Biometrika (2017), **104**, 3, *pp.* 633–647
Printed in Great Britain
doi: 10.1093/biomet/asx032
Advance Access publication 12 July 2017

Robust reduced-rank regression

BY Y. SHE

Department of Statistics, Florida State University, 117 N. Woodward Avenue, Tallahassee, Florida 32306, U.S.A

yshe@stat.fsu.edu

AND K. CHEN

Department of Statistics, University of Connecticut, 215 Glenbrook Road U-4120, Storrs, Connecticut 06269, U.S.A.

kun.chen@uconn.edu

SUMMARY

In high-dimensional multivariate regression problems, enforcing low rank in the coefficient matrix offers effective dimension reduction, which greatly facilitates parameter estimation and model interpretation. However, commonly used reduced-rank methods are sensitive to data corruption, as the low-rank dependence structure between response variables and predictors is easily distorted by outliers. We propose a robust reduced-rank regression approach for joint modelling and outlier detection. The problem is formulated as a regularized multivariate regression with a sparse mean-shift parameterization, which generalizes and unifies some popular robust multivariate methods. An efficient thresholding-based iterative procedure is developed for optimization. We show that the algorithm is guaranteed to converge and that the coordinatewise minimum point produced is statistically accurate under regularity conditions. Our theoretical investigations focus on non-asymptotic robust analysis, demonstrating that joint rank reduction and outlier detection leads to improved prediction accuracy. In particular, we show that redescending ψ-functions can essentially attain the minimax optimal error rate, and in some less challenging problems convex regularization guarantees the same low error rate. The performance of the proposed method is examined through simulation studies and real-data examples.

Some key words: Low-rank matrix approximation; Non-asymptotic analysis; Robust estimation; Sparsity.

1. INTRODUCTION

Given n observations of m response variables and p predictors, denoted by $y_i \in \mathbb{R}^m$ and $x_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$, we consider the multivariate regression model

$$Y = XB^* + \mathcal{E},$$

where $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $X = (x_1, \ldots, x_n)^{\mathrm{T}}$, $B^* \in \mathbb{R}^{p \times m}$ is an unknown coefficient matrix, and $\mathcal{E} = (e_1, \ldots, e_n)^{\mathrm{T}} \in \mathbb{R}^{n \times m}$ is a random error matrix. Such a high-dimensional multivariate problem, in which both p and m may be comparable to or even exceed the sample size n, has drawn increasing attention in both applied and theoretical statistics.

Conventional least-squares linear regression ignores the multivariate nature of the problem and may fail when p is large relative to n. Dimension reduction holds the key to characterizing the dependence between responses and predictors in a parsimonious way. Reduced-rank regression (Anderson, 1951; Izenman, 1975) achieves this by restricting the rank of the coefficient matrix, i.e., by solving the problem

$$\min_{B\in\mathbb{R}^{p\times m}} \mathrm{tr}\{(Y - XB)\Gamma(Y - XB)^{\mathrm{T}}\} \quad \text{subject to } r(B) \leqslant r, \tag{1}$$

where $\mathrm{tr}(\cdot)$ and $r(\cdot)$ denote trace and rank, and Γ is a prespecified positive-definite weighting matrix (Reinsel & Velu, 1998). The ranks are typically much smaller than m and p. A global solution to (1) can be obtained explicitly. See Reinsel & Velu (1998) for a comprehensive account of reduced-rank regression under the classical large-n asymptotic regime. Finite-sample theories on rank selection and estimation accuracy of the penalized form of reduced-rank regression were developed by Bunea et al. (2011). The nuclear norm and Schatten p-norms can also be used to promote sparsity of the singular values of B or XB; see Yuan et al. (2007), Koltchinskii et al. (2011), Rohde & Tsybakov (2011), Agarwal et al. (2012), Foygel et al. (2012) and Chen et al. (2013), among others. Reduced-rank regression is closely connected with principal component analysis, canonical correlation analysis, partial least squares, matrix completion, and many other multivariate methods (Izenman, 2008).

Although reduced-rank regression can substantially reduce the number of free parameters in multivariate problems, it is extremely sensitive to outliers, which are bound to occur; so in real-world data analysis, the low-rank structure could easily be masked or distorted. This is even more serious in high-dimensional or big-data applications. For example, in cancer genetics, multivariate regression is commonly used to explore the associations between genotypical and phenotypical characteristics (Vounou et al., 2010), where employing rank regularization can help to reveal latent regulatory pathways linking the two sets of variables; but pathway recovery should not be distorted by abnormal samples or subjects. As another example, financial time series, even after stationarity transformation, often contain anomalies or have heavier tails than those of a normal distribution, which may jeopardize the recovery of common market behaviours and asset return forecasting; see § 3 in the Supplementary Material.

In this work, we deem explicit outlier detection to be as important as robust low-rank estimation. Indeed, the reduced-rank component may not be of direct interest in some applications, as it often represents common background information shared across the response variables, while capturing unusual changes or jumps is helpful. The robustification of low-rank matrix estimation is nontrivial. A straightforward idea might be to use a robust loss function ρ in place of the squared error loss in (1), leading to

$$\min_{B} \sum_{i=1}^{n} \rho\{\|\Gamma^{1/2}(y_i - B^{\mathrm{T}}x_i)\|_2\} \quad \text{subject to } r(B) \leqslant r, \tag{2}$$

but such an estimator may be difficult to compute. To the best of our knowledge, even when ρ is Huber's loss function (Huber, 1981) there is no algorithm for solving (2), let alone when it involves nonconvex losses, which are known to be more effective in dealing with multiple gross outliers with possibly high leverage values. Another motivation is that non-asymptotic theory on the topic is limited. Classical robust analysis, ignoring the low-rank constraint, deals with either deterministic worst-case studies or large-n asymptotics with p and m held fixed, which may not meet modern needs.

