Statistics and Its Interface Volume 9 (2016) 469–484

# Model diagnostics in reduced-rank estimation

Kun Chen*,†

Reduced-rank methods are very popular in high-dimensional multivariate analysis for conducting simultaneous dimension reduction and model estimation. However, the commonly-used reduced-rank methods are not robust, as the underlying reduced-rank structure can be easily distorted by only a few data outliers. Anomalies are bound to exist in big data problems, and in some applications they themselves could be of the primary interest. While naive residual analysis is often inadequate for outlier detection due to potential masking and swamping, robust reduced-rank estimation approaches could be computationally demanding. Under Stein's unbiased risk estimation framework, we propose a set of tools, including leverage score and generalized information score, to perform model diagnostics and outlier detection in large-scale reduced-rank estimation. The leverage scores give an exact decomposition of the so-called model degrees of freedom to the observation level, which lead to exact decompositions of many commonly-used information criteria; the resulting quantities are thus named information scores of the observations. The proposed information score approach provides a principled way of combining the residuals and leverage scores for anomaly detection. Simulation studies confirm that the proposed diagnostic tools work well. A pattern recognition example with hand-writing digital images and a time series analysis example with monthly U.S. macroeconomic data further demonstrate the efficacy of the proposed approaches.

AMS 2000 subject classifications: Primary 62M10; secondary 62J12.

Keywords and phrases: Big data, Information score, Model diagnostics, Multivariate regression, Outlier detection, Reduced-rank estimation.

## 1. INTRODUCTION

With $n$ independent observations of response $\mathbf{y}_i \in \mathbb{R}^q$ and predictor $\mathbf{x}_i \in \mathbb{R}^p$, we consider a multivariate regression model

$$\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E}, \tag{1}$$

where $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times q}$ is the response matrix, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$ is the predictor matrix, $\mathbf{C} \in \mathbb{R}^{p\times q}$ is an unknown coefficient matrix, and $\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times q}$ is an error matrix consisting of independent and identically distributed (i.i.d.) zero-mean error vectors. Such multivariate learning problems, in which either the model dimensions $(p, q)$ or the sample size $n$ can be very large, have become increasingly common, due to the unprecedented data explosion with ever increasing volume and complexity. For example, in genetics, it is of great interest to understand how gene expression profiles are related to DNA copy number variations [50]. In a study of human lung disease, the goal is to use detailed segmental lung airway measurements from CT-scanned images to predict various lung pulmonary function test results [15, 16]. Many matrix approximation methods such as principal component analysis (PCA) can be formulated as a special case of Model 1, for which $\mathbf{X}$ becomes $\mathbf{I}_p$, the $p$-dimensional identity matrix. A typical application is in image denoising and compression, in which a primary task is to extract true image signals from corrupted noisy inputs [4].

To make sense of the dependency between the possibly high-dimensional responses and predictors, it is often assumed that the coefficient matrix $\mathbf{C}$ admits certain low-dimensional structure, the exploitation of which may mitigate the curse of dimensionality, enhance model interpretability, and improve model predictive performance. Along this line, various regularized estimation methods have been developed. Several ridge regression methods relied on shrinkage estimation to overcome the deficiencies of multicollinearity [23]. Most sparse multivariate regression methods focused on predictor selection by exploring certain sparse patterns in $\mathbf{C}$ [38, 40, 37]. While the above multivariate methods are usually readily extended from their univariate counterparts, the reduced-rank estimation methodology possesses a genuine multivariate flavor [2, 25, 39]. The seminal reduced-rank regression (RRR) [2] achieved dimension reduction by imposing a low-rank constraint on $\mathbf{C}$, which could dramatically reduce the number of free parameters and induce an appealing latent variable interpretation. Bunea *et al.* [5] extended RRR to high dimensional settings. Yuan *et al.* [53] proposed the convex nuclear-norm penalized method to achieve simultaneous rank reduction and shrinkage estimation; see, also, Negahban and Wainwright [36], Lu *et al.* [29], Mukherjee and Zhu [34] and Chen *et al.* [13]. The reduced-rank representation has been further extended to enable variable selection [11, 6, 14], and it can be distilled from many other multivariate tools such as PCA, canonical correlation analysis, and matrix completion [9, 27]. It is evident that the reduced-rank methods have greatly facilitated

*This work is partially supported by the National Institutes of Health (U01–HL114494).
†This work is partially supported by the National Science Foundation (DMS-1613295).

scientific investigations in various disciplines, e.g., finance [39], ecology [12], neuroscience [55], genetics [31], etc.

In a nutshell, reduced-rank estimation has become an indispensable ingredient in modern statistical learning. However, despite of its effectiveness, reduced-rank estimation can be very sensitive to the presence of gross outliers and influential points. In real applications, anomalies are bound to appear, and they could severely distort or even completely destroy the underlying reduced-rank structure of interest. As such, the nonrobustness of reduced-rank estimation may greatly jeopardize its applicability in big data applications. For example, in the aforementioned lung disease study, reduced-rank estimation can be applied to identify a few latent pathways as some linear combinations of the lung airway measurements, to link the airway tree alternation to certain lung disease status. However, the identified pathways and airway features could be distorted by a few abnormal samples that deviate from the assumed model. Ignoring the heterogeneous effects caused by these samples could lead to misunderstanding about the actual disease mechanism. Yet in another type of applications, the primary interest may be to capture certain unusual signals and rare jumps. For example, to detect motion of certain objects using surveillance video frames, a reduced-rank model component is designed to extract the common background of the images, and it is what remains capture the motion of the objects.

Therefore, outlier detection is critical in reduced-rank analysis. This task, however, is notoriously difficult, due to both the high-dimensionality of the problem and the nonlinearity/nonconvexity of the low-rank structure. In the context of unsupervised learning, extensive research has been devoted to robust PCA [51, 8]. In supervised learning, She and Chen [46] proposed robust reduced-rank regression ($\mathrm{R}^4$) to conduct joint outlier detection and robust low-rank regression, and the method was shown to have a strong connection with M-estimation and possess attractive finite-sample robustness guarantees. However, there is no free lunch: in general these robust methods have a much increased computation cost comparing to the plain RRR, as the estimators no longer admit explicit form and certain iterative procedure is required in optimization [7].

An overlooked yet effective approach for robust reduced-rank modeling is to perform model diagnostics in the regular reduced-rank estimation, which can be both easy to implement and computationally efficient. Many existing model diagnostic tools, such as studentized residual, leverage score and Cook's distance [17], were mainly developed for linear models in low-dimensional setups. Zhu *et al.* [54] considered perturbation and scaled Cook's distance. Several methods considered extensions of these tools to multivariate regression models [18, 41, 56]. However, to the best of our knowledge, model diagnostics have not been studied in high-dimensional reduced-rank estimation.

In this paper, we develop a set of model diagnostic tools in large-scale reduced-rank regression problems, including leverage score (LEV), generalized information score (GIS), and Cook's distance (CD). In Section 2, we present a general class of reduced-rank estimators and reveal its nonrobustness against data corruption. The Stein's unbiased risk estimation framework [47] is reviewed in Section 3.1, based on which we derive the leverage scores of the data observations for reduced-rank estimation in Section 3.2. This task is nontrivial, as in nonlinear models the leverage scores depend on both the responses and the predictors and in general do not admit explicit form. Motivated by Mukherjee *et al.* [33], we show that all the ingredients required for computing the reduced-rank estimator and the leverage scores can be obtained from a single singular value decomposition (SVD), and the computation is efficient with careful manipulations of matrix operations. The model residuals and the leverage scores deliver important and complementary messages about how the observed data points fit to the model. In Section 3.3, we consider how to best use the two measures together to achieve a unified assessment of each observation. Intriguingly, the leverage scores of all the observations give an exact decomposition of the so-called model degrees of freedom, which then lead to exact decomposition of an unbiased estimator of the true model prediction error as well as many information criteria such as AIC [1] and BIC [43]. For a given information criterion, we term the decomposed quantity for each individual observation as its information score, which measures the contribution from this observation to the overall information criterion and hence also indicates the outlying effect of this observation. As such, the proposed information score approach provides a principled way of combining model residuals and leverage scores for evaluating the observations, and the idea is generally applicable in linear/nonlinear estimation problems. In Section 3.4, we consider approximating the Cook's distance in reduced-rank estimation. Numerical studies presented in Sections 4 and 5 confirm that using the proposed diagnostic measures is indeed a simple yet effective way for anomaly detection in large-scale data analysis. Some concluding remarks are provided in Section 6.

## 2. REDUCED-RANK ESTIMATION

Consider the multivariate regression model in (1). Let $\widehat{\mathbf{Y}}_{LS}$ be the ordinary least squares estimator (LS) of the regression component,

$$\widehat{\mathbf{Y}}_{LS} = \mathbf{X}\widehat{\mathbf{C}}_{LS} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \underset{n\times\bar{r}}{\mathbf{W}}\,\underset{\bar{r}\times\bar{r}}{\mathbf{D}}\,\underset{\bar{r}\times q}{\mathbf{V}}{}^{\mathrm{T}}, \tag{2}$$

where $(\cdot)^{-}$ denotes the Moore-Penrose inverse, and $\mathbf{WDV}^{\mathrm{T}}$ is the SVD of $\widehat{\mathbf{Y}}_{LS}$, i.e., $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$, $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$, and $\mathbf{D} = \mathbf{diag}\{d_i, i = 1, ..., \bar{r}\}$ with $d_1 > \cdots > d_{\bar{r}} > 0$. Here $\widehat{\mathbf{Y}}_{LS}$ is of rank $\bar{r} \le \min(r_x, q)$, with $r_x = r(\mathbf{X})$ being the rank of the design matrix. Without loss of generality, we assume the nonzero singular values of $\widehat{\mathbf{Y}}_{LS}$ are distinct, so that the SVD of $\widehat{\mathbf{Y}}_{LS}$ is unique up to the signs of the singular vectors.

We consider reduced-rank estimation by minimizing the following singular-value penalized least squares criterion

$$\frac{1}{2}\|\mathbf{Y}-\mathbf{XC}\|_F^2+\sum_{k=1}^{\min(r_x,q)}\rho(d_k(\mathbf{XC});\lambda\omega_k), \tag{3}$$

where $d_k(\cdot)$ denotes the $k$th largest singular value of the enclosed matrix, $\rho$ is certain sparsity-inducing penalty function, $\lambda\geq 0$ is the penalty level, and $\omega_k$s are some prespecified adaptive weights satisfying $0\leq\omega_1\leq\omega_2\cdots\leq\omega_{\min(r_x,q)}$. Minimizing (3) is the same as minimizing $\|\widehat{\mathbf{Y}}_{LS}-\mathbf{B}\|_F^2/2+\sum\rho(d_k(\mathbf{B});\lambda\omega_k)$ subject to $\mathbf{B}=\mathbf{XC}$, owing to the Pythagoras' theorem. As shown by She [45] and Chen *et al.* [13], with a wide class of $\rho$, the constraint $\mathbf{B}=\mathbf{XC}$ can be dropped, since the solution of $\mathbf{B}$, denoted as $\widehat{\mathbf{Y}}(\lambda)$, still belongs to the column space of $\mathbf{X}$ and is explicitly given by singular value thresholding. This produces a general class of reduced-rank estimators:

$$\begin{aligned}\widehat{\mathbf{Y}}(\lambda)=\mathbf{X}\widehat{\mathbf{C}}(\lambda)&=\mathbf{W}\Theta^{\sigma}(\mathbf{D};\lambda)\mathbf{V}^{\mathrm{T}}\\&=(\mathbf{WDV}^{\mathrm{T}})\mathbf{VD}^{-1}\Theta^{\sigma}(\mathbf{D};\lambda)\mathbf{V}^{\mathrm{T}}\\&=\widehat{\mathbf{Y}}_{LS}\sum_{k=1}^{\bar{r}}(\Theta(d_k;\lambda\omega_k)/d_k)\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}\\&=\widehat{\mathbf{Y}}_{LS}\sum_{k=1}^{\bar{r}}s(d_k;\lambda\omega_k)\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}.\end{aligned} \tag{4}$$

The function $\Theta^{\sigma}(\cdot;\lambda)$ denotes an arbitrary singular value thresholding rule associated with $\rho$ [45]; for the diagonal matrix $\mathbf{D}$, $\Theta^{\sigma}(\mathbf{D};\lambda)=\mathbf{diag}\{\Theta(d_k;\lambda\omega_k),k=1,\ldots,\bar{r}\}$, where $\Theta(\cdot;\lambda)$ is an arbitrary thresholding rule associated with $\rho$, defined as some odd monotone unbounded shrinkage function. See details in She [44], in which a general functional link between $\rho$ and $\Theta$ was established. Here we have also defined $s(d_k;\lambda\omega_k)=\Theta(d_k;\lambda\omega_k)/d_k$, and consequently they satisfy $1\geq s(d_1;\lambda h_1)\geq\cdots\geq s(d_{\bar{r}};\lambda h_{\bar{r}})\geq 0$. For simplicity, we may write $s(d_k;\lambda\omega_k)=s(d_k;\lambda)=s_k$ when no confusion arises.

