:

$$\min_{j \in \mathcal{M}_*} |\beta_j| \ge cn^{-\kappa} \quad \text{and} \quad \min_{j \in \mathcal{M}_*} |\mathrm{cov}(\beta_j^{-1} Y, X_j)| \ge c, \tag{28}$$

in which $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x}), \kappa, \tau, \ge 0$, and $c > 0$ is a constant. Under some regularity conditions, Fan & Lv (2008) showed that if $2\kappa + \tau < 1$, there exists a constant $\theta \in (2\kappa + \tau, 1)$ such that when $d_n \sim n^\theta$, we have for some $C > 0$

$$P(\mathcal{M}_* \subset \widehat{\mathcal{M}}_d) = 1 - O(e^{-Cn^{1-2\kappa}/\log n}). \tag{29}$$

This shows that SIS has the sure screening property even in ultra-high dimensions. With SIS, we can reduce exponentially growing dimensionality to a relatively large-scale $d_n \ll n$, retaining all the important variables in the reduced model $\widehat{\mathcal{M}}_d$ with a significant probability.

The above results have been extended by Fan & Song (2010) to cover non-Guassian covariates and non-Gaussian response. In the context of generalized linear models, they showed that independence screening through the use of marginal likelihood ratios or marginal regression coefficients possesses a sure screening property with the selected model size explicitly controlled. In particular, they do not impose elliptical symmetry of the distribution of covariates nor conditions on the covariance $\mathbf{\Sigma}$ of covariates. The latter property is a huge advantage over the penalized likelihood method, which requires restrictive conditions on the covariates.

There are other related methods of marginal screening. Huang et al. (2008) introduced marginal bridge regression, Hall & Miller (2009) proposed a generalized correlation for feature ranking, and Fan et al. (2011a) developed nonparametric screening using the B-spline basis. All these methods require that one choose a thresholding parameter. Zhao & Li (2010) proposed the use of an upper quantile of marginal utilities for decoupled (via random permutation) responses and covariates, called PSIS (principled sure independence screening), to select the thresholding parameter. The idea is to randomly permute the covariates and response so that they have no relation and then to compute the marginal utilities based on the permuted data and select the upper α quantile of the marginal utilities as the thresholding parameter. The choice of α is related to the false selection rate. A stringent screening procedure would take $\alpha = 0$, namely, the maximum of the marginal utilities for the randomly decoupled data. Hall et al. (2009) presented independence learning rules by tilting methods and empirical likelihood and proposed a bootstrap method to assess the accuracy of feature ranking in classification.

5.2. A Two-Scale Framework

When the dimensionality is reduced to a moderate scale d with a sure screening method such as SIS, the well-developed variable selection techniques can be applied to the reduced feature space. This provides a powerful tool of SIS-based methods for ultra-high-dimensional variable selection.

By its nature, SIS uses only the marginal information of predictors and does not look at their joint behavior. Fan & Lv (2008) noticed three potential issues of simple SIS that can make the sure screening property fail to hold. First, it may miss an important predictor that is marginally uncorrelated or weakly correlated but jointly correlated with the response. Second, it may select some unimportant predictors that are highly correlated with the important predictors and exclude important predictors that are relatively weakly related to the response. Third, the issue of collinearity among predictors is an intrinsic difficulty of the variable selection problem. To address these issues, Fan & Lv (2008) proposed iterative SIS, which iteratively applies the idea of large-scale screening and moderate-scale selection. This idea was extended and improved by Fan et al. (2009b) as follows.

Suppose that we wish to find a sparse $\boldsymbol{\beta}$ to minimize the objective

$$Q_n(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}), \tag{30}$$

where L is the loss function, which can be the quadratic loss, robust loss, log-likelihood, or quasi-likelihood. It is usually convex in $\boldsymbol{\beta}$. The first step is to apply the marginal screening, using the marginal utilities

$$L_j = \min_{\beta_0, \beta_j} n^{-1} \sum_{i=1}^{n} L(Y_i, \beta_0 + X_{i,j}\beta_j) \qquad (31)$$

or the magnitude $|\widehat{\beta}_j|$ of the minimizer of Equation 31 itself (assuming covariates are standardized in this case) to rank the covariates. This results in the active set $\mathcal{A}_1$. The thresholding parameter can be selected using the permutation method of Zhao & Li (2010) in Section 5.1. Now apply a penalized likelihood method