We propose a novel robust reduced-rank regression method for concurrent robust modelling and outlier identification. We explicitly introduce a sparse mean-shift outlier component and formulate a shrinkage multivariate regression in place of (2), where p and/or m can be much larger than n. The proposed robust reduced-rank regression provides a general framework and includes M-estimation and principal component pursuit (Huber, 1981; Hampel et al., 2005; Zhou et al., 2010; Candès et al., 2011). All the techniques developed in this work apply to high-dimensional sparse regression with a single response. In § 2 we show that low-rank estimation can be ruined by a single rogue point, and propose a robust reduced-rank estimation framework. A universal connection between the proposed robustification and conventional M-estimation is established, regardless of the size of p, m or n. In § 3 we conduct finite-sample theoretical studies of the proposed robust estimators, with the aim of extending classical robust analysis to multivariate data with possibly large p and/or m. A computational algorithm developed in § 4 is easy to implement and leads to a coordinatewise minimum point with theoretical guarantees. Applications to real data are demonstrated in § 5. All the proofs and results of simulation studies, as well as a financial example, are given in the Supplementary Material.

The following notation will be used throughout the paper. We denote by $\mathbb{N}$ the set of natural numbers. We use $a \wedge b$ to denote $\min(a, b)$ and e to denote the Euler constant. Let $[n] = \{1, \ldots, n\}$. Given any matrix A, $\mathcal{P}_A$ denotes the orthogonal projection matrix onto the range of A, i.e., $A(A^{\mathrm{T}}A)^{-}A^{\mathrm{T}}$, where $^{-}$ stands for the Moore–Penrose pseudo-inverse. When there is no ambiguity, we also use $\mathcal{P}_A$ to denote the column space of A. Let $\|A\|_{\mathrm{F}}$ denote the Frobenius norm and $\|A\|_2$ the spectral norm, and let $\|A\|_0 = \|\mathrm{vec}(A)\|_0 = |\{(i,j) : A(i,j) \neq 0\}|$ with $|\cdot|$ denoting the cardinality of the enclosed set. For $A = (a_1 \ \ldots \ a_n)^{\mathrm{T}} \in \mathbb{R}^{n \times m}$, $\|A\|_{2,1} = \sum_{i=1}^{n} \|a_i\|_2$ and $\|A\|_{2,0} = \sum_{i=1}^{n} 1_{\|\alpha_i\| \neq 0}$, which gives the number of nonzero rows of A. Given $\mathcal{J} \subset [n]$, we often denote $\sum_{i \in \mathcal{J}} \|a_i\|_2$ by $\|A_{\mathcal{J}}\|_{2,1}$. Threshold functions are defined as follows.

Definition 1 (Threshold function). *A threshold function is a real-valued function* $\Theta(t; \lambda)$ *defined for* $-\infty < t < \infty$ *and* $0 \leqslant \lambda < \infty$ *such that* (i) $\Theta(-t; \lambda) = -\Theta(t; \lambda)$*;* (ii) $\Theta(t; \lambda) \leqslant \Theta(t'; \lambda)$ *for* $t \leqslant t'$*;* (iii) $\lim_{t \to \infty} \Theta(t; \lambda) = \infty$*; and* (iv) $0 \leqslant \Theta(t; \lambda) \leqslant t$ *for* $0 \leqslant t < \infty$.

Definition 2 (Multivariate threshold function). *Given any* Θ, $\vec{\Theta}$ *is defined for any vector* $a \in \mathbb{R}^m$ *by* $\vec{\Theta}(a; \lambda) = a\Theta(\|a\|_2; \lambda)/\|a\|_2$ *for* $a \neq 0$ *and* 0 *otherwise. For any matrix* $A = (a_1 \ \ldots \ a_n)^{\mathrm{T}} \in \mathbb{R}^{n \times m}$, $\vec{\Theta}(A; \lambda) = \{\vec{\Theta}(a_1; \lambda) \ \ldots \ \vec{\Theta}(a_n; \lambda)\}^{\mathrm{T}}$.

2. Robust reduced-rank regression

2·1. *Motivation*

Although reduced-rank regression is associated with a highly nonconvex problem (1), a global minimizer $\hat{B}$ can be obtained in explicit form. Given any r such that $1 \leqslant r \leqslant \min(m, q)$ with $q = r(X)$,

$$\hat{B}(r) = \mathcal{R}(X, Y, \Gamma, r) = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}Y\Gamma^{1/2}\mathcal{P}_{V(X,Y,\Gamma,r)}\Gamma^{-1/2}, \tag{3}$$

where $V(X, Y, \Gamma, r)$ is formed by the leading r eigenvectors of $\Gamma^{1/2}Y^{\mathrm{T}}\mathcal{P}_X Y\Gamma^{1/2}$; see, e.g., Reinsel & Velu (1998) for a detailed justification. When $\Gamma = I$, we abbreviate $\mathcal{R}(X, Y, I, r)$ to $\mathcal{R}(X, Y, r)$. The reduced-rank regression estimator is denoted by $\hat{B}(r)$ to emphasize its dependence on the regularization parameter.