In (3), the singular values of the regression component $\mathbf{XC}$ are penalized rather than those of $\mathbf{C}$. It is mainly due to this setup, (3) is able to produce the explicit reduced-rank solution given in (4). This class of estimators shares the same set of singular vectors with the LS estimator, but their singular value estimates are some shrunk or thresholded versions of the estimated singular values from LS. Some commonly-used penalty forms in (3) include the rank penalty [5] and the nuclear-norm penalty [53]. Specifically, a rank-penalized criterion can be expressed as

$$\frac{1}{2}\|\mathbf{Y}-\mathbf{XC}\|_F^2+\frac{\lambda^2}{2}\sum_{k=1}^{\min(r_x,q)}I(d_k(\mathbf{XC})\neq 0), \tag{5}$$

where $I(\cdot)$ is the indicator function. Here $\sum_{k=1}^{\min(r_x,q)}I(d_k(\mathbf{XC})\neq 0)=r(\mathbf{XC})$, and penalizing $r(\mathbf{XC})$ is equivalent to penalizing $r(\mathbf{C})$ [13]. The solution of (5) is obtained by a singular value hard-thresholding operation, i.e.,

$$s(d_k;\lambda)=I(d_k>\lambda),\qquad k=1,...,\min(r_x,q).$$

Equivalently, the set of reduced-rank regression (RRR) estimators, obtained by minimizing $\|\mathbf{Y}-\mathbf{XC}\|_F^2$ subject to $r(\mathbf{C})\leq r$, for $r=1,\ldots,\min(r_x,q)$, spans the solution path of (5). The adaptive nuclear-norm penalized criterion (ANN) is

$$\frac{1}{2}\|\mathbf{Y}-\mathbf{XC}\|_F^2+\lambda\sum_{k=1}^{\min(r_x,q)}\omega_k d_k(\mathbf{XC}), \tag{6}$$

where the weights can be chosen as $\omega_k=d_k^{-\gamma}(\widehat{\mathbf{Y}}_{LS})=d_k^{-\gamma}$ with $\gamma$ being a prespecified nonnegative constant. The estimator is given by an adaptive singular value soft-thresholding operation, i.e.,

$$s(d_k;\lambda)=(1-\lambda d_k^{-\gamma-1})_+,\qquad k=1,...,\min(r_x,q),$$

where $x_+=\max(0,x)$ is the nonnegative part of $x$.

Outliers are bound to occur in practice, especially in big data applications where the reduced-rank methodology is supposed to be the most effective. However, reduced-rank estimation methods can be highly nonrobust. Using the notation of break-down point from robust estimation literature, the following results state that it takes only a single outlier to destroy the reduced-rank estimators in (3). More details on the nonrobustness of reduced-rank estimation can be found in She and Chen [46].

**Proposition 1.** *[46] Let $\widehat{\mathbf{C}}(\lambda;\mathbf{X},\mathbf{Y})$ denote the reduced-rank estimator by solving (3), with finite data $(\mathbf{X},\mathbf{Y})$ and some tuning parameters $\lambda\geq 0$, $0\leq\omega_1\leq\ldots\leq\omega_{\min(r_x,q)}<\infty$. Define its break down point as*

$$\begin{aligned}&\tau(\widehat{\mathbf{C}}(\lambda;\mathbf{X},\mathbf{Y}))\\=&\frac{1}{N}\min\Bigg\{k\in\{0,1,\ldots,N\}:\\&\qquad\sup_{\tilde{\mathbf{Y}}:\|\tilde{\mathbf{Y}}-\mathbf{Y}\|_0\leq k}\|\mathbf{X}\widehat{\mathbf{C}}(\lambda;\mathbf{X},\tilde{\mathbf{Y}})\|_F=\infty\Bigg\},\end{aligned}$$

*where $N=nq$, and $\|\mathbf{A}\|_0=\sum_i\sum_j I(a_{ij}\neq 0)$ for any matrix $\mathbf{A}=(a_{ij})$. Then we have, $\tau(\widehat{\mathbf{C}}(\lambda;\mathbf{X},\mathbf{Y}))=1/N$.*

## 3. DIAGNOSTIC TOOLS FOR REDUCED-RANK ESTIMATION

From Proposition 1, it is pressing to consider outlier detection and robust estimation problems in reduced-rank estimation. Our focus in this paper is on developing some simple yet effective model diagnostic tools, which preferably can be computed as some byproducts from the estimation procedure outlined in Section 2. In real applications, robust

estimation methods [46] and the proposed diagnostic tools are often complementary to each other, and together they may better fulfill the needs of handling large data problems.

### 3.1 Some key concepts from Stein's unbiased risk estimation

We begin with a brief description of the Stein's Unbiased Risk Estimation (SURE) theory [47, 19], in the general context of multivariate regression. To formulate, assume that we observe $\mathbf{Y} \in \mathbb{R}^{n\times q}$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$. Consider the model $\mathbf{Y} = \boldsymbol{\mu}(\mathbf{X}) + \mathbf{E}$, where $\boldsymbol{\mu}(\mathbf{X})$ is the mean function, and the entries of $\mathbf{E}$, $e_{ij}$, are i.i.d. with mean zero and variance $\sigma^2$. For an estimation procedure $f$, denote its fitted value of $\mathbf{Y}$ as $\widehat{\mathbf{Y}} = f(\mathbf{X}, \mathbf{Y})$.

Define the standardized apparent error and true prediction error for the $(i, j)$th observation as

$$\mathrm{err}_{ij} = (y_{ij} - \widehat{y}_{ij})^2/\sigma^2, \qquad \mathrm{Err}_{ij} = \mathbb{E}_0(y^0_{ij} - \widehat{y}_{ij})^2/\sigma^2,$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, q$, where the expectation $\mathbb{E}_0$ is over a new observation $y^0_{ij} \sim (\mu_{ij}, \sigma^2)$ independent of $\mathbf{Y}$. It is shown that

$$\mathbb{E}(\mathrm{Err}_{ij}) = \mathbb{E}\left\{\mathrm{err}_{ij} + 2\frac{\mathrm{cov}(\widehat{y}_{ij}, y_{ij})}{\sigma^2}\right\}. \tag{7}$$

That is, $\mathrm{err}_i$ has to be adjusted by a covariance term in order to unbiasedly estimate $\mathrm{Err}_{ij}$. Assuming normality of $e_{ij}$, Stein's lemma reveals that $\mathrm{cov}(\widehat{y}_{ij}, y_{ij}) = \sigma^2\mathbb{E}(\partial\widehat{y}_{ij}/\partial y_{ij})$, provided that $\partial\widehat{y}_{ij}/\partial y_{ij}$ exists [47]. This leads to Stein's unbiased risk estimate,

$$\gamma_{ij} = \widehat{\mathrm{Err}}_{ij} = \mathrm{err}_{ij} + 2\frac{\partial\widehat{y}_{ij}}{\partial y_{ij}}, \tag{8}$$

where we name $\gamma_{ij}$ as the SURE information score for the $(i, j)$th observation (to be elaborated later).

The SURE information criterion for estimating the total prediction error is then given by

$$\sum_{i=1}^{n}\sum_{j=1}^{q}\gamma_{ij} = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{\sigma^2} + 2\sum_{i=1}^{n}\sum_{j=1}^{q}\frac{\partial\widehat{y}_{ij}}{\partial y_{ij}}. \tag{9}$$

The first term is the training error, and it alone inevitably underestimates the true prediction error. The second term is for bias correction [22], and it is closely related to model complexity: the harder we fit the data, the stronger each $y_{ij}$ affects its own prediction, thereby increasing the complexity. From (7) and (9), the degrees of freedom of an estimation procedure $f$ is defined as $df(f) = \sum_{i=1}^{n}\sum_{j=1}^{q}\mathrm{cov}(\widehat{y}_{ij}, y_{ij})/\sigma^2$, and an unbiased estimator is

$$\widehat{df}(f) = \sum_{i=1}^{n}\sum_{j=1}^{q}\frac{\partial\widehat{y}_{ij}}{\partial y_{ij}}. \tag{10}$$

These definitions and concepts hold generally for both linear and nonlinear models.

### 3.2 Leverage score

From the SURE theory, the leverage score of an observation in an estimation procedure is defined as the self-sensitivity or self-influence of the observation,

$$l_{ij} = \frac{\partial\widehat{y}_{ij}}{\partial y_{ij}}, \qquad i = 1, \ldots, n, j = 1, \ldots, q.$$

In multivariate problems, it is also of interest to measure the overall leverage effect of each multivariate observation. We propose to use the sum of the individual leverage scores of the $q$ observations constituting a multivariate observation, i.e.,

$$l_i = \sum_{j=1}^{q} l_{ij} = \sum_{j=1}^{q}\frac{\partial\widehat{y}_{ij}}{\partial y_{ij}}, \qquad i = 1, \ldots, n,$$

as the leverage score for the multivariate observation.

The above definition is a generalization of the familiar concept of leverage value in classic linear regression models. For a linear estimation procedure, it holds that $\widehat{\mathbf{Y}} = \mathbf{HY}$ with the hat matrix $\mathbf{H} = (h_{ij})_{n\times n}$ being a function of $\mathbf{X}$ but free of $\mathbf{Y}$. Using the above definition, the scores are indeed given by the diagonal elements of $\mathbf{H}$, i.e., $l_{ij} = h_{ii}$. As such, in linear estimation, the leverage scores only characterize the properties of $\mathbf{X}$. However, in nonlinear methods such as reduced-rank regression, the leverage scores in general depend on both $\mathbf{Y}$ and $\mathbf{X}$, and they may no longer admit any explicit form. As a consequence, the computation of the leverage scores may be much more complicated. To the best of our knowledge, the usages of the individual leverage scores in reduced-rank estimation have never been investigated in the literature. On the other hand, the leverage score is closely related to the measure of model complexity, which has been extensively studied. Specifically, as shown in (10), the sum of the leverage scores provides an unbiased estimator of the degrees of freedom, $\widehat{df}(\widehat{\mathbf{Y}}) = \sum_{i=1}^{n}\sum_{j=1}^{q} l_{ij}$. Mukherjee *et al.* [33] studied the model complexity of reduced-rank estimation, and an exact unbiased estimator of the degrees of freedom was derived as a function of the estimated singular values; see, also, Yuan [52] and Candès *et al.* [10].