$$n^{-1} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{X}_{i,\mathcal{A}_1}^T \boldsymbol{\beta}_{\mathcal{A}_1}) + \sum_{j \in \mathcal{A}_1} p_\lambda(|\beta_j|) \qquad (32)$$

to further select a subset of active variables, resulting in $\mathcal{M}_1$. The next step is the conditional screening. Given the active set of covariates $\mathcal{M}_1$, what are the conditional contributions of those variables that were not selected in the first step? This leads us to define the conditional marginal utilities:

$$L_{j|\mathcal{M}_1} = \min_{\beta_0, \boldsymbol{\beta}_{\mathcal{M}_1}, \beta_j} n^{-1} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{X}_{i,\mathcal{M}_1}^T \boldsymbol{\beta}_{\mathcal{M}_1} + X_{ij}\beta_j). \qquad (33)$$

Note that for the quadratic loss, when $\boldsymbol{\beta}_{\mathcal{M}_1}$ is fixed at the minimizer from Equation 32, such a method reduces to the residual-based approach of Fan & Lv (2008). The current approach avoids the generalization of the concept of residuals to other complicated models and fully uses the conditional inference, but it involves more intensive computation in the conditional screening. We recruit additional variables by using the marginal utilities $L_{j|\mathcal{M}_1}$ or the magnitude of the minimizer (Equation 33). This is again a large-scale screening step, giving an active set $\mathcal{A}_2$, and the thresholding parameter can be chosen by the permutation method. The next step is then the moderate-scale selection. The potentially useful variables are now in the set $\mathcal{M}_1 \cup \mathcal{A}_2$. Apply the penalized likelihood technique to the problem:

$$\sum_{i=1}^{n} n^{-1} L(Y_i, \beta_0 + \mathbf{X}_{i,\mathcal{M}_1}^T \boldsymbol{\beta}_{\mathcal{M}_1} + \mathbf{X}_{i,\mathcal{A}_2}^T \boldsymbol{\beta}_{\mathcal{A}_2}) + \sum_{j \in \mathcal{M}_1 \cup \mathcal{A}_2} p_\lambda(|\beta_j|). \qquad (34)$$

This results in the selected variables $\mathcal{M}_2$. Note that some variables selected in the previous step $\mathcal{M}_1$ can be deleted in this step. Iterate the conditional large-scale screening followed by the moderate-scale selection until $\mathcal{M}_{\ell-1} = \mathcal{M}_\ell$ or the maximum number of iterations is reached. This takes account of the joint information of predictors in the selection and avoids solving large-scale optimization problems. The success of such a two-scale method and its theoretical properties are documented by Fan & Lv (2008), Fan et al. (2009b), Zhao & Li (2010), Fan & Song (2010), and Fan et al. (2011a).

6. CONCLUSIONS

Above we briefly survey some recent developments of sparse high-dimensional modeling and discuss some applications in economics and finance. In particular, the recent developments in ultra-high-dimensional variable selection can be widely applied to the statistical analysis of large-scale economic and financial problems. Those sparse modeling problems deserve further study both theoretically and empirically. We focus on regularization methods including PLS and penalized likelihood. The role of penalty functions and the impact of dimensionality on sparse high-dimensional modeling are revealed and discussed. Sure independent screening has been introduced to reduce dimensionality and the problems of collinearity. It is a fundamental element of the promising two-scale framework in ultra-high-dimensional econometrics modeling.

Sparse models are ideal and generally biased. Yet they have proven to be effective in many large-scale applications. The biases are typically small as variables are selected from a large pool to best approximate the true model. High-dimensional statistical learning facilitates undoubtedly the understanding and derivation of the often-complex nature of statistical relationships among the explanatory variables and the response. In particular, the key notion of sparsity, which helps reduce the intrinsic complexity at little cost of statistical efficiency and computation, provides powerful tools to explore relatively low-dimensional structures among huge amounts of candidate models. It should have a marked impact on econometric theory and practice, from econometric modeling to fundamental understanding of economic problems. New novel modeling techniques are needed to address the challenges in the frontiers of economics and finance, and other social science problems.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We sincerely thank Prof. Kenneth J. Arrow and Prof. James J. Heckman for their kind invitation to write this article. We also appreciate the helpful comments of the reviewers. Prof. Fan's research was partially supported by National Science Foundation grants DMS-0704337 and DMS-0714554 and National Institutes of Health grant R01-GM072611. Prof. Lv's research was partially supported by National Science Foundation CAREER Award DMS-0955316 and grant DMS-0806030.