Outliers are unavoidable in real data. We define the finite-sample breakdown point for an arbitrary estimator $\hat{B}$ in the spirit of Donoho & Huber (1983): given finite data (X, Y, Γ) and an estimator $\hat{B}(X, Y, \Gamma)$, its breakdown point is

$$\epsilon^*(\hat{B}) = \frac{1}{n} \min \left\{ k \in \mathbb{N} \cup \{0\} : \sup_{\tilde{Y} \in \mathbb{R}^{n \times m} : \|\tilde{Y} - Y\|_0 \leqslant k} \|X\hat{B}(X, \tilde{Y}, \Gamma)\|_{\mathrm{F}} = +\infty \right\}.$$

In addition to the reduced-rank regression estimator $\hat{B}(r)$, we take into account a general low-rank estimator obtained by imposing a singular value penalty

$$\hat{B}(\lambda) \in \arg\min_{B} \frac{1}{2} \mathrm{tr}\{(Y - XB)\Gamma(Y - XB)^{\mathrm{T}}\} + \sum_{s=1}^{p \wedge m} P(\sigma_s^{B\Gamma^{1/2}}; \lambda). \tag{4}$$

Here λ is a regularization parameter and the $\sigma_s^{B\Gamma^{1/2}}$ are the singular values of $B\Gamma^{1/2}$. The penalty P is constructed from an arbitrary thresholding rule $\Theta(\cdot\,; \lambda)$ by

$$P(t; \lambda) - P(0; \lambda) = P_\Theta(t; \lambda) + q(t; \lambda), \quad P_\Theta(t; \lambda) = \int_0^{|t|} [\sup\{s : \Theta(s; \lambda) \leqslant u\} - u]\, \mathrm{d}u, \tag{5}$$

for some nonnegative $q(\cdot\,; \lambda)$ satisfying $q\{\Theta(s; \lambda); \lambda\} = 0$ for all $s \in \mathbb{R}$.

THEOREM 1. *Given any finite (X, Y, Γ) and $r \geqslant 1$ with Γ positive definite and $X \neq 0$, let $\hat{B}(r)$ be a reduced-rank regression estimator which solves (1). Then its finite-sample breakdown point is exactly $1/n$. Furthermore, for any $\hat{B}(\lambda)$ as in (4), $\epsilon^*\{\hat{B}(\lambda)\} = 1/n$ still holds for any finite value of λ.*

The result indicates that a single outlier can completely ruin low-rank matrix estimation, whether one applies a rank constraint or, say, a Schatten p-norm penalty. This limits the use of ordinary rank reduction in big-data applications. Because with the low-rank constraint, directly using a robust loss function as in (2) may result in nontrivial computational and theoretical challenges, we will apply a novel additive robustification, motivated by She & Owen (2011).

2·2. *The additive framework*

We introduce a multivariate mean-shift regression model to explicitly encompass outliers:

$$Y = XB^* + C^* + \mathcal{E}, \tag{6}$$

where $B^* \in \mathbb{R}^{p \times m}$ gives the matrix of coefficients, $C^* \in \mathbb{R}^{n \times m}$ describes the outlying effects on Y, and $\mathcal{E} \in \mathbb{R}^{n \times m}$ has independently and identically distributed rows following $N(0, \Sigma)$. Obviously, this leads to an overparameterized model, so we must regularize the unknown matrices appropriately. We assume that B^* has low rank and C^* is a sparse matrix with only a few nonzero entries because outliers are inconsistent with the majority of the data. Given a positive-definite weighting matrix Γ, we propose the robust reduced-rank regression problem

$$\min_{B,C} \frac{1}{2} \mathrm{tr}\{(Y - XB - C)\Gamma(Y - XB - C)^{\mathrm{T}}\} + P(C; \lambda) \quad \text{subject to } r(B) \leqslant r. \tag{7}$$

Here $P(\cdot\,;\lambda)$ is a sparsity-promoting penalty function with λ to adjust the amount of shrinkage, but it can also be a constraint, such as in (12). The following form of P can handle elementwise outliers:

$$P(C;\lambda) = \sum_{i=1}^{n}\sum_{k=1}^{m} P(|c_{i,k}|;\lambda). \tag{8}$$

It is more common in robust statistics to assume outlying samples, or outlying rows in (Y, X), which corresponds to

$$P(C;\lambda) = \sum_{i=1}^{n} P(\|c_i\|_2;\lambda),$$

where c_i^{T} is the ith row vector of C. Unless otherwise specified, we consider row-wise outliers. But all our algorithms and analyses can handle elementwise outliers after simple modification.

In the literature on reduced-rank regression, it is common to regard the weighting matrix Γ as known (Reinsel & Velu, 1998; Yuan et al., 2007; Izenman, 2008). The choice of Γ is flexible and is usually based on a pilot covariance estimate $\hat{\Sigma}$. For example, it can be $\hat{\Sigma}^{-1}$ when $\hat{\Sigma}$ is nonsingular, or a regularized version $(\hat{\Sigma} + \delta I)^{-1}$ for some $\delta > 0$. Although it sounds intriguing to consider jointly estimating the high-dimensional mean and the even higher-dimensional covariance matrix in the presence of outliers, this is beyond the scope of the present paper. When a reliable estimate of Σ is unavailable, a standard practice in finance and econometric forecasting is to reduce Γ to a diagonal matrix or, equivalently, an identity matrix after robustly scaling the response variables. For ease of presentation, we shall take Γ to be the identity matrix, unless otherwise noted, and mainly focus on the following robust reduced-rank regression criterion:

$$\min_{B,C} \frac{1}{2}\|Y - XB - C\|_{\mathrm{F}}^2 + P(C;\lambda) \quad \text{subject to } r(B) \leqslant r. \tag{9}$$

We show that the proposed additive outlier characterization indeed comes with a robustness guarantee and, interestingly, generalizes M-estimation to the multivariate rank-deficient setting. We write $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $C = (c_1, \ldots, c_n)^{\mathrm{T}}$.