While Mukherjee *et al.* [33] mainly focused on the model complexity problem in reduced-rank estimation, here we have a different focus, i.e., we compute and study the properties of the leverage scores in reduced-rank estimation, for the purpose of model diagnostics. More generally, we consider the computation of $\partial\widehat{y}_{i^*j^*}/\partial y_{ij}$, for any $1 \le i^*, i < n$ and $1 \le j^*, j \le q$, which characterizes the influence of the $(i, j)$th observation on the estimation of the $(i^*, j^*)$th observation. That is, we study the possible usages of the divergence matrix $\partial\mathrm{vec}(\widehat{\mathbf{Y}})/\partial\mathrm{vec}(\mathbf{Y}) \in \mathbb{R}^{nq\times nq}$, where $\mathrm{vec}(\cdot)$ denotes the vectorization operator. Obtaining the entire divergence matrix is often unnecessary in real applications, but some of its elements are very informative in model diagnostics. In par-

ticular, the divergence matrix gives the individual leverage scores as its diagonal elements, and we will see later that it also provides the ingredients to construct and approximate other diagnostic measures such as the Cook's distance [17].

Now consider the derivation of the divergence matrix $\partial \mathrm{vec}(\widehat{\mathbf{Y}})/\partial \mathrm{vec}(\mathbf{Y})$ in reduced-rank estimation. The reduced-rank estimator $\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}(\lambda)$ and the least squares estimator $\widehat{\mathbf{Y}}_{LS}$ differ in their estimated singular values, so the former can be regarded as certain low-rank approximation of the latter. Direct computation of the divergence matrix involves taking derivatives of the singular values and singular vectors with respect to the entries of $\widehat{\mathbf{Y}}_{LS}$. In high-dimensional scenarios, such derivatives may not be well-defined as $\widehat{\mathbf{Y}}_{LS}$ may not be of full column (row) rank. Following Mukherjee *et al.* [33], a reparameterization approach can be used to avoid this difficulty.

We assume that $\widehat{\mathbf{Y}}_{LS}$ is of rank $\bar{r} = \min(r_x, q)$. This assumption generally holds even when the sample size $n$ exceeds both $p$ and $q$. Let $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{Q}\mathbf{S}^2\mathbf{Q}^{\mathrm{T}}$ be the eigen decomposition of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, i.e., $\mathbf{Q} \in \mathbb{R}^{p\times r_x}$, $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$, and $\mathbf{S} \in \mathbb{R}^{r_x\times r_x}$ is a diagonal matrix with positive diagonal elements. Then, the inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ can be written as $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-} = \mathbf{Q}\mathbf{S}^{-2}\mathbf{Q}^{\mathrm{T}}$. Define

$$\mathbf{A} = \mathbf{S}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

It then follows that $\mathbf{A} \in \mathbb{R}^{r_x\times q}$ admits an SVD of the form $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U} \in \mathbb{R}^{r_x\times \bar{r}}$, $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$, and $\mathbf{V}$, $\mathbf{D}$ are defined in (2). The matrix $\mathbf{A}$ shares the same set of singular values and right singular vectors with $\widehat{\mathbf{Y}}_{LS}$, as $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \widehat{\mathbf{Y}}_{LS}^{\mathrm{T}}\widehat{\mathbf{Y}}_{LS} = \mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. Therefore,

$$\begin{aligned}\widehat{\mathbf{Y}}(\lambda) &= \mathbf{X}\mathbf{Q}\mathbf{S}^{-1}\mathbf{A}\sum_{k=1}^{\bar{r}} s(d_k,\lambda)\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}\\ &= \mathbf{X}\mathbf{Q}\mathbf{S}^{-1}\mathbf{U}\widehat{\mathbf{D}}(\lambda)\mathbf{V}^{\mathrm{T}}\\ &= \mathbf{X}\mathbf{Q}\mathbf{S}^{-1}\widehat{\mathbf{A}}(\lambda), \qquad (11)\end{aligned}$$

where $\widehat{\mathbf{D}}(\lambda) = \mathbf{diag}\{s(d_k,\lambda)d_k; k = 1,...,\bar{r}\}$ and $\widehat{\mathbf{A}}(\lambda) = \mathbf{U}\widehat{\mathbf{D}}(\lambda)\mathbf{V}^{\mathrm{T}}$. Using the new expression of $\widehat{\mathbf{Y}}(\lambda)$ in (11) and following Mukherjee *et al.* [33], we have the following results on the explicit form of the divergence matrix.

**Theorem 1.** *Assume $r(\widehat{\mathbf{Y}}_{LS}) = \bar{r} = \min(r_x, q)$. Suppose the singular values of $\widehat{\mathbf{Y}}_{LS}$ are all distinct, i.e., $d_1 > \cdots d_{\bar{r}} > 0$. Then for the reduced-rank estimator $\widehat{\mathbf{Y}}(\lambda)$ in (4), the divergence matrix exists and is given by*

$$\begin{aligned}&\frac{\partial vec(\widehat{\mathbf{Y}}(\lambda))}{\partial vec(\mathbf{Y})}\\ &= (\mathbf{I}_q \otimes \mathbf{X}\mathbf{Q}\mathbf{S}^{-1})\left(\frac{\partial vec(\widehat{\mathbf{A}}(\lambda))}{\partial vec(\mathbf{A})}\right)(\mathbf{I}_q \otimes \mathbf{S}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}).\end{aligned}$$

*The entries of $\partial vec(\widehat{\mathbf{A}}(\lambda))/\partial vec(\mathbf{A})$ are given by*

$$\begin{aligned}\frac{\partial \widehat{\mathbf{A}}(\lambda)}{\partial a_{ij}} &= \frac{\partial \mathbf{A}}{\partial a_{ij}}\sum_{k=1}^{\hat{r}} s_k\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}} + \mathbf{A}\sum_{k=1}^{\hat{r}} s_k\frac{\partial \mathbf{v}_k}{\partial a_{ij}}\mathbf{v}_k^{\mathrm{T}}\\ &\quad + \mathbf{A}\sum_{k=1}^{\hat{r}} s_k\mathbf{v}_k\frac{\partial \mathbf{v}_k^{\mathrm{T}}}{\partial a_{ij}} + \mathbf{A}\sum_{k=1}^{\hat{r}}\frac{\partial s_k}{\partial a_{ij}}\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}\\ &= \mathbf{Z}^{(ij)}\mathbf{V}^{(\hat{r})}\mathbf{D}^{(\hat{r})-1}\widehat{\mathbf{D}}^{(\hat{r})}\mathbf{V}^{(\hat{r})\top}\\ &\quad - \mathbf{A}\sum_{k=1}^{\hat{r}}\Big\{s_k(\mathbf{A}^{\mathrm{T}}\mathbf{A} - d_k^2\mathbf{I})^{-}(\mathbf{A}^{\mathrm{T}}\mathbf{Z}^{(ij)}\\ &\quad + \mathbf{Z}^{(ij)\top}\mathbf{A})\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}\Big\}\\ &\quad - \mathbf{A}\sum_{k=1}^{\hat{r}}\Big\{s_k\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{Z}^{(ij)}\\ &\quad + \mathbf{Z}^{(ij)\top}\mathbf{A})(\mathbf{A}^{\mathrm{T}}\mathbf{A} - d_k^2\mathbf{I})^{-}\Big\}\\ &\quad + \mathbf{A}\sum_{k=1}^{\hat{r}}\Big\{s_k'\{\frac{1}{2d_k}\mathbf{v}_k^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{Z}^{(ij)}\\ &\quad + \mathbf{Z}^{(ij)\top}\mathbf{A})\mathbf{v}_k\}\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}\Big\},\end{aligned}$$

*where $\hat{r} = r(\widehat{\mathbf{Y}}(\lambda)) = \max\{k : s_k > 0.\}$, $s_k = s(d_k;\lambda\omega_k)$, $s_k' = \partial s_k/\partial d_k$, and $\mathbf{Z}^{(ij)} = \partial\mathbf{A}/\partial a_{ij}$ is an $r_x \times q$ matrix of zeros with only its $(i,j)$th entry being one.*

Under the assumptions $\bar{r} = r(\widehat{\mathbf{Y}}_{LS}) = \min(r_x, q)$ and $d_1 > \cdots > d_{\bar{r}} > 0$, the construction of the $\mathbf{A}$ matrix reduces the dimensionality of the problem, and ensures that the partial derivatives are well-defined. A sketch of proof is given in the Appendix.

All the elements required for computing the divergence matrix are obtained from a QR decomposition of the Gram matrix and the SVD of the LS estimator. Furthermore, for the computation of the leverage scores, it suffices to only compute $\partial\widehat{\mathbf{a}}_j/\partial\mathbf{a}_j$, with $\widehat{\mathbf{A}} = [\widehat{\mathbf{a}}_1,\ldots,\widehat{\mathbf{a}}_q]$ and $\mathbf{A} = [\mathbf{a}_1,\ldots,\mathbf{a}_q]$.

**Corollary 1.** *Denote $\mathbf{A}^{(k)} = (\mathbf{A}^{\mathrm{T}}\mathbf{A} - d_k\mathbf{I})^{-1}$ for $k = 1,\ldots,\hat{r}$ and $\mathbf{A}^{[j]} = [\mathbf{a}_j^{(1)},\ldots,\mathbf{a}_j^{(\hat{r})}]$ for $j = 1,\ldots,q$. Also denote $\mathbf{V} = [\mathbf{v}_1,\ldots,\mathbf{v}_{\hat{r}}] = [\tilde{\mathbf{v}}_1^{\mathrm{T}},\ldots,\tilde{\mathbf{v}}_q^{\mathrm{T}}]^{\mathrm{T}}$. Let $\mathbf{z}_j$ be an $q\times 1$ vector of zeros with only its $j$th entry being one. Then,*

$$\begin{aligned}\frac{\partial\widehat{\mathbf{y}}_j(\lambda)}{\partial\mathbf{y}_j} &= \mathbf{X}\mathbf{Q}\mathbf{S}^{-1}\left(\frac{\partial\widehat{\mathbf{a}}_j(\lambda)}{\partial\mathbf{a}_j}\right)\mathbf{S}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}, \qquad j = 1,\ldots,q,\\ \frac{\partial\widehat{\mathbf{a}}_j(\lambda)}{\partial\mathbf{a}_j} &= \tilde{\mathbf{v}}_j^{\mathrm{T}}\tilde{\mathbf{v}}_j\mathbf{I} + \mathbf{A}\left(\sum_{k=1}^{\hat{r}} s_k v_{jk}^2\mathbf{A}^{(k)} + \mathbf{A}^{[j]}\,\mathbf{diag}(\tilde{\mathbf{v}}_j)\mathbf{V}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\right)\\ &\quad + \mathbf{A}\left\{\sum_{k=1}^{\hat{r}} s_k\mathbf{v}_k\mathbf{a}_j^{(k)^{\mathrm{T}}}(v_{jk}\mathbf{A} + \mathbf{z}_j^{\mathrm{T}}\otimes\mathbf{A}\mathbf{v}_k)^{\mathrm{T}}\right\}\\ &\quad + \mathbf{A}\left(\sum_{k=1}^{\hat{r}}\frac{s_k' v_{jk}^2}{d_k}\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}\right)\mathbf{A}^{\mathrm{T}}.\end{aligned}$$

With careful manipulation of the matrix operations, the computation of the leverage scores is efficient.

### 3.3 Generalized information score

The SURE criterion in (9) is commonly used for model assessment and selection as it reflects the true predictive performance of an estimation procedure. Our primary interest here, however, is on the individual SURE information scores $\gamma_{ij}$ defined in (8), which provide an exact decomposition of the total true prediction error. Each $\gamma_{ij}$ unbiasedly estimates the true prediction error for $y_{ij}^0$, which can be regarded as the contribution of the $(i,j)$th observation to the total error. The $\gamma_{ij}$ becomes large when the $(i,j)$th residual is large and/or the $(i,j)$th leverage score is large. As such, the SURE information scores provide an intriguing way to combine model residuals and leverage scores!