LITERATURE CITED

Antoniadis A. 1996. Smoothing noisy data with tapered coiflets series. *Scand. J. Stat.* 23:313–30

Antoniadis A, Fan J. 2001. Regularization of wavelet approximations. *J. Am. Stat. Assoc.* 96:939–67

Bai J. 2003. Inferential theory for factor models of large dimensions. *Econometrica* 71:135–71

Bai J, Ng S. 2008. Large dimensional factor analysis. *Found. Trends Econom.* 3(2):89–163

Barron A, Birge L, Massart P. 1999. Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* 113:301–413

Belloni A, Chernozhukov V. 2009. *Post-L1-penalized estimators in high-dimensional linear regression models*. Unpublished manuscript, Duke Univ./Mass. Inst. Technol.

Belloni A, Chernozhukov V. 2011. ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Stat.* 39(1):82–130

Bernanke B, Boivin J, Eliasz PS. 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *Q. J. Econ.* 120(1):387–422

Bickel PJ. 2008. Discussion of "Sure independence screening for ultrahigh dimensional feature space." *J. R. Stat. Soc. B* 70:883–84

Bickel PJ, Levina E. 2008a. Regularized estimation of large covariance matrices. *Ann. Stat.* 36:199–227

Bickel PJ, Levina E. 2008b. Covariance regularization by theresholding. *Ann. Stat.* 36:2577–604

Bickel PJ, Ritov Y, Tsybakov A. 2009. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* 37:1705–32

Bradic J, Fan J, Wang W. 2011. Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. *J. R. Stat. Soc. B.* 73:In press

Breiman L. 1995. Better subset regression using the non-negative garrote. *Technometrics* 37:373–84

Breiman L. 1996. Heuristics of instability and stabilization in model selection. *Ann. Stat.* 24:2350–83

Brodie J, Daubechies I, De Mol C, Giannone D, Loris I. 2009. Sparse and stable Markowitz portfolios. *Proc. Natl. Acad. Sci. USA* 106(30):12267–72

Cai T, Zhang C-H, Zhou H. 2010. Optimal rates of convergence for covariance matrix estimation. *Ann. Stat.* 38:2118–44

Calomiris CW, Longhofer SD, Miles W. 2008. *The foreclosure-house price nexus: lessons from the 2007–2008 housing turmoil*. NBER Work. Pap. 14294

Campbell J, Lo A, MacKinlay C. 1997. *The Econometrics of Financial Markets*. Princeton, NJ: Princeton Univ. Press

Candes E, Tao T. 2007. The Dantzig selector: statistical estimation when *p* is much larger than *n*. *Ann. Stat.* 35:2313–404

Chamberlain G. 1983. Funds, factors and diversification in arbitrage pricing theory. *Econometrica* 51:1305–23

Chamberlain G, Rothschild M. 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* 51:1281–304

Cochrane JH. 2005. *Asset Pricing*. Princeton, NJ: Princeton Univ. Press. Rev. ed.

Cox DR. 1972. Regression models and life-tables. *J. R. Stat. Soc. B* 34:187–220

Craven P, Wahba G. 1978. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* 31(4):377–403

Daubechies I, Defrise M, De Mol C. 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.* 57:1413–57

DeMiguel V, Garlappi L, Nogales FJ, Uppal R. 2009. A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. *Manage. Sci.* 55(5):798–812

Donoho DL, Elad M, Temlyakov V. 2006. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory* 52:6–18

Duan J-C, Sun J, Wang T. 2010. *Multiperiod corporate default prediction: a forward intensity approach*. Unpublished manuscript, Natl. Univ. Singapore

Duffie D, Eckner A, Horel G, Saita L. 2009. Frailty correlated default. *J. Finance* 64:2089–123

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. *Ann. Stat.* 32:407–99

El Karoui N. 2008. Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Stat.* 36:2717–56

Engle RF, Watson MW. 1981. A one-factor multivariate time series model of metropolitan wage rates. *J. Am. Stat. Assoc.* 76:774–81