THEOREM 2. (i) *Suppose that $\Theta(\cdot\,;\lambda)$ is an arbitrary thresholding rule satisfying Definition 1, and let P be any penalty associated with Θ through (5). Consider*

$$\min_{B,C} \frac{1}{2}\|Y - XB - C\|_{\mathrm{F}}^2 + \sum_{i=1}^{n} P(\|c_i\|_2;\lambda) \quad \textit{subject to } r(B) \leqslant r. \tag{10}$$

For any fixed B, a globally optimal solution for C is $C(B) = \vec{\Theta}(Y - XB;\lambda)$. By profiling out C with $C(B)$, (10) can be expressed as an optimization problem with respect to B only, and it is equivalent to the robust M-estimation problem

$$\min_{B} \sum_{i=1}^{n} \rho(\|y_i - B^{\mathrm{T}}x_i\|_2;\lambda) \quad \textit{subject to } r(B) \leqslant r, \tag{11}$$

where the robust loss function ρ is given by

$$\rho(t;\lambda) = \int_0^{|t|} \psi(u;\lambda)\,\mathrm{d}u, \quad \psi(t;\lambda) = t - \Theta(t;\lambda).$$

(ii) *Given $\varrho \in \{0, 1, \ldots, n\}$, consider*

$$\min_{B,C} \frac{1}{2}\|Y - XB - C\|_{\mathrm{F}}^2 \quad \textit{subject to} \;\; r(B) \leqslant r, \;\; \|C\|_{2,0} \leqslant \varrho. \tag{12}$$

Similarly, (12), after profiling out C, can be expressed as an optimization problem with respect to B only, and it is equivalent to the rank-constrained trimmed least-squares problem

$$\min_{B} \frac{1}{2}\sum_{i=1}^{n-\varrho} r_{(i)} \quad \textit{subject to} \;\; r(B) \leqslant r, \;\; r_i = \|y_i - B^{\mathrm{T}}x_i\|_2,$$

where $r_{(1)}, \ldots, r_{(n)}$ are the order statistics of $r_1, \ldots, r_n$ satisfying $|r_{(1)}| \leqslant \cdots \leqslant |r_{(n)}|$.

Remark 1. Theorem 2 connects P to ρ through Θ. As is well known, changing the squared error loss to a robust loss amounts to designing a set of multiplicative weights for $y_i - B^{\mathrm{T}}x_i$ $(i = 1, \ldots, n)$. Our additive robustification achieves the same robustness but leaves the original loss function untouched. The connection is also valid in the case of elementwise outliers, with P and ρ applied in an elementwise manner. In fact, the identity constructed in Lemma 2 of the Supplementary Material,

$$\frac{1}{2}\{r - \Theta(r;\lambda)\}^2 + P_\Theta\{\Theta(r;\lambda);\lambda\} = \int_0^{|r|} \psi(t;\lambda)\,\mathrm{d}t \quad (r \in \mathbb{R}),$$

implies that the equivalence holds much more generally, with B subject to an arbitrary constraint or penalty, regardless of the number of response variables and the number of predictors. This extends the main result of She & Owen (2011) to multiple-response models with p possibly larger than n.

Remark 2. Theorem 2 holds for all thresholding rules, and commonly used convex and nonconvex penalties are all covered by (5). For example, the convex group ℓ_1 penalty $\lambda\sum\|c_i\|_2$ is associated with the soft-thresholding $\Theta_S(s;\lambda) = \mathrm{sgn}(s)(|s| - \lambda)_+$. The group ℓ_0 penalty $(\lambda^2/2)\sum_{i=1}^n 1_{\|c_i\|_2 \neq 0}$ can be obtained from (5) with the hard-thresholding $\Theta_H(s;\lambda) = s1_{|s|>\lambda}$ and with $q(t;\lambda) = 0{\cdot}5(\lambda - |t|)^2 1_{0<|t|<\lambda}$. Our Θ-P coupling framework also covers ℓ_p for $0 < p < 1$, the smoothly clipped absolute deviation penalty (Fan & Li, 2001), the minimax concave penalty (Zhang, 2010a), and the capped ℓ_1 penalty (Zhang, 2010b) as particular instances; see She (2012).

Remark 3. The universal link between (10) and (11) provides insight into the choice of regularization. It is easy to verify that the ℓ_1-norm penalty as commonly used in variable selection leads to Huber's loss, which is prone to masking and swamping and may fail with even moderately leveraged outliers occurring. To handle gross outliers, redescending ψ-functions are often advocated, which amounts to using nonconvex penalties in (10). For example, Hampel's three-part ψ (Hampel et al., 2005) can be shown to yield Fan and Li's smoothly clipped absolute deviation penalty; the skipped mean ψ corresponds to the exact ℓ_0 penalty; and rank-constrained least trimmed squares can be rephrased as the ℓ_0-constrained form in (12). Our approach not only provides a unified way to robustify low-rank matrix estimation but also facilitates theoretical analysis and computation of reduced-rank M-estimators in high dimensions.