Motivated by the connection between the SURE information score and information criterion, we propose a generalized information score approach for model diagnostics. Consider an information criterion of the following form

$$\mathrm{IC}(f,w) = \sum_{i=1}^{n}\sum_{j=1}^{q} g(y_{ij};\widehat{y}_{ij}) + w\sum_{i=1}^{n}\sum_{j=1}^{q}\frac{\partial\widehat{y}_{ij}}{\partial y_{ij}},$$

where $f$ denotes an estimation procedure, $g$ is a function measuring the model estimation error or the lack of fit, and $w$ controls the penalty on the model complexity. Generally, $g$ can be chosen as $-2\log L(\widehat{y}_{ij};y_{ij})$ where $L$ is the likelihood function. We now define the generalized information score (GIS) for the $(i,j)$th observation as

$$\gamma_{ij}(w) = g(y_{ij};\widehat{y}_{ij}) + w\frac{\partial\widehat{y}_{ij}}{\partial y_{ij}}, \qquad i=1,\ldots,n, j=1,\ldots,q,$$

so that $\mathrm{IC}(f,w) = \sum_{i=1}^{n}\sum_{j=1}^{q}\gamma_{ij}(w)$. When the data are contaminated, the outliers may have different influences on the model estimation. In general, if an outlier dominates the model estimation, its lack of fit term $g(y_{ij};\widehat{y}_{ij})$ tends to be small, which may lead to "masking". However, the leverage score of this observation may become large due to its dominance. On the other hand, a normal observation may happen to have relatively large lack of fit term, which may lead to "swamping". However, the leverage score of this observation may tend to be small, as it does not fit the estimated model well. Therefore, the merit of the proposed information score is to integrate the lack of fit measure and the leverage measure in a principled way, to achieve an objective assessment of the true outlying effect of each observation.

The proposed approach directly applies to the multivariate reduced-rank estimation, i.e.,

(12)
$$\gamma_{ij}(w) = \frac{(y_{ij}-\widehat{y}_{ij})^2}{\sigma^2} + wl_{ij}, \qquad i=1,\ldots,n, j=1,\ldots,q,$$

in which the leverage scores $l_{ij}$ can be obtained from Theorem 1. This decomposition applies to many well-known information criteria. To list a few, $\gamma_{ij}(2)$ gives the AIC score [1] or the SURE score, and $\gamma_{ij}(\log(nq))$ gives the BIC score [43]. Fan and Tang [20] proposed a generalized information criterion (GIC) for model selection in high-dimensional penalized regression; correspondingly, $\gamma_{ij}(\log\log(nq)\log(pq))$ gives the GIC score. We also define the score for any subset of observations as the sum of the corresponding entrywise scores,

$$\gamma_{\mathcal{A}}(w) = \sum_{(i,j)\in\mathcal{A}}\gamma_{ij}(w),$$

where $\mathcal{A}\subset\{1,\ldots,n\}\times\{1,\ldots,q\}$. In particular, $\gamma_i(w) = \sum_{j=1}^{q}\gamma_{ij}(w)$ gives the information score for the $i$th multivariate observation.

Each information score is a weighted sum of the lack of fit term and the leverage score, and it characterizes the contribution from each single observation to the overall information criterion. As such, an observation with an extreme information score can be viewed as an outlier. In practice, the error standard deviation $\sigma$ may be unknown; we have used the median absolute deviation of the residuals from a fitted model to construct a robust estimate of $\sigma$ [42], which performs well in our numerical studies. It is well-known that the information criteria mentioned above have different behaviors in model selection. For high-dimensional models, the GIC criterion [20] was shown to enjoy many desirable theoretical properties. To fix the idea, we thus mainly focus on the GIS scores computed from the GIC criterion in the sequel.

### 3.4 Cook's distance

Cook's distance is a commonly-used diagnostic measure in regression analysis [17]. For each observation, its Cook's distance is computed from examining the impact of deleting the observation on model estimation. An observation with large residual or high leverage value tends to have high impact on model estimation, which usually results in a large Cook's distance. Therefore, in practice, the observations with large values of Cook's distance may require a close examination and some special treatment.

In reduced-rank estimation, following the same spirit, we define the Cook's distance for the $(i,j)$th observation as

(13)
$$o_{ij}^{*} = \frac{\sum_{t=1}^{n}\sum_{g=1}^{q}(\widehat{y}_{tg}-\widehat{y}_{tg}^{-(ij)})^2}{r_x\mathrm{MSE}},$$

where MSE denotes the mean squared error, and $\widehat{y}_{tg}^{-(ij)}$ denotes the estimator of $y_{tg}$ when the $(i,j)$th response $y_{ij}$ is removed. That is, $\widehat{\mathbf{Y}}^{-(ij)} = (\widehat{y}_{tg}^{-(ij)})_{n\times q} = \mathbf{X}\widehat{\mathbf{C}}^{-(ij)}$ is obtained from

(14)
$$\min_{\mathbf{C}}\frac{1}{2}\|\mathcal{P}_{\Omega_{ij}}(\mathbf{Y}-\mathbf{X}\mathbf{C})\|_F^2 + \sum_{k=1}^{\min(r_x,q)}\rho(d_k(\mathbf{X}\mathbf{C}),\lambda\omega_k),$$

where $\Omega_{ij}$ denotes the index set consisting of the indices of all the entries of $\mathbf{Y}$ except for the $(i,j)$th entry, and $\mathcal{P}_{\Omega_{ij}}$

denote the orthogonal projection onto the linear space of matrices supported on $\Omega_{ij}$. The problem in (14) is closely related to the matrix completion problem [9]. It does not have an explicit solution in general and usually solving it is much more computationally intensive than solving (3). As a consequence, the exact calculation of the $nq$ many $o_{ij}$s would be prohibitively expensive.

We consider a close approximation to (13), without refitting the model by solving (14). Based on the idea of the first-order Taylor approximation, we show that $o_{ij}^*$ can be approximated by

$$o_{ij} = \frac{\sum_{t=1}^{n}\sum_{g=1}^{q}(\frac{\partial \widehat{y}_{tg}}{\partial y_{ij}})^2}{(1-l_{ij})^2 r_x \mathrm{MSE}}\widehat{e}_{ij}^2, \tag{15}$$

where $\widehat{e}_{ij} = y_{ij} - \widehat{y}_{ij}$ is the residual for the $(i,j)$th observation from solving (3). The derivation is provided in the Appendix. In the special case of least squares estimation, i.e., $\widehat{\mathbf{Y}} = \mathbf{HY}$, the above computation becomes exact so that $o_{ij}^* = o_{ij}$, and since $\partial \widehat{y}_{tg}/\partial y_{ij} = 0$ when $g \neq j$ and $\partial \widehat{y}_{tg}/\partial y_{ij} = h_{ti}$ when $g = j$, the expression of $o_{ij}$ reduces to a familiar form [17]

$$\frac{(\sum_{t=1}^{n} h_{ti}^2)\widehat{e}_{ij}^2}{(1-h_{ii})^2 r_x \mathrm{MSE}} = \frac{h_{ii}\widehat{e}_{ij}^2}{(1-h_{ii})^2 r_x \mathrm{MSE}}.$$

For detecting rowwise outliers, Cook's distance can be defined for each multivariate observation $\mathbf{y}_i$, by aggregating the entrywise Cook's distances defined above, i.e., we define

$$o_i = \sum_{j=1}^{q} o_{ij}, \qquad i = 1, \ldots, n, \tag{16}$$

as the Cook's distance for the $i$th multivariate observation.

We remark that another way of defining rowwise Cook's distance is based on the change due to deleting an entire multivariate observation,

$$o_i^* = \frac{\sum_{t=1}^{n} \|\widehat{\mathbf{y}}_t - \widehat{\mathbf{y}}_t^{-(i)}\|^2}{r_x \mathrm{MSE}},$$

where MSE denotes the mean squared error, and $\widehat{\mathbf{y}}_t^{-(i)}$ denotes the estimator of $\mathbf{y}_t$ when the $i$th response vector $\mathbf{y}_i$ is not observed. The computation requires refitting the reduced-rank model $n$ times, each time with $n-1$ multivariate observations. Based on vector-wise Taylor approximation, $o_i^*$ can be approximated by

$$\frac{\sum_{t=1}^{n}(\widehat{\mathbf{y}}_i - \mathbf{y}_i)^{\mathrm{T}}(\mathbf{I} - \frac{\partial \widehat{\mathbf{y}}_i}{\partial \mathbf{y}_i})^{-\mathrm{T}}(\frac{\partial \widehat{\mathbf{y}}_t}{\partial \mathbf{y}_i})^{\mathrm{T}}(\frac{\partial \widehat{\mathbf{y}}_t}{\partial \mathbf{y}_i})(\mathbf{I} - \frac{\partial \widehat{\mathbf{y}}_i}{\partial \mathbf{y}_i})^{-1}(\widehat{\mathbf{y}}_i - \mathbf{y}_i)}{r_x \mathrm{MSE}}.$$

However, this definition is harder to compute comparing to (16) and we also find that it is numerically less stable. We thus mainly use the simpler definition given in (16) for detecting multivariate outliers.

The Cook's distance of an observation given in (15) not only relates to its self-influence or leverage score, but also relates to how it influences the prediction of other observations. This is very different from linear models, in which the Cook's distance only depends on self-influences. Although all the needed quantities for computing Cook's distances admit explicit forms as presented in Theorem 1, the computation involves the entire divergence matrix, which can be much more computationally demanding than the computation of the leverage scores or the information scores. As such, the usage of (15) may be limited in large-scale problems, i.e, when both $r_x$ and $q$ are large.

# 4. SIMULATION

## 4.1 Setups

We consider a high-dimensional setup, with $n = 50$, $p = 1000$, $q = 30$, $r^* = 3$ and $r_x = 30$. The design matrix $\mathbf{X}$ is generated as $\mathbf{X} = \mathbf{X}_1\mathbf{X}_2$, where all the entries of $\mathbf{X}_1 \in \mathbb{R}^{n\times r_x}$ are i.i.d samples from $\mathrm{N}(0,1)$, and $\mathbf{X}_2 \in \mathbb{R}^{r_x\times p}$ is constructed by generating its $r_x$ rows as i.i.d. samples from $\mathrm{N}(\mathbf{0}, \Sigma)$, where $\Sigma$ is the covariance matrix with diagonal elements 1 and off-diagonal elements 0.5. The coefficient matrix $\mathbf{C}$ is generated as $\mathbf{C} = \mathbf{C}_1\mathbf{C}_2^{\mathrm{T}}$, where $\mathbf{C}_1 \in \mathbb{R}^{p\times r^*}$, $\mathbf{C}_2 \in \mathbb{R}^{q\times r^*}$ and all entries in $\mathbf{C}_1$ and $\mathbf{C}_2$ are i.i.d. samples from $\mathrm{N}(0,1)$. The rows of the noise matrix $\mathbf{E}$ are generated as i.i.d. samples from $\mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The $\sigma^2$ is set to control the signal to noise ratio (SNR), defined as $d_{r^*}(\mathbf{XC})/\|\mathbf{E}\|_F$, so that SNR $\in \{0.5, 1, 1.5\}$. The outlier-free response matrix $\mathbf{Y}$ is then generated as $\mathbf{Y} = \mathbf{XC} + \mathbf{E}$.

We then create some additive outliers in $\mathbf{Y}$. Without loss of generality, suppose the first $n_o = 5$ rows are the outlier rows. Specifically, the $j$th entry in any outlier row of $\mathbf{Y}$ is either added or subtracted $4\widehat{\sigma}_j^2$, where $\widehat{\sigma}_j^2$ is the standard deviation of the $j$th row of $\mathbf{XC}$. To make the problem more challenging, we also replace each of the first $n_l$ rows of $\mathbf{X}$ with the same value 10, where $n_l \in \{0, 2\}$, so there may be some highly influential outliers. As such, the low rank signal $\mathbf{XC}$ is contaminated by both random errors and gross outliers.