Fama E, French K. 1992. The cross-section of expected stock returns. *J. Finance* 47:427–65

Fama E, French K. 1993. Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* 33:3–56

Fan J, Fan Y. 2008. High-dimensional classification using features annealed independence rules. *Ann. Stat.* 36:2605–37

Fan J, Fan Y, Lv J. 2008. High dimensional covariance matrix estimation using a factor model. *J. Econom.* 147:186–97

Fan J, Feng Y, Song R. 2011a. Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Am. Stat. Assoc.* In press

Fan J, Feng Y, Wu Y. 2009a. Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Stat.* 3:521–41

Fan J, Guo S, Hao N. 2011b. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc.* In press

Fan J, Li R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 96:1348–60

Fan J, Li R. 2002. Variable selection for Cox's proportional hazards model and frailty model. *Ann. Stat.* 30:74–99

Fan J, Lv J. 2008. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. B* 70:849–911

Fan J, Lv J. 2011. Non-concave penalized likelihood with NP-dimensionality. *IEEE Trans. Inf. Theory.* In press

Fan J, Lv J. 2010. A selective overview of variable selection in high dimensional feature space. *Stat. Sinica* 20:101–48

Fan J, Peng H. 2004. Nonconcave penalized likelihood with diverging number of parameters. *Ann. Stat.* 32:928–61

Fan J, Samworth R, Wu Y. 2009b. Ultrahigh dimensional variable selection: beyond the linear model. *J. Mach. Learn. Res.* 10:1829–53

Fan J, Song R. 2010. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Stat.* 38:3567–604

Fan J, Zhang J, Yu K. 2011c. Asset allocation and risk assessment with gross exposure constraints for vast portfolios. Manuscript submitted

Frank IE, Friedman JH. 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35:109–48

Friedman J, Hastie T, Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9:432–41

Fu WJ. 1998. Penalized regression: the bridge versus the lasso. *J. Comput. Graph. Stat.* 7:397–416

Hall P, Marron JS, Neeman A. 2005. Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. B* 67:427–44

Hall P, Miller H. 2009. Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Stat.* 18(3):533–50

Hall P, Pittelkow Y, Ghosh M. 2008. Theoretic measures of relative performance of classifiers for high dimensional data with small sample sizes. *J. R. Stat. Soc. B* 70:158–73

Hall P, Titterington DM, Xue J-H. 2009. Tilting methods for assessing the influence of components in a classifier. *J. R. Stat. Soc. B* 71:783–803

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer-Verlag. 2nd ed.

Himmelberg C, Mayer C, Sinai T. 2005. Assessing high house prices: bubbles, fundamentals and misperceptions. *J. Econ. Perspect.* 19(4):67–92

Huang J, Horowitz J, Ma S. 2008. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Stat.* 36:587–613

Huang JZ, Liu N, Pourahmadi M, Liu L. 2006. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* 93:85–98

Hunter DR, Li R. 2005. Variable selection using MM algorithms. *Ann. Stat.* 33:1617–42

Jagannathan R, Ma T. 2003. Risk reduction in large portfolios: why imposing the wrong constraints helps. *J. Finance* 58(4):1651–83

Jarrow RA. 2009. Credit risk models. *Annu. Rev. Financ. Econ.* 1:37–68

Johnstone IM. 2001. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* 29:295–327

Johnstone IM, Lu AY. 2004. *Sparse principal components analysis*. Unpublished manuscript, Stanford Univ./Renaissance Technol.

Kim Y, Choi H, Oh HS. 2008. Smoothly clipped absolute deviation on high dimensions. *J. Am. Stat. Assoc.* 103:1665–73

Kim Y, Kwon S. 2009. *On the global optimum of the SCAD penalized estimator*. Unpublished manuscript, Seoul Natl. Univ.