2·3. *Connections and extensions*

Before we dive into theoretical study, it is worth pointing out some connections and extensions of the proposed framework. First, one can set Γ equal to the inverse covariance matrix of the response variables to perform robust canonical correlation analysis; see Reinsel & Velu (1998). Although we mainly focus on the rank-constrained form, there is no difficulty in extending our discussion to

$$\min_{B,C} \frac{1}{2}\|Y - XB - C\|_{\mathrm{F}}^2 + \sum_{s=1}^{p\wedge m} P_B(\sigma_s^B; \lambda_B) + P_C(C; \lambda_C), \tag{13}$$

where the σ_s^B are the singular values of B, and P_B and P_C are sparsity-inducing penalties.

Our robust reduced-rank regression subsumes a special but important case, $Y = B+C+\mathcal{E}$. This problem is perhaps less challenging than its supervised counterpart, but has wide applications in computer vision and machine learning (Wright et al., 2009; Candès et al., 2011).

Finally, our method can be extended to reduced-rank generalized linear models; see, for example, Yee & Hastie (2003) and She (2013) for computational details. In these scenarios, directly robustifying the loss can be messy, but a sparse outlier term can always be introduced without altering the form of the given loss, so that many algorithms designed for fitting ordinary generalized linear models can be seamlessly applied.

3. Non-asymptotic robust analysis

Theorem 2 provides robustness and some helpful intuition for the proposed method, but it might not be enough from a theoretical point of view. For example, can one justify the need for robustification in estimating a matrix of low rank? Is using redescending ψ-functions still preferable in rank-deficient settings? Unlike in traditional robust analysis, we cannot assume an infinite sample size and a fixed number of predictors or response variables, because p and/or m can be much larger than n in modern applications. Conducting non-asymptotic robust analysis would be desirable. The finite-sample results in this section contribute to this type of robust analysis.

For simplicity we assume that the model is given by $Y = XB^* + C^* + \mathcal{E}$, where $\mathcal{E}$ has independent and identically distributed $N(0, \sigma^2)$ entries, and consider the robust reduced-rank regression problem defined in (9). The noise distribution can be more general. For example, in all of the following theorems except Theorem 5, $\mathcal{E}$ can be sub-Gaussian. Given an estimator $(\hat{B}, \hat{C})$, we focus on its prediction accuracy measured by $M(\hat{B} - B^*, \hat{C} - C^*)$, where

$$M(B, C) = \|XB + C\|_{\mathrm{F}}^2.$$

This predictive learning perspective is always legitimate in evaluating the performance of an estimator, and requires no signal strength or model uniqueness assumptions. The ℓ_2-recovery of $M(\hat{B} - B^*, \hat{C} - C^*)$ is fundamental, and such a bound, together with additional regularity assumptions, can easily be adapted to obtain estimation error bounds in different norms as well as selection consistency (Ye & Zhang, 2010; Lounici et al., 2011); see Theorem 10 in the Supplementary Material, for instance. Given a penalty function P or, equivalently, a robust loss ρ, we will study the performance of the set of global minimizers to show the ultimate power of the associated method; but our techniques of proof apply more generally (see, e.g., Theorem 7).

For any $C = (c_1, \ldots, c_n)^{\mathrm{T}}$, define

$$\mathcal{J}(C) = \{i : c_i \neq 0\}, \quad J(C) = |\mathcal{J}(C)| = \|C\|_{2,0}.$$

We let $r^* = r(B^*)$ denote the rank of the true coefficient matrix, and $J^* = J(C^*)$ the number of nonzero rows in C^*, i.e., the number of outliers. Let $q = r(X)$.

To handle problems in arbitrary dimensions, we construct some finite-sample oracle inequalities (Donoho & Johnstone, 1994). The first theorem considers a general penalty $P(C;\lambda) = \sum_{i=1}^{n} P(\|c_i\|_2;\lambda)$. Here we assume that $P(\cdot\,;\lambda)$ takes λ as the threshold parameter and satisfies

$$P(0;\lambda) = 0, \quad P(t;\lambda) \geqslant P_H(t;\lambda), \tag{14}$$

where $P_H(t;\lambda) = (-t^2/2 + \lambda|t|)1_{|t|<\lambda} + (\lambda^2/2)1_{|t|\geqslant\lambda}$. The latter inequality is natural in view of (5), because a shrinkage estimator with λ as the threshold is always bounded above by the hard-thresholding function $\Theta_H(\cdot\,,\lambda)$. From Theorem 2, (14) covers all ψ-functions bounded below by the skipped mean $\psi_H(s;\lambda) = s1_{|s|\leqslant\lambda}$ for any $s \geqslant 0$.

THEOREM 3. *Let $\lambda = A\sigma(m+\log n)^{1/2}$ with A a constant, and let $(\hat{B}, \hat{C})$ be a global minimizer of (9). Then, for any sufficiently large A, the following oracle inequality holds for any $(B, C) \in \mathbb{R}^{p\times m} \times \mathbb{R}^{n\times m}$ satisfying $r(B) \leqslant r$:*

$$E\{M(\hat{B} - B^*, \hat{C} - C^*)\} \lesssim M(B - B^*, C - C^*) + \sigma^2(q+m)r + P(C;\lambda) + \sigma^2, \tag{15}$$

where $\lesssim$ means that the inequality holds up to a multiplicative constant.