We mainly consider the rank-penalized reduced-rank regression estimator from (5), as a prototype of the general class of reduced-rank estimators in (4). In practice, the model rank is usually determined based on either cross validation or some information criterion. However, with the contaminated data, the estimation of the initial rank can be unstable. To better understand the behavior of the diagnostic measures under potential rank misspecification, we consider three scenarios, i.e., $\widehat{r} = 1, 3, 5$. We consider using mean squared residuals (RES), leverage score (LEV), Cook's distance (CD), and GIS score (GIS) for model diagnostics and outlier detection. Their behaviors in both rowwise outlier detection and entrywise outlier detection are investigated. The experiment is replicated 100 times in each model setup and parameter setting.

To make the comparison fair, each method is used to identify 16% of data points who yield the largest diagnostic

*Table 1. Simulation: results for rowwise outlier detection, with SNR = 0.5. For each method, 8 multivariate observations with the largest diagnostic measures are treated as outliers. The best possible FPR is 37.5%*

| Rank | | FNR | FPR | CDR | RK | PRE | FNR | FPR | CDR | RK | PRE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #leverage= 0 | | | | | #leverage= 2 | | | | |
| | RES | 22.2% | 51.4% | 25.0% | 1.00 | 68.03 | 46.6% | 66.6% | 0.0% | 1.01 | 85.09 |
| | LEV | 15.6% | 47.3% | 26.0% | 1.00 | 56.69 | 1.8% | 38.6% | 75.0% | 1.00 | 27.94 |
| 1 | CD | 0.0% | 37.5% | 64.0% | 1.01 | 29.65 | 0.8% | 38.0% | 67.0% | 1.02 | 28.12 |
| | GIS | 0.0% | 37.5% | 82.0% | 1.01 | 29.63 | 0.0% | 37.5% | 74.0% | 1.01 | 27.53 |
| | ORE | 0.0% | 0.0% | 100.0% | 1.02 | 28.16 | 0.0% | 0.0% | 100.0% | 1.02 | 26.21 |
| | RES | 19.2% | 49.5% | 7.0% | 1.02 | 55.61 | 38.2% | 61.4% | 3.0% | 1.18 | 55.60 |
| | LEV | 30.4% | 56.5% | 1.0% | 1.01 | 77.65 | 25.6% | 53.5% | 2.0% | 1.00 | 57.88 |
| 3 | CD | 5.8% | 41.1% | 34.0% | 1.02 | 34.56 | 15.4% | 47.1% | 22.0% | 1.13 | 36.18 |
| | GIS | 0.4% | 37.8% | 46.0% | 1.01 | 28.53 | 0.0% | 37.5% | 66.0% | 1.01 | 28.51 |
| | ORE | 0.0% | 0.0% | 100.0% | 1.03 | 26.97 | 0.0% | 0.0% | 100.0% | 1.01 | 27.51 |
| | RES | 22.2% | 51.4% | 5.0% | 1.06 | 60.05 | 37.0% | 60.6% | 3.0% | 1.15 | 57.44 |
| | LEV | 49.4% | 68.4% | 0.0% | 1.02 | 119.26 | 38.8% | 61.8% | 0.0% | 1.02 | 88.85 |
| 5 | CD | 28.4% | 55.3% | 5.0% | 1.04 | 73.69 | 41.6% | 63.5% | 2.0% | 1.10 | 71.10 |
| | GIS | 7.2% | 42.0% | 14.0% | 1.03 | 37.17 | 10.8% | 44.3% | 9.0% | 1.06 | 36.96 |
| | ORE | 0.0% | 0.0% | 100.0% | 1.02 | 27.42 | 0.0% | 0.0% | 100.0% | 1.01 | 28.07 |

values. That is, we identify 8 multivariate observation rows using rowwise detection methods and identify 240 data entries using entrywise detection methods. For rowwise outlier detection, we report the false negative rate (FNR), the false positive rate (FPR), and the correct detection rate (CDR). Specifically, FNR is defined as the percentage of unidentified outliers among all the true outliers, FPR is defined as the percentage of incorrectly identified observations among all the detected outliers. We note that the best possible FPR in our setup is 37.5%, as each method always identifies 16% of all the observations but at most 62.5% of them are true outliers. When the 5 data rows with the largest diagnostic measures are exactly the 5 true outlier rows, we call it a correct detection; CDR is then defined as the proportion of time the correct detection occurs among all simulation runs. For entrywise outlier selection, we only report the FNR rates for simplicity.

Once the outliers are identified, a common practice is to discard the outliers and use the rest of the data to refit the model. In case of entrywise outliers, reduced-rank estimation becomes a matrix completion problem, which no longer admits explicit solution. As fitting robust reduced rank models is not the focus of this paper, we refer interested readers to Candès *et al.* [8] and She and Chen [46]. On the other hand, in case of rowwise outliers, the estimation method remains the same. Therefore, for rowwise detection, we report the average prediction error (PRED) and rank estimate (RK) from the refitted model, with the detected outliers being discarded and the rank being selected by GIC. We also report the performance of an oracle estimator (ORE), which is based on discarding only the true outliers.

## 4.2 Results

Tables 1–3 and Figures 1–2 show the results for rowwise outlier detection. Consider first the impact of rank specification. In the presence of gross outliers, as the specified rank gets larger, the task of outlier detection becomes harder in general. This is because the outliers are more likely to be falsely accommodated in a model of higher rank.

The naive approach of outlier detection, i.e., using the residuals, performs poorly in all scenarios. This is mainly due to the masking effect: some of the gross outliers may dominate the reduced-rank estimation, so their residuals can be quite small. LEV is designed to capture such masking effects, and it behaves better than RES when the specified rank is low, but its performance becomes much worse when the rank is overspecified. These results clearly demonstrate that there is a tradeoff between the residuals and the leverage scores: when an outlier dominates the estimation, it may have a relatively small residual but a large leverage score; on the other hand, if its outlying effect is not accounted, the outlier may lead to a relatively large residual but a small leverage score. Therefore, it is critical to combine both RES or LEV for outlier detection.

Indeed, both CID and GIS perform much better than RES and LEV across all scenarios. When the specified rank is low, they both archive very low false negative rates and prediction error rates, comparable to those of the oracle estimator. Overall, GIS is more robust than CID against rank specification and the presence of high-leverage outliers. In Figures 1 and 2, the boxplots of the diagnostic measures for the five outliers and another five randomly selected observations are shown. The outliers have much larger leverage scores and residuals than the good observations in general. However, the distributions of the RES or LEV scores of the outliers have some visible overlap with those of the good observations. In contrast, the distributions of CD or GIS scores of the outliers are much more separable from those of the good observations. We have also experimented with

*Table 2. Simulation: results for rowwise outlier detection, with SNR = 1*

| Rank | | FNR | FPR | CDR | RK | PRE | FNR | FPR | CDR | RK | PRE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #leverage= 0 | | | | | #leverage= 2 | | | | |
| 1 | RES | 21.8% | 51.1% | 16.0% | 1.08 | 68.30 | 42.8% | 64.3% | 0.0% | 1.16 | 71.99 |
| | LEV | 9.8% | 43.6% | 32.0% | 1.15 | 31.49 | 1.0% | 38.1% | 71.0% | 1.17 | 22.82 |
| | CD | 0.0% | 37.5% | 69.0% | 1.18 | 22.68 | 2.8% | 39.3% | 64.0% | 1.17 | 25.04 |
| | GIS | 0.2% | 37.6% | 70.0% | 1.20 | 22.98 | 0.6% | 37.9% | 55.0% | 1.18 | 24.55 |
| | ORE | 0.0% | 0.0% | 100.0% | 1.30 | 20.71 | 0.0% | 0.0% | 100.0% | 1.24 | 20.60 |
| 3 | RES | 21.2% | 50.8% | 8.0% | 1.20 | 58.16 | 35.8% | 59.9% | 2.0% | 1.51 | 40.66 |
| | LEV | 32.2% | 57.6% | 2.0% | 1.01 | 71.47 | 23.2% | 52.0% | 1.0% | 1.06 | 52.96 |
| | CD | 8.0% | 42.5% | 28.0% | 1.26 | 33.80 | 17.0% | 48.1% | 20.0% | 1.36 | 29.31 |
| | GIS | 1.2% | 38.3% | 40.0% | 1.22 | 22.78 | 1.8% | 38.6% | 48.0% | 1.26 | 22.99 |
| | ORE | 0.0% | 0.0% | 100.0% | 1.41 | 18.50 | 0.0% | 0.0% | 100.0% | 1.36 | 19.91 |
| 5 | RES | 20.0% | 50.0% | 3.0% | 1.15 | 49.91 | 34.6% | 59.1% | 1.0% | 1.43 | 43.22 |
| | LEV | 47.6% | 67.3% | 0.0% | 1.02 | 115.33 | 39.8% | 62.4% | 0.0% | 1.06 | 97.87 |
| | CD | 19.6% | 49.8% | 6.0% | 1.15 | 52.80 | 34.8% | 59.3% | 6.0% | 1.44 | 48.56 |
| | GIS | 4.8% | 40.5% | 16.0% | 1.21 | 27.43 | 15.6% | 47.3% | 19.0% | 1.34 | 30.47 |
| | ORE | 0.0% | 0.0% | 100.0% | 1.41 | 19.74 | 0.0% | 0.0% | 100.0% | 1.32 | 20.52 |

*Table 3. Simulation: results for rowwise outlier detection, with SNR = 1.5*

| Rank | | FNR | FPR | CDR | RK | PRE | FNR | FPR | CDR | RK | PRE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #leverage= 0 | | | | | #leverage= 2 | | | | |
| 1 | RES | 23.2% | 52.0% | 20.0% | 1.73 | 61.87 | 45.2% | 65.8% | 1.0% | 1.78 | 68.79 |
| | LEV | 11.0% | 44.4% | 33.0% | 2.18 | 22.90 | 2.6% | 39.1% | 68.0% | 2.72 | 5.87 |
| | CD | 0.2% | 37.6% | 72.0% | 2.96 | 1.78 | 4.2% | 40.1% | 50.0% | 3.02 | 2.74 |
| | GIS | 0.2% | 37.6% | 70.0% | 2.92 | 2.01 | 0.8% | 38.0% | 40.0% | 2.98 | 1.14 |
| | ORE | 0.0% | 0.0% | 100.0% | 3.00 | 0.56 | 0.0% | 0.0% | 100.0% | 3.00 | 0.50 |
| 3 | RES | 20.2% | 50.1% | 8.0% | 1.85 | 45.91 | 36.0% | 60.0% | 2.0% | 2.35 | 27.73 |
| | LEV | 27.8% | 54.9% | 0.0% | 1.40 | 61.95 | 21.8% | 51.1% | 3.0% | 1.49 | 44.00 |
| | CD | 5.6% | 41.0% | 26.0% | 2.45 | 17.01 | 18.8% | 49.3% | 14.0% | 2.74 | 15.33 |
| | GIS | 0.6% | 37.9% | 30.0% | 2.93 | 3.88 | 1.4% | 38.4% | 33.0% | 2.88 | 2.65 |
| | ORE | 0.0% | 0.0% | 100.0% | 3.00 | 0.53 | 0.0% | 0.0% | 100.0% | 3.00 | 0.51 |
| 5 | RES | 17.4% | 48.4% | 9.0% | 1.81 | 40.90 | 33.0% | 58.1% | 4.0% | 2.58 | 24.79 |
| | LEV | 49.0% | 68.1% | 0.0% | 1.08 | 122.61 | 37.4% | 60.9% | 0.0% | 1.16 | 75.46 |
| | CD | 13.6% | 46.0% | 7.0% | 2.07 | 34.76 | 31.8% | 57.4% | 7.0% | 2.55 | 27.56 |
| | GIS | 5.4% | 40.9% | 21.0% | 2.48 | 13.88 | 14.2% | 46.4% | 11.0% | 2.90 | 6.92 |
| | ORE | 0.0% | 0.0% | 100.0% | 3.00 | 0.54 | 0.0% | 0.0% | 100.0% | 3.00 | 0.48 |

discarding 20% or 24% observations, and the conclusions are similar, and hence these results are omitted.