Koltchinskii V. 2008. Sparse recovery in convex hulls via entropy penalization. *Ann. Stat.* 37(3):1332–59

Lam C, Fan J. 2009. Sparsistency and rates of convergence in large covariance matrices estimation. *Ann. Stat.* 37:4254–78

Lando D. 1998. On Cox processes and credit risky securities. *Rev. Deriv. Res.* 2:99–120

Ledoit O, Wolf M. 2004. A well conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* 88:365–411

Levina E, Rothman AJ, Zhu J. 2008. Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Stat.* 2:245–63

Lv J, Fan Y. 2009. A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Stat.* 37:3498–528

Markowitz HM. 1952. Portfolio selection. *J. Finance* 7:77–91

Markowitz HM. 1959. *Portfolio Selection: Efficient Diversification of Investments*. New York: Wiley & Sons

Meier L, van de Geer S, Bühlmann P. 2008. The group lasso for logistic regression. *J. R. Stat. Soc. B* 70:53–71

Meinshausen N, Bühlmann P. 2006. High dimensional graphs and variable selection with the Lasso. *Ann. Stat.* 34:1436–62

Ng S, Moench E. 2011. A factor analysis of housing market dynamics in the U.S. and the regions. *Econom. J.* In press

Osborne MR, Presnell B, Turlach BA. 2000. On the LASSO and its dual. *J. Comput. Graph. Stat.* 9:319–37

Rapach DE, Strass JK. 2007. Forecasting real housing price growth in the eighth district states. *Reg. Econ. Dev.* 3(2):33–42

Rosset S, Zhu J. 2007. Piecewise linear regularized solution paths. *Ann. Stat.* 35:1012–30

Rothman AJ, Bickel PJ, Levina E, Zhu J. 2008. Sparse permutation invariant covariance estimation. *Electron. J. Stat.* 2:494–515

Sims CA. 1980. Macroeconomics and reality. *Econometrica* 48(1):1–48

Stein C. 1975. *Estimation of a covariance matrix*. Presented as Rietz Lecture, IMS Annu. Meet., 39th, Atlanta, Georgia

Stock JH, Watson MW. 2001. Vector autoregressions. *J. Econ. Perspect.* 15(4):101–15

Stock JH, Watson MW. 2005. *Implications of dynamic factor models for VAR analysis*. NBER Work. Pap. 11467

Stock JH, Watson MW. 2006. Forecasting with many predictors. In *Handbook of Economic Forecasting*, Vol. 1, ed. G Elliott, C Granger, A Timmermann, pp. 515–54. Amsterdam: North-Holland

Stock JH, Watson MW. 2010. The evolution of national and regional factors in U.S. housing construction. In *Volatility Time Series Econometrics*, ed. T Bollerslev, J Russell, M Watson, pp. 35–62. New York: Oxford Univ. Press

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58:267–88

Tibshirani R. 1997. The lasso method for variable selection in the Cox model. *Stat. Med.* 16:385–95

van de Geer S. 2008. High-dimensional generalized linear models and the lasso. *Ann. Stat.* 36:614–45

Wainwright MJ. 2006. Sharp thresholds for high-dimensional and noisy recovery of sparsity. *Tech. Rep.*, Dep. Stat., Univ. Calif., Berkeley

Wang H, Li B, Leng C. 2009. Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. B* 71(3):671–83

Wang H, Li R, Tsai CL. 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94:553–68

Wu TT, Lange K. 2008. Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* 2:224–44

Wu WB, Pourahmadi M. 2003. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* 90:831–44

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B* 68:49–67

Yuan M, Lin Y. 2007. Model selection and estimation in the Gaussian graphical model. *Biometrika* 94:19–35

Zhang CH. 2010. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* 38 (2):894–942

Zhang HH, Lu W. 2007. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 94:691–703

Zhao P, Yu B. 2006. On model selection consistency of Lasso. *J. Mach. Learn. Res.* 7:2541–63

Zhao SD, Li Y. 2010. *Principled sure independence screening for Cox models with ultra-high-dimensional covariates*. Unpublished manuscript, Harvard Univ.

Zou H. 2006. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101:1418–29

Zou H. 2008. A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* 95:241–47

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67:301–20

Zou H, Hastie T, Tibshirani R. 2006. Sparse principal component analysis. *J. Comput. Graph. Stat.* 15:265–86

Zou H, Li R. 2008. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.* 36:1509–66