COROLLARY 1. *Under the same conditions as in Theorem 3, if $r \geqslant 1$ and P is a bounded nonconvex penalty satisfying $P(t;\lambda) \lesssim \lambda^2$ for any $t \in \mathbb{R}$, then*

$$\begin{aligned} &E\{M(\hat{B} - B^*, \hat{C} - C^*)\} \\ &\quad\lesssim \inf_{(B,C):\, r(B)\leqslant r} \big\{M(B - B^*, C - C^*) + \sigma^2(q+m)r + \sigma^2 J(C)m + \sigma^2 J(C)\log n\big\}. \end{aligned} \tag{16}$$

Remark 4. Both (15) and (16) involve a bias term $M(B - B^*, C - C^*)$. Upon setting $r = r^*$, $B = B^*$ and $C = C^*$ in, say, (16), we obtain a prediction error bound of the order

$$\sigma^2(q+m)r^* + \sigma^2 J^*(m + \log n). \tag{17}$$

On the other hand, the presence of the bias term ensures applicability of robust reduced-rank regression to weakly sparse C^*, and similarly r may also deviate from r^* to some extent, as a benefit derived from the bias-variance trade-off.

Remark 5. Our scheme of proof can also be used to show similar conclusions for the doubly penalized form (13) and the doubly constrained form (12), under the general assumption that the noise matrix has sub-Gaussian marginal tails. The following theorem states the result for (12), which is one of our favoured forms in practical data analysis.

THEOREM 4. *Let $(\hat{B}, \hat{C})$ be a solution to (12). With the convention $0\log 0 = 0$, we have*

$$\begin{aligned} &E\{M(\hat{B} - B^*, \hat{C} - C^*)\} \\ &\quad\lesssim \inf_{r(B)\leqslant r, J(C)\leqslant\varrho} M(B - B^*, C - C^*) + \sigma^2\{(q+m)r + \varrho m + \varrho\log(en/\varrho)\} + \sigma^2. \end{aligned}$$

Theorem 4 reveals some breakdown point information as a by-product. Specifically, fixing $\bar{Y} = XB$, we contaminate Y in the set $\mathcal{B}(\varrho) = \{Y \in \mathbb{R}^{n\times m} : Y = \bar{Y} + C + \mathcal{E},\ \|C\|_{2,0} \leqslant \varrho\}$, where $\mathrm{vec}(\mathcal{E})$ is sub-Gaussian and $\varrho \in \mathbb{N} \cup \{0\}$. Given any estimator $(\hat{B}, \hat{C})$ which implicitly depends on Y, we define its risk-based finite-sample breakdown point by $\epsilon^*(\hat{B}, \hat{C}) = (1/n) \times \min\{\varrho : \sup_{Y\in\mathcal{B}(\varrho)} E\{M(\hat{B}-B, \hat{C}-C)\} = +\infty\}$, where the randomness of the estimator is well accounted for by taking the expectation. Then, for the estimator defined by (12), it follows from Theorem 4 that $\epsilon^* \geqslant (\varrho + 1)/n$.

We emphasize that neither Theorem 3 nor Theorem 4 places any requirement on X, in contrast to Theorem 6 below.

Remark 6. The benefit of applying a redescending ψ is clearly shown by Theorem 3. As an example, for Huber's ψ, which corresponds to the popular convex ℓ_1 penalty due to Theorem 2, $P(C;\lambda)$ on the right-hand side of (15) is unbounded, while Hampel's three-part ψ gives a finite rate as seen in (16). Furthermore, we show that in a minimax sense, the error rate obtained in Corollary 1 is essentially optimal. Consider the signal class

$$\mathcal{S}(r, J) = \{(B^*, C^*) : r(B^*) \leqslant r,\ J(C^*) \leqslant J\} \quad (1 \leqslant r \leqslant q \wedge m,\ 1 \leqslant J \leqslant n/2).$$

Let $\ell(\cdot)$ be a nondecreasing loss function with $\ell(0) = 0, \ell \not\equiv 0$.

THEOREM 5. *Let $Y = XB^* + C^* + \mathcal{E}$ where $\mathcal{E}$ has independently and identically distributed $N(0, \sigma^2)$ entries. Assume that $n \geqslant 2$, $1 \leqslant J \leqslant n/2$, $1 \leqslant r \leqslant q \wedge m$, $r(q + m - r) \geqslant 8$, and $\sigma_{\min}(X)/\sigma_{\max}(X)$ is a positive constant, where $\sigma_{\max}(X)$ and $\sigma_{\min}(X)$ denote the largest and smallest nonzero singular values of X, respectively. Then there exist positive constants $\tilde{c}$ and c, depending on $\ell(\cdot)$ only, such that*

$$\inf_{(\hat{B},\hat{C})} \sup_{(B^*,C^*)\in\mathcal{S}(r,J)} E\big(\ell\big[M(\hat{B} - B^*, \hat{C} - C^*)/\{\tilde{c}P_o(J, r)\}\big]\big) \geqslant c > 0,$$

where $(\hat{B}, \hat{C})$ denotes any estimator of (B^*, C^*) and

$$P_o(J, r) = \sigma^2\{r(q + m) + Jm + J\log(en/J)\}.$$

We give some examples of ℓ to illustrate the conclusion. Using the indicator function $\ell(u) = 1_{u\geqslant 1}$, for any estimator $(\hat{B}, \hat{C})$, $M(\hat{B} - B^*, \hat{C} - C^*) \gtrsim \sigma^2\{r(q + m) + Jm + J\log(en/J)\}$ holds with positive probability. For $\ell(u) = u$, Theorem 5 shows that the risk $E\{M(\hat{B} - B^*, \hat{C} - C^*)\}$ is bounded from below by $P_o(J, r)$ up to some multiplicative constant. Therefore, (17) attains the minimax optimal rate up to a mild logarithmic factor, showing the advantage of utilizing redescending ψ-functions in robust low-rank estimation. The analysis is non-asymptotic and applies to any n, p and m.