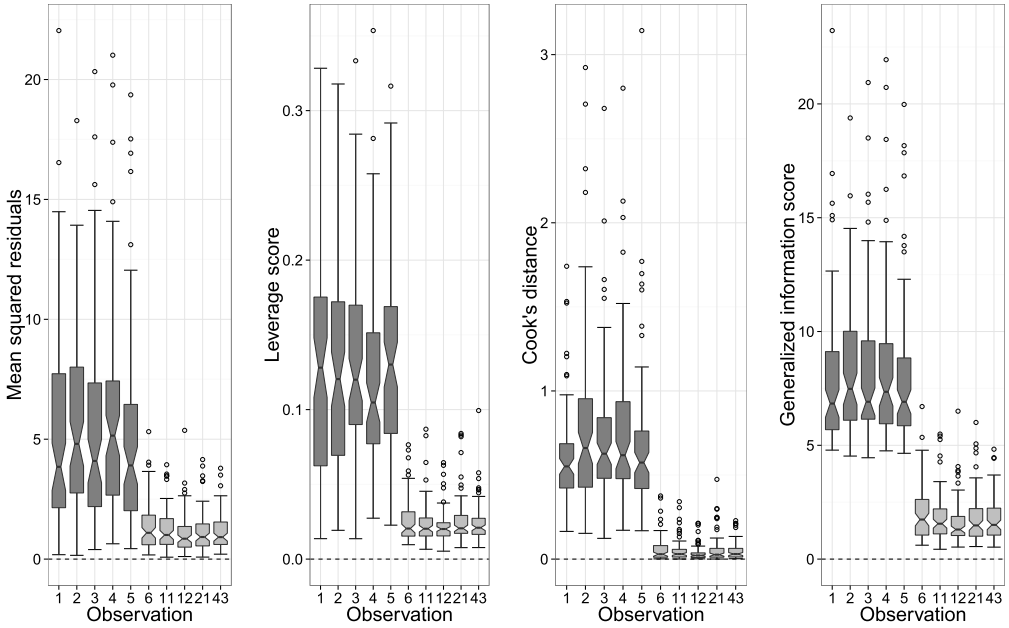
Table 4 reports the results from conducting entrywise outlier detection using the same simulation setups. Overall, the false negative rates computed from entrywise observations are much higher than those computed from rowwise observations. One plausible reason is that the outlying effect of each entry is much weaker comparing to that of an entire outlying row. Besides, there could be a large variability among the entrywise observations, leading to a large variability among their diagnostic measures. Nevertheless, the entrywise LEV scores are still informative, and both CD and GIS consistently outperform RES, especially when the model rank is not overspecified. We note that the presence of gross outliers usually makes the estimated rank smaller; in our example, when GIC is used to select the rank with the contaminated data, the estimated rank is almost always 1.

In summary, RES and LEV both provide important information on the observations, but in general neither of them alone yields satisfactory performance in outlier detection. CD and GIS provide two principled ways of jointly using the residuals and the leverage scores, and the latter is more robust against model under-fitting or over-fitting. The proposed model diagnostic tools indeed provide a convenient way for anomaly detection in reduced-rank estimation.

# 5. APPLICATIONS

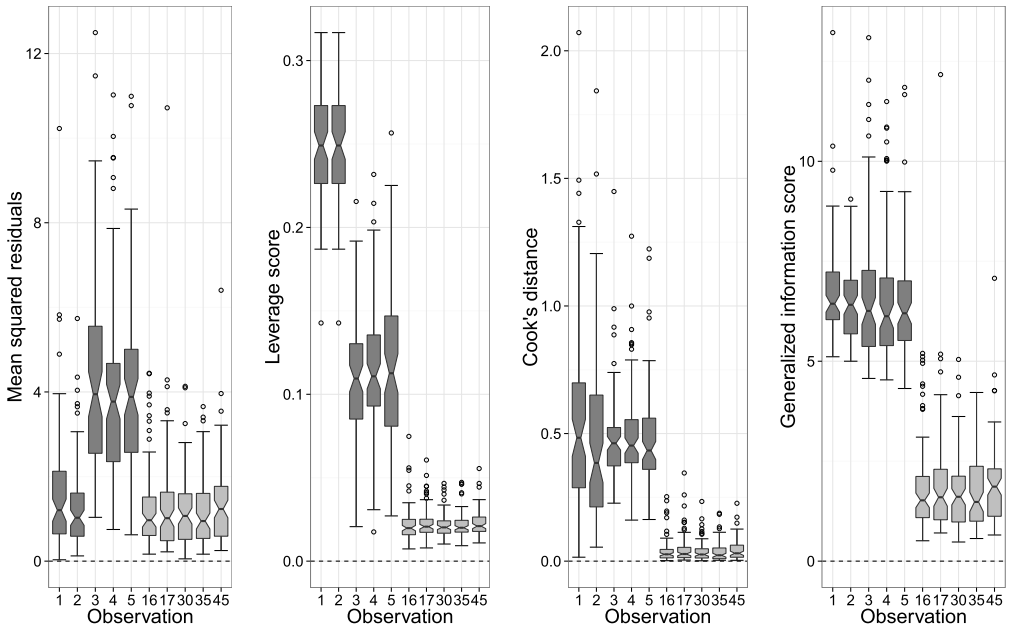
## 5.1 Handwritten digits data

We consider an example in pattern recognition. The data are images of handwritten digits from the MNIST database [28]. There are over 60,000 samples of handwritten digits from approximately 250 writers. The digits have been size-




(a) Residual (b) Leverage (c) Cook's D (d) GIS

*Figure 1. Simulation: boxplots of the rowwise diagnostic measures, with SNR* $= 1$, $\widehat{r} = 1$, *and no high-leverage point. Each subfigure consists of ten boxplots for the 5 outliers and the other 5 randomly selected good observations.*




(a) Residual (b) Leverage (c) Cook's D (d) GIS

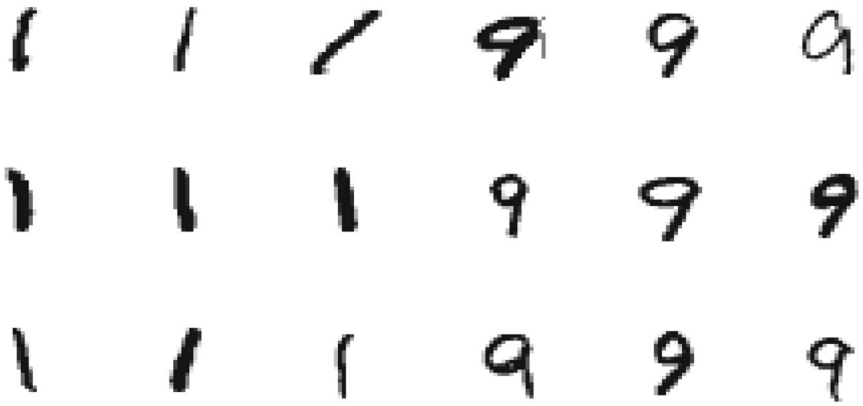
*Figure 2. Simulation: boxplots of the rowwise diagnostic measures, with SNR* $= 1$, $\widehat{r} = 1$, *and 2 high-leverage points. Each subfigure consists of ten boxplots for the 5 outliers and the other 5 randomly selected good observations.*

normalized and centered in a fixed-size image. Each sample contains $28 \times 28$ numerical grey levels computed from its black and white original image, which is then vectorized to form a data vector of length $q = 28^2 = 784$.

There are, in total, 6,742 samples of digit one and 5,949 samples of digit nine. We randomly select 200 samples of digit one and 10 samples of digit 9, and represent the resulting data as a matrix $\mathbf{Y}$ of dimension $n = 210$ and $q = 784$.

*Table 4. Simulation: false negative rates in entrywise outlier detection*

| SNR | Rank | RES | LEV | CD | GIS | RES | LEV | CD | GIS |
|---|---|---|---|---|---|---|---|---|---|
| | | #leverage= 0 | | | | #leverge= 2 | | | |
| 0.5 | 1 | 73.6% | 53.0% | 61.1% | 58.9% | 81.1% | 44.3% | 62.0% | 56.8% |
| | 3 | 70.7% | 67.2% | 65.5% | 61.9% | 76.1% | 59.3% | 67.5% | 58.7% |
| | 5 | 67.2% | 72.2% | 63.4% | 61.5% | 72.6% | 63.6% | 67.2% | 63.2% |
| 1 | 1 | 72.8% | 52.2% | 61.2% | 60.4% | 79.8% | 43.4% | 61.7% | 57.2% |
| | 3 | 65.5% | 65.5% | 61.4% | 58.6% | 71.5% | 59.3% | 64.2% | 56.8% |
| | 5 | 62.9% | 70.3% | 59.8% | 58.6% | 68.0% | 63.8% | 63.9% | 62.6% |
| 1.5 | 1 | 71.3% | 52.5% | 60.4% | 60.4% | 79.5% | 44.0% | 62.0% | 58.4% |
| | 3 | 64.6% | 64.8% | 60.1% | 57.8% | 71.3% | 59.4% | 65.2% | 59.5% |
| | 5 | 60.0% | 71.8% | 57.7% | 56.2% | 67.7% | 62.0% | 63.6% | 62.4% |



(a) Digit 1 (b) Digit 9

*Figure 3. Handwritten digits data: a typical set of images of digit 1 and digit 9, randomly selected from the MNIST dataset.*

As such, $\mathbf{Y}$ is expected to be well approximated by a low-rank matrix up to a few rowwise outliers. It is then interesting to conduct a low-rank matrix approximation analysis of $\mathbf{Y}$, and examine whether the 10 images of digit nine can be identified using the developed model diagnostic tools. Here this matrix approximation problem is a special case of the reduced-rank regression problem, corresponding to $n = p$ and $\mathbf{X} = \mathbf{I}_p$.

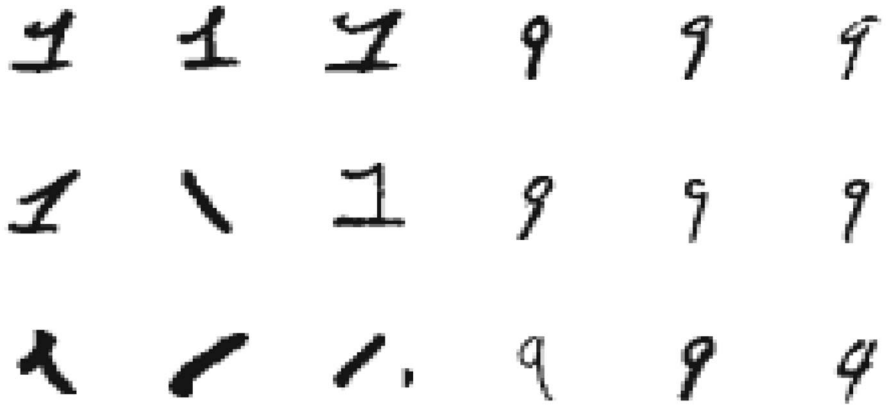
We consider three methods for outlier detection, i.e., RES, LEV and GIS. (CD is not included as its performance is similar to that of GIS but its computation is not as scalable). Each method is used to identify 15 samples which give the largest diagnostic values. We then count how many images of digit 9 are undetected, and compute the false negative rate (FNR) as the percentage of missed ones out of the 10 images of digit 9. This random data generation process and the subsequent reduced-rank estimation procedure are repeated 300 times.

Figure 3 shows a typical set of images of digit 1 and digit 9, randomly selected from the MNIST samples. In general, the two types of patterns are quite distinctive, so the images of digit 9 can be considered as severe outliers. On the other hand, there is a large variability among the hand-writing patterns. In the MNIST dataset, there exists patterns of either 1 or 9 that are abnormal, which can potentially blur the distinction between the two digits.