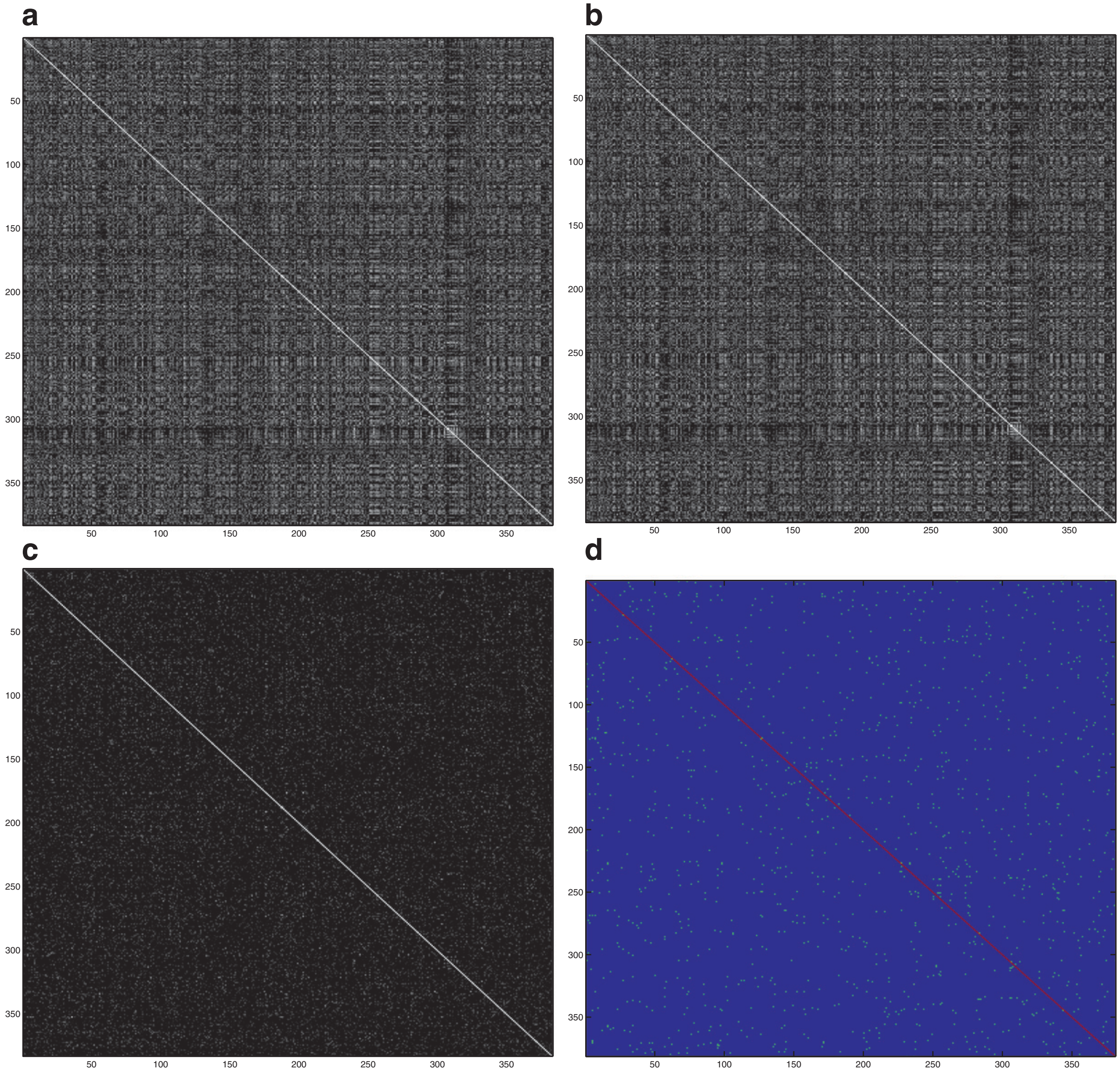

Figure 1

Cross-county correlation captured by a sparse vector autoregressive model. (*a*) The spatial correlation of 352 home-price appreciation (HPA) data series. (*b*) The residual correlation of an ordinary least-squares (OLS) approach, using only the national factor, not neighboring HPAs. For panels *a* and *b*, the correlation is from OLS residuals, conditional on national home price index (HPI). (*c*) The spatial correlation of residuals with national HPA and neighborhood selection. The correlation is from penalized least-squares residuals, conditional on national HPI. (*d*) Neighborhoods with nonzero regression coefficients. For each county, only three to four neighboring counties are chosen on average. The correlation is from sparse structure.