Convex methods are however sometimes useful. In some less challenging problems, where some incoherence regularity condition is satisfied by the augmented design matrix, Huber's ψ can achieve the same low error rate. The result of the following theorem can be extended to any subadditive penalties with the associated ψ sandwiched between Huber's ψ and ψ_H.

THEOREM 6. *Let $(\hat{B}, \hat{C}) = \arg\min_{(B,C)} \|Y - XB - C\|_{\mathrm{F}}^2/2 + \lambda\|C\|_{2,1}$ subject to $r(B) \leqslant r$, with $\lambda = A\sigma(m + \log n)^{1/2}$ where A is a large enough constant. Then*

$$\begin{aligned} E\{M(\hat{B} - B^*, \hat{C} - C^*)\} \lesssim M(B - B^*, C - C^*) + \sigma^2 \\ + \sigma^2(q + m)r + \sigma^2K^2J(C)(m + \log n) \end{aligned} \tag{18}$$

for any (B, C) *with* $\mathrm{rank}(B) \leqslant r$, *if given* $\mathcal{J} = \mathcal{J}(C)$, X *satisfies* $(1+\vartheta)\|C'_{\mathcal{J}}\|_{2,1} \leqslant \|C'_{\mathcal{J}^c}\|_{2,1} + K|\mathcal{J}|^{1/2}\|(I-\mathcal{P}_r)C'\|_{\mathrm{F}}$ *for all* C' *and all* $\mathcal{P}_r$ *such that* $\mathcal{P}_r \subset \mathcal{P}_X$ *and* $r(\mathcal{P}_r) \leqslant 2r$, *where* $K \geqslant 0$ *and* ϑ *is a positive constant.*

Compared with (16), (18) has an additional factor of K on the right-hand side. Some numerical experiments on the magnitude of K, presented in the Supplementary Material, show that the error bound obtained in Theorem 6 is comparable to (16) in some settings. Also, under a different regularity condition, an estimation error bound on B^* can be obtained. See the Supplementary Material for more details.

Remark 7. The results obtained can be used to argue the necessity of robust estimation when outliers occur. Similar to Theorem 3, we can show that the ordinary reduced-rank regression, which sets $\hat{C} = 0$, satisfies

$$E\{M(\hat{B} - B^*, \hat{C} - C^*)\} \lesssim \inf_{r(B)\leqslant r} \|XB - (XB^* + C^*)\|_{\mathrm{F}}^2 + \sigma^2(q+m)r + \sigma^2. \tag{19}$$

Taking $r = r^*$, the error bound of the reduced-rank regression, evaluated at the optimal B satisfying $XB = XB^* + \mathcal{P}_{XB^*}C^*$ and $r(B) \leqslant r$, is of order

$$\sigma^2(q+m)r^* + \|(I - \mathcal{P}_{XB^*})C^*\|_{\mathrm{F}}^2. \tag{20}$$

Because XB^* has low rank, $I - \mathcal{P}_{XB^*}$ is not null in general. Notable outliers that can affect the projection subspace in performing rank reduction tend to occur in the orthogonal complement of the range of XB^*, and so (20) can be arbitrarily large, which echoes the deterministic breakdown-point conclusion in Theorem 1.

To control the size of the bias term, a better way is to use a larger rank value in the presence of outliers. Concretely, setting $B = B^* + (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}C^*$ in (19) yields

$$\sigma^2 J^* q + \sigma^2 J^* m + \sigma^2(q+m)r^* + \|(I - \mathcal{P}_X)C^*\|_{\mathrm{F}}^2, \tag{21}$$

where we have used $r(B) \leqslant r^* + J^*$. When $p > n$ we have $\mathcal{P}_X = I$, and so (21) offers an improvement over (20) by giving a finite error rate of $\sigma^2 J^* q + \sigma^2 J^* m + \sigma^2(q+m)r^*$. But our robust reduced-rank regression guarantees a consistently lower rate at $\sigma^2 J^* \log n + \sigma^2 J^* m + \sigma^2(q+m)r^*$, since $\sigma^2 J^* q \gg \sigma^2 J^* \log n$. The performance gain can be dramatic in big-data applications, where the design matrix is huge and typically multiple outliers are bound to occur.

4. COMPUTATION AND TUNING

In this section we show that compared with the M-characterization in Theorem 2, the additive formulation (6) simplifies computation and parameter tuning. Let us consider a penalized form of the robust reduced-rank regression problem

$$\min_{B,C} F(B, C) = \frac{1}{2}\|Y - XB - C\|_{\mathrm{F}}^2 + \sum_{i=1}^{n} P(\|c_i\|_2; \lambda) \quad \text{subject to } r(B) \leqslant r. \tag{22}$$

The penalties of interest may be nonconvex in light of the theoretical results in § 3, as stringent incoherence assumptions associated with convex penalties can be much relaxed or even removed.

Assuming that P is constructed by (5), a simple procedure for solving (22) is as described in Algorithm 1, where the two matrices C and B are alternately updated with the other held fixed until convergence. Here the multivariate thresholding, $\vec{\Theta}$, is defined based on Θ; cf. Definitions 1 and 2.

Algorithm 1. A robust reduced-rank regression algorithm.