The average FNR rates are 19.1%, 27.9% and 13.3%, for using RES, LEV and GIS. The results are consistent with the findings from the simulation study. Although using the leverage scores alone does not lead to good outlier detection performance, the scores indeed provide valuable information complementary to that provided by the residuals. By combining the information from both the residuals and the leverage scores, GIS has the best detection performance of all three methods. In Figure 4, we also show a set of images of digit 9 that are not selected as outliers and a set of images of digit 1 that are selected as outliers. It is clear that these hand-writing patterns are different from the typical patterns of the digits. Interestingly, the selected patterns of digit 1 tend to have extra strokes, and the unselected patterns of digit 9 tend to be very "skinny".

## 5.2 U.S. macroeconomic data

We consider a U.S. macroeconomic dataset consisting of monthly observations on 132 U.S. macroeconomic variables between 1959 and 2003 [48]. These time series cover 14 dif-



(a) Selected images of digit 1 (b) Not selected images of digit 9

*Figure 4. Handwritten digits data: a set of images of digit 9 that are not selected as outliers and a set of images of digit 1 that are selected as outliers.*

ferent categories of economic measurements, such as real output and income, employment and hours, consumption, real inventories, stock prices, exchange rates, interest rates, etc. In our analysis, five time series with missing values are removed, and the rest series are transformed individually by taking logarithms and/or differencing, resulting in $q = 127$ time series that are approximately stationary.

Following Stock and Watson [48], we consider a vector autoregressive model (VAR) of order $h = 6$ to analyze the data, so that the macroeconomic conditions at each month are modeled by what happened in the past six months. A VAR model can be expressed as a multivariate regression form $\mathbf{Y} = \mathbf{XC} + \mathbf{E}$, in which $\mathbf{Y}$ consists of the observed multivariate time series, and $\mathbf{X}$ consists of the lagged time series up to lag $h$. Accordingly, a common way of conducting VAR model estimation is via the least squares method, possibly coupled with certain regularization in dealing with high dimensionality [21, 30, 24]. In particular, upon assuming $\mathbf{C}$ is of low rank, the model becomes a reduced-rank VAR model [39, 26], which is widely used in analyzing multivariate time series data. Here in our analysis, we use the RRR method and the model rank is chosen as $r = 3$. Since the model dimensions $p = 762$, $q = 127$ are relatively large comparing to the sample size $n = 533$, applying reduced-rank estimation dramatically reduces the VAR model complexity.

Generally, the usage of the reduced-rank VAR model analysis is to capture the main patterns and common factors that drive the macroeconomic developments in U.S. over time. However, the U.S. economy encountered several large disturbances between 1953 and 2003, during which the economic activities may no longer follow their general behavior; see, e.g., the report on U.S. Business Cycle Expansions and Contractions, published by the National Bureau of Economic Research [35]. These disturbances may distort the estimation of the general macroeconomic behavior in the VAR analysis. It is thus interesting to see whether performing model diagnostics of the fitted VAR model can reveal these periods of economic disturbances.

We plot the series of GIS scores, the LEV scores and the RES values from the fitted VAR model in Figure 5. The points above the dashed line in each panel are the 5% most extreme values of the corresponding diagnostic measure. The regions of dark points indicate four historical major economic recessions occurred during the period, as documented by the National Bureau of Economic Research (NBER). By NBER, "a recession is a significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in real GDP, real income, employment, industrial production, and wholesale-retail sales." The first one was from early 1960 to February 1961, occurred after the Federal Reserve began raising interest rates in 1959. The second was from November 1973 to March 1975, which was caused by quadrupled oil prices and dramatically increased government spending due to the Vietnam War. Both the 1973 oil crisis and the 1973–1974 stock market crash happened during this recession. The third recession was in early 1980s, happened after the 1979 energy crisis caused by the sharply increased oil price due to Iranian Revolution. The decade of 1990 was one of the longest periods of growth in U.S. history, but this long growth was brought to the end by the collapse of the "dot-com bubble" and the September 11th terrorism attacks, causing the fourth recession from March 2001 to the end of 2001.

As seen from Figure 5, all the four recessions are clearly visible from the GIS series. Although the start of the first recession was recorded as April 1960 by NBER, all three series show that there were already some anomalies in late 1959. This may be explained by the fact that the main cause of this recession was the raising interest rates by the Fed-

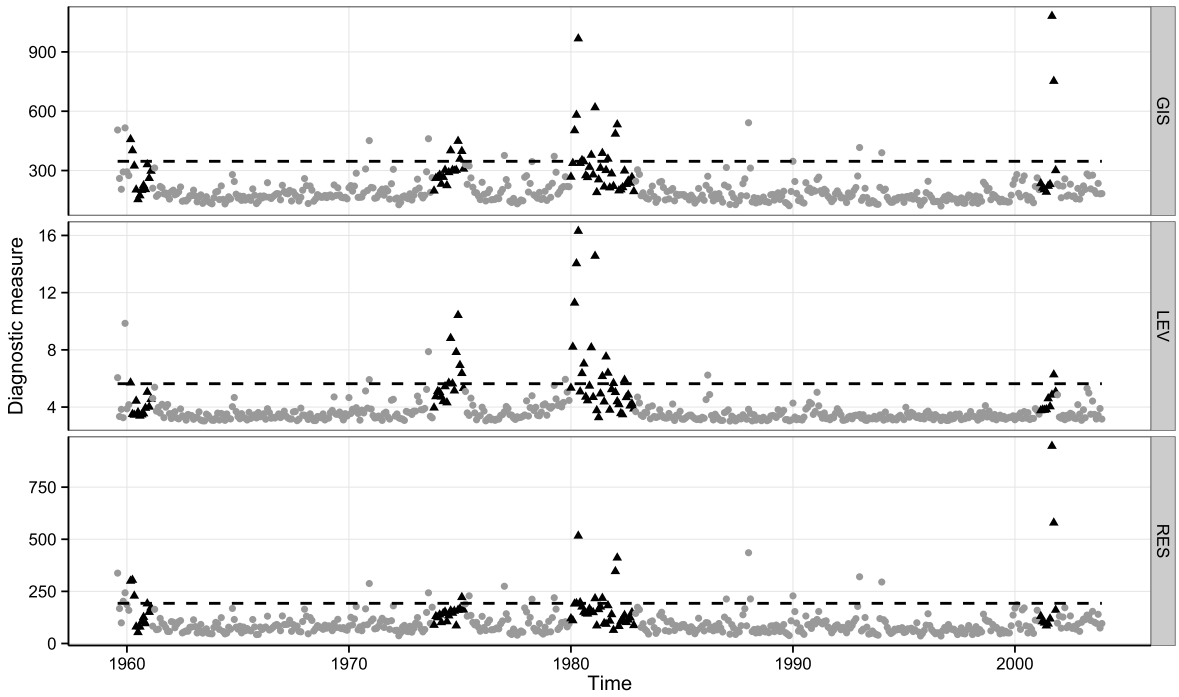



*Figure 5. U.S. macroeconomic data: series of GIS, LEV and RES from the fitted VAR model, from the top to the bottom panel. The points above the dashed line in each panel are the 5% most extreme values of the corresponding diagnostic measure. The regions of dark triangular points indicate four major economic recessions occurred during the period 1959–2003.*

eral Reserve in 1959. In contrast to the GIS series, neither the LEV series nor the RES series alone can adequately reveal all the recessions. In particular, the second recession is barely seen from the RES series, while the fourth recession is barely seen from the LEV series. These interesting patterns show that the 1973–1975 recession has dominated the VAR estimation, resulting in small residuals and large leverage scores; the early 2000 recession does not fit the VAR model well, resulting in large residuals and small leverage scores. We have also tested models with different ranks, i.e., $r = 1$ and $r = 5$, corresponding to either under-fitting or over-fitting. While LEV and RES show a clear trade-off in different models, the four recession periods are consistently revealed by GIS. This example again demonstrates the importance and effectiveness of utilizing the residuals and the leverage scores in anomaly detection.

## 6. DISCUSSION

We propose a set of model diagnostic tools for the nonlinear and nonrobust high-dimensional reduced-rank estimation methods. The explicit form of the leverage score is derived based on the SURE framework and its importance in outlier detection is demonstrated. In particular, the proposed generalized information score approach provides a principled way for incorporating the residuals and the leverage scores for anomaly detection. The approach is applicable in many other high-dimensional estimation problems, whenever the leverage scores can be readily obtained. For example, the leverage scores of a lasso estimator can be obtained as the diagonal elements of the projection matrix constructed from only the selected predictors [57, 49].

There are many directions for future research. It is pressing to thoroughly study the theoretical properties of the proposed leverage scores and the generalized information scores. An important problem is to study how to choose a proper weight to combine the sum of squared residuals and the leverage scores for ensuring the outlier detection consistency. This problem is closely related to the problem of choosing an appropriate penalty rate on model complexity in an information criterion in order to achieve consistent model selection [20]. Studying the probability distributions of the diagnostic measures may also be fruitful, which can provide formal guidelines and cutoff points for declaring outliers. The proposed diagnostic tools, including the leverage score and the Cook's distance, only provide one particular way to extend these classic quantities/concepts to handle high-dimensional multivariate problems; it is interesting to investigate other alternative extensions that may be more suitable in multivariate problems. In real applications, the model diagnostic approach and the joint estimation approach can be utilized together in many ways. For example, the proposed approaches can be used to screen out apparent outliers, and the leverage scores or information scores can

be used to construct certain adaptive weights to facilitate robust estimation [46].

## ACKNOWLEDGMENTS

The author is grateful to the referees and the editors for their valuable comments and suggestions.

## APPENDIX

### Sketch of proof of Theorem 1

We acknowledge that the main steps for establishing the results in Theorem 1 are from Mukherjee *et al.* [33], although the focus there was on getting a simple expression of the sum of the leverage scores.

From $\mathbf{A} = \mathbf{S}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ and $\widehat{\mathbf{Y}}(\lambda) = \mathbf{XQS}^{-1}\mathbf{A}(\lambda)$,

$$\begin{aligned}\frac{\partial \mathrm{vec}(\widehat{\mathbf{Y}}(\lambda))}{\partial \mathrm{vec}(\mathbf{Y})} &= (\mathbf{I}_q \otimes \mathbf{XQS}^{-1})\left(\frac{\partial \mathrm{vec}(\widehat{\mathbf{A}}(\lambda))}{\partial \mathrm{vec}(\mathbf{Y})}\right)\\ &= (\mathbf{I}_q \otimes \mathbf{XQS}^{-1})\left(\frac{\partial \mathrm{vec}(\widehat{\mathbf{A}}(\lambda))}{\partial \mathrm{vec}(\mathbf{A})}\right)\left(\frac{\partial \mathrm{vec}(\mathbf{A})}{\partial \mathrm{vec}(\mathbf{Y})}\right)\\ &= (\mathbf{I}_q \otimes \mathbf{XQS}^{-1})\left(\frac{\partial \mathrm{vec}(\widehat{\mathbf{A}}(\lambda))}{\partial \mathrm{vec}\mathbf{A})}\right)\\ &\quad \times (\mathbf{I}_q \otimes \mathbf{S}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}).\end{aligned}$$

From Theorem 1 of Mukherjee *et al.* [33],

$$\begin{aligned}\frac{\partial \mathbf{v}_k}{\partial a_{ij}} &= -(\mathbf{A}^{\mathrm{T}}\mathbf{A} - d_k^2\mathbf{I})^{-}(\mathbf{A}^{\mathrm{T}}\mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)^{\mathrm{T}}}\mathbf{A})\mathbf{v}_k,\\ \frac{\partial d_k}{\partial a_{ij}} &= \frac{1}{2d_k}\mathbf{v}_k^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{Z}^{(ij)} + \mathbf{Z}^{(ij)^{\mathrm{T}}}\mathbf{A})\mathbf{v}_k.\end{aligned}$$