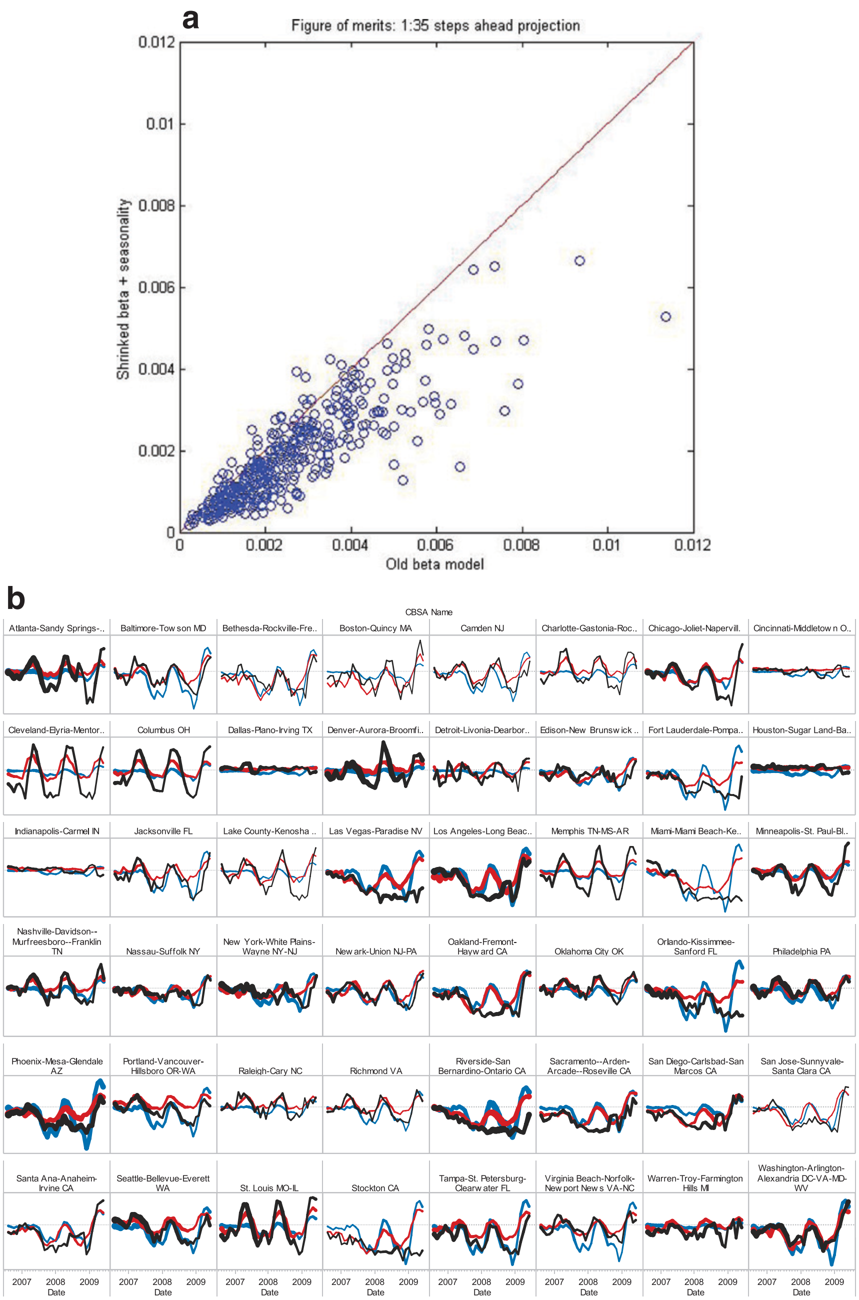

Figure 2

(*a*) Forecast error comparison over 352 counties. For each dot, the *x* axis represents the error by ordinary least squares (OLS) with only a national factor, and the *y* axis represents the error by penalized least squares (PLS) with additional neighborhood information. If the dot lies below the 45° line, PLS outperforms OLS. (*b*) Forecast comparison for the largest counties during the test period. The blue lines represent OLS with national house-price appreciation (HPA) data, the red lines represent PLS with additional neighborhood information, and the black lines are the historical HPAs. The thickness of the lines is proportional to repeated sales.

Annual Review of
Economics

Volume 3, 2011

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Economics* articles may be found at http://econ.annualreviews.org