Since $\widehat{\mathbf{A}}(\lambda) = \mathbf{A}\sum_{k=1}^{\hat{r}} s_k\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}$, the results in Theorem 1 can then be obtained by applying the chain rule,

$$\begin{aligned}\frac{\partial \widehat{\mathbf{A}}(\lambda)}{\partial a_{ij}} &= \frac{\partial \mathbf{A}}{\partial a_{ij}}\sum_{k=1}^{\hat{r}} s_k\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}} + \mathbf{A}\sum_{k=1}^{\hat{r}} s_k\frac{\partial \mathbf{v}_k}{\partial a_{ij}}\mathbf{v}_k^{\mathrm{T}}\\ &\quad + \mathbf{A}\sum_{k=1}^{\hat{r}} s_k\mathbf{v}_k\frac{\partial \mathbf{v}_k^{\mathrm{T}}}{\partial a_{ij}} + \mathbf{A}\sum_{k=1}^{\hat{r}}\frac{\partial s_k}{\partial a_{ij}}\mathbf{v}_k\mathbf{v}_k^{\mathrm{T}}.\end{aligned}$$

### Derivation of the approximated Cook's distance

We write the reduced-rank estimator from solving (3) as $\widehat{\mathbf{Y}}(y_{ij}; i = 1, \ldots, n, j = 1, \ldots, q)$, to emphasize that it is a function of $\mathbf{Y}$. It follows that

$$\widehat{\mathbf{Y}}^{-(ij)} = \mathbf{X}\widehat{\mathbf{C}}^{-(ij)} = \widehat{\mathbf{Y}}(y_{11}, \ldots, \widehat{y}_{ij}^{-(ij)}, \ldots, y_{nq}).$$

That is, if $\widehat{y}_{ij}^{-(ij)}$ were known, $\widehat{\mathbf{Y}}^{-(ij)}$ can be obtained by solving (3) with $y_{ij}$ replaced by $\widehat{y}_{ij}^{-(ij)}$. This observation relates $\widehat{y}_{tg}^{-(i,j)}$ to $\widehat{y}_{tg}$, for any $t = 1, \ldots, n$ and $g = 1, \ldots, q$. Based on Taylor expansion,

$$\widehat{y}_{tg}^{-(ij)} \approx \widehat{y}_{tg} + \frac{\partial \widehat{y}_{tg}}{\partial y_{ij}}(\widehat{y}_{ij}^{-(ij)} - y_{ij}),$$

and it follows that

$$(\widehat{y}_{tg}^{-(ij)} - \widehat{y}_{tg})^2 \approx \left(\frac{\partial \widehat{y}_{tg}}{\partial y_{ij}}\right)^2(\widehat{y}_{ij}^{-(ij)} - y_{ij})^2. \tag{17}$$

On the other hand,

$$\widehat{y}_{ij}^{-(ij)} \approx \widehat{y}_{ij} + \frac{\partial \widehat{y}_{ij}}{\partial y_{ij}}(\widehat{y}_{ij}^{-(ij)} - y_{ij}) = \widehat{y}_{ij} + l_{ij}(\widehat{y}_{ij}^{-(ij)} - y_{ij}),$$

which implies that

$$y_{ij}^{-(ij)} - y_{ij} \approx \frac{1}{1 - l_{ij}}(\widehat{y}_{ij} - y_{ij}). \tag{18}$$

Combining (17) and (18), we see that $o_{ij}^*$ can be approximated by (15).

Similarly, for approximating the rowwise Cook's distance in (13), the result follows from

$$\begin{aligned}&(\widehat{\mathbf{y}}_t^{-(i)} - \widehat{\mathbf{y}}_t)^{\mathrm{T}}(\widehat{\mathbf{y}}_t^{-(i)} - \widehat{\mathbf{y}}_t)\\ \approx&(\widehat{\mathbf{y}}_i^{-(i)} - \mathbf{y}_i)^{\mathrm{T}}\left(\frac{\partial \widehat{\mathbf{y}}_t}{\partial \mathbf{y}_i}\right)^{\mathrm{T}}\left(\frac{\partial \widehat{\mathbf{y}}_t}{\partial \mathbf{y}_i}\right)(\widehat{\mathbf{y}}_i^{-(i)} - \mathbf{y}_i),\end{aligned}$$

and

$$\widehat{\mathbf{y}}_i^{-(i)} - \mathbf{y}_i \approx \left(\mathbf{I} - \frac{\partial \widehat{\mathbf{y}}_i}{\partial \mathbf{y}_i}\right)^{-1}(\widehat{\mathbf{y}}_i - \mathbf{y}_i).$$

*Received 2 April 2015*

## REFERENCES

[1] Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723. MR0423716

[2] Anderson, T. W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, **22**, 327–351. MR0042664

[3] Brown, P. J. and Zidek, J. V. (1982) Multivariate regression shrinkage estimators with unknown covariance matrix. *Scandinavian Journal of Statistics*, **9**, 209–215. MR0695283

[4] Buades, A., Coll, B. and Morel, J. M. (2005) A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, **4**, 490–530. MR2162865

[5] Bunea, F., She, Y. and Wegkamp, M. (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics*, **39**, 1282–1309. MR2816355

[6] Bunea, F., She, Y. and Wegkamp, M. (2012) Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *Annals of Statistics*, **40**, 2359–2388. MR3097606

[7] Cai, J.-F., Candès, E. J. and Shen, Z. (2010) A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, **20**, 1956–1982. MR2600248

[8] Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011) Robust principal component analysis? *Journal of the ACM*, **58**, 1–37. MR2811000

[9] Candès, E. J. and Recht, B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717–772. MR2565240

[10] Candès, E. J., Sing-Long, C. and Trzasko, J. D. (2013) Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, **61**, 4643–4657. MR3105401

[11] Chen, K., Chan, K.-S. and Stenseth, N. C. (2012) Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B*, **74**, 203–221. MR2899860

[12] Chen, K., Chan, K.-S. and Stenseth, N. C. (2014) Source-sink reconstruction through regularized multicomponent regression analysis–with application to assessing whether north sea cod larvae contributed to local fjord cod in skagerrak. *Journal of the American Statistical Association*, **109**, 560–573. MR3223733

[13] Chen, K., Dong, H. and Chan, K.-S. (2013) Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, **100**, 901–920. MR3142340

[14] Chen, L. and Huang, J. Z. (2012) Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, **107**, 1533–1545. MR3036414

[15] Choi, S., Hoffman, E. A., Wenzel, S. E., Tawhai, M. H., Yin, Y., Castro, M. and Lin, C.-L. (2013) Registration-based assessment of regional lung function via volumetric ct images of normal subjects vs. severe asthmatics. *Journal of Applied Physiology*, **115**, 730–742.

[16] Choi, S., Hoffmann, E. A., Wenzel, S. E., Castro, M., Fain, S. B., Jarjour, N. N., Schiebler, M. L., Chen, K. and Lin, C.-L. (2015) Quantitative assessment of multiscale structural and functional alterations in asthmatic populations. *Journal of Applied Physiology*, **118**, 1286–1298.

[17] Cook, R. D. (1977) Detection of influential observation in linear regression. *Technometrics*, **19**, 15–18. MR0436478

[18] Diaz-Garcia, J. A. and Gonzalez-Farias, G. (2004) A note on the Cook's distance. *Journal of Statistical Planning and Inference*, **120**, 119–136. MR2026486

[19] Efron, B. (2004) The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, **99**, 619–642. MR2090899

[20] Fan, Y. and Tang, C. Y. (2013) Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B*, **75**, 531–552. MR3065478

[21] Hamilton, J. (1994) *Time series analysis*, vol. 2. Cambridge Univ Press. MR1278033

[22] Hastie, T. J., Tibshirani, R. J. and Friedman, J. H. (2008) *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer. MR2722294

[23] Hoerl, A. E. and Kennard, R. W. (2000) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **42**.

[24] Hsu, N.-J., Hung, H.-L. and Chang, Y.-M. (2008) Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis*, **52**, 3645–3657. MR2427370

[25] Izenman, A. J. (1975) Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, **5**, 248–264. MR0373179

[26] Izenman, A. J. (2008) *Mordern multivariate statistical techniques*. Springer. MR2445017

[27] Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011) Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, **39**, 2302–2329. MR2906869

[28] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (2001) Gradient-based learning applied to document recognition. In *Intelligent Signal Processing* (eds. Haykin, S. and Kosko, B.), 306–351. IEEE Press.

[29] Lu, Z., Monteiro, R. D. C. and Yuan, M. (2012) Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Math. Program.*, **131**, 163–194. MR2886145

[30] Lütkepohl, H. (2006) *New introduction to multiple time series analysis*. New York: Springer.

[31] Ma, X., Xiao, L. and Wong, W. H. (2014) Learning regulatory programs by threshold svd regression. *Proceedings of the National Academy of Sciences*, **111**, 15675–15680.

[32] McDonald, G. C. and Galarneau, D. I. (1975) A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, **70**, 407–416.

[33] Mukherjee, A., Chen, K., Wang, N. and Zhu, J. (2015) On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, **102**, 457–477. MR3371016

[34] Mukherjee, A. and Zhu, J. (2011) Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining*, **4**, 612–622. MR2862506

[35] NBER (2009) US Business Cycle Expansions and Contractions. *National Bureau of Economic Research*.

[36] Negahban, S. and Wainwright, M. J. (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, **39**, 1069–1097. MR2816348

[37] Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2011) Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, **39**, 1–47. MR2797839

[38] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R. and Wang, P. (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, **4**, 53–77. MR2758084

[39] Reinsel, G. C. and Velu, P. (1998) *Multivariate reduced-rank regression: Theory and applications*. New York: Springer. MR1719704

[40] Rothman, A. J., Levina, E. and Zhu, J. (2010) Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, **19**, 947–962. MR2791263

[41] Rousseeuw, P. J., Aelst, S. V., Driessen, K. V., Professor, P. J. R. I., Van, S. and Belgium, A. F. (2000) Robust multivariate regression. *Technometrics*, **46**, 293–305. MR2082499

[42] Rousseeuw, P. J. and Croux, C. (1993) Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**. MR1245360

[43] Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464. MR0468014

[44] She, Y. (2009) Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Statist.*, **3**, 384–415. MR2501318

[45] She, Y. (2013) Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface*, **6**, 197–209. MR3066685

[46] She, Y. and Chen, K. (2015) Robust reduced rank regression. *arXiv e-prints arXiv:1509.03938*.

[47] Stein, C. M. (1981) Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**. MR0630098

[48] Stock, J. H. and Watson, M. W. (2005) Implications of dynamic factor models for var analysis. Working Paper 11467, *National Bureau of Economic Research*.

[49] Tibshirani, R. J. and Taylor, J. (2012) Degrees of freedom in lasso problems. *Ann. Statist.*, **40**, 1198–1232. MR2985948

[50] Witten, D. M., Tibshirani, R. J. and Hastie, T. J. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

[51] Wright, J., Ganesh, A., Rao, S., Peng, Y. and Ma, Y. (2009) Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in Neural Information Processing Systems 22* (eds. Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I. and Culotta, A.), 2080–2088.

[52] Yuan, M. (2011) Degrees of freedom in low rank matrix estimation. *Unpublished Manuscript*.

[53] Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007) Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B*, **69**, 329–346. MR2323756

[54] Zhu, H., Ibrahim, J. G. and Cho, H. (2012) Perturbation and scaled cook's distance. *Annals of Statistics*, **40**, 785–811. MR2933666

[55] Zhu, H., Khondker, Z., Lu, Z. and Ibrahim, J. G. (2014) Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, **109**, 977–990. MR3265670

[56] Zhu, L., Zhu, R. and Song, S. (2008) Diagnostic checking for multivariate regression models. *Journal of Multivariate Analysis*, **99**, 1841–1859. MR2466539

[57] Zou, H., Hastie, T. J. and Tibshirani, R. J. (2007) On the degree of freedom of the lasso. *Annals of Statistics*, **35**, 2173–2192. MR2363967

Kun Chen
Department of Statistics
University of Connecticut
215 Glenbrook Rd. U-4120
Storrs, CT 06269-4120
USA
E-mail address: kun.chen@uconn.edu