Input X, Y, $C^{(0)}$, $B^{(0)}$, Θ, $t = 0$
Repeat
(a) $t \leftarrow t + 1$
(b) $C^{(t+1)} \leftarrow \vec{\Theta}(Y - XB^{(t)}; \lambda)$
(c) $B^{(t+1)} \leftarrow \mathcal{R}(X, Y - C^{(t+1)}, r)$, as defined in (3)
Until convergence

Step (b) performs simple multivariate thresholding operations and Step (c) performs reduced-rank regression on the adjusted response matrix $Y - C^{(t+1)}$. We need not explicitly compute B to update C in the iterative process. In fact, we need only $XB^{(t)}$, which depends on X through $\mathcal{P}_X$, or I when $p \gg n$. The eigenvalue decomposition called in (3) has low computational complexity because the rank values of practical interest are often small. Algorithm 1 is simple to implement and cost-effective. For example, even for $p = 1200$ and $n = m = 100$, it takes only about 40 seconds to compute a whole solution path for a two-dimensional grid of 100 λ values and 10 rank values.

THEOREM 7. *Let Θ be an arbitrary thresholding rule, and let F be as defined in (22), where P is associated with Θ through (5). Then, given any $\lambda \geqslant 0$ and $r \geqslant 0$, the proposed algorithm has the property that $F(B^{(t)}, C^{(t)}) \geqslant F(B^{(t+1)}, C^{(t+1)})$ for all t, and so $F(B^{(t)}, C^{(t)})$ converges as $t \to \infty$. Furthermore, under the assumptions that $\vec{\Theta}(\cdot\,; \lambda)$ is continuous in the closure of $\{Y - XB^{(t)}\}$ and $\{B^{(t)}\}$ is uniformly bounded, any accumulation point of $(B^{(t)}, C^{(t)})$ is a coordinatewise minimum point, and is a stationary point when $q(\cdot\,; \lambda) \equiv 0$; hence $F(B^{(t)}, C^{(t)})$ converges monotonically to $F(B^*, C^*)$ for some coordinatewise minimum point (B^*, C^*).*

The algorithm can be slightly modified to deal with (8), (12), and (13). For example, we can replace $\vec{\Theta}$ by Θ, applied componentwise, to handle elementwise outliers. The ℓ_0-penalized form with $P(C; \lambda) = (\lambda^2/2)\|C\|_{2,0}$, as well as the constrained form (12), will be used in data analysis and simulation. In implementation, they correspond to applying hard-thresholding and quantile-thresholding operators (She et al., 2013).

In common with most high-breakdown algorithms in robust statistics, we recommend using the multi-sampling iterative strategy (Rousseeuw & van Driessen, 1999). In many practical applications, however, we have found that the initial values can be chosen rather freely. Indeed, Theorem 8 shows that if the problem is regular, our algorithm guarantees low statistical error even without the multi-start strategy.

In the following theorem, given Θ, define $\mathcal{L}_\Theta = 1 - \operatorname{ess\,inf}\{\mathrm{d}\Theta^{-1}(u; \lambda)/\mathrm{d}u : u \geqslant 0\}$, where ess inf is the essential infimum. By definition, $\mathcal{L}_\Theta \leqslant 1$. We use $P_{2,\Theta}(C; \lambda)$ as shorthand for $\sum_{i=1}^n P_\Theta(\|c_i\|_2; \lambda)$ and set $r = (1 + \alpha)r^*$ with $\alpha \geqslant 0$ and $r^* \geqslant 1$.

THEOREM 8. *Let $(\hat{B}, \hat{C})$ be any solution satisfying $\hat{B} = \mathcal{R}(X, Y - \hat{C}, r)$ and $\hat{C} = \vec{\Theta}(Y - X\hat{B}; \lambda)$ with $\hat{B}$ of rank r and $\vec{\Theta}$ continuous at $Y - X\hat{B}$. Let Θ be associated with a bounded nonconvex penalty as described in Corollary 1 and let $\lambda = A\sigma(m + \log n)^{1/2}$ with A being a large enough*

constant. Assume that $(1+\alpha)^{-1/2}\|XB - XB^*\|_{\mathrm{F}}^2 + \mathcal{L}_\Theta\|C - C^*\|_{\mathrm{F}}^2 + \vartheta P_{2,H}(C - C^*;\lambda) \leqslant (2-\delta)M(B-B^*, C-C^*) + 2P_{2,\Theta}(C;\lambda) + \zeta P_{2,0}(C^*;\lambda)$ *holds for all* (B, C) *satisfying* $r(B) \leqslant r$, *where* $\zeta \geqslant 0$, $\delta > 0$ *and* $\vartheta > 0$ *are constants. Then* $E\{M(\hat{B} - B^*, \hat{C} - C^*)\} \lesssim \sigma^2(1+\alpha)(q+m)r^* + \sigma^2 J^* m + \sigma^2 J^* \log n$.

To choose an optimal rank for B and an optimal row support for C jointly, crossvalidation would seem to be an option. However, it lacks theoretical support in the robust low-rank setting, and for large-scale problems crossvalidation can be quite expensive. Motivated by Theorem 5, we propose the predictive information criterion

$$\log\|Y - XB - C\|_{\mathrm{F}}^2 + \frac{1}{mn}\big[A_1\{Jm + (m+q-r)r\} + A_2 J\log(en/J)\big], \tag{23}$$

where $\|Y - XB - C\|_{\mathrm{F}}^2$ is the residual sum of squared errors, $r = r(B)$, $J = \|C\|_{2,0}$, and e denotes the Euler constant. The term $Jm + (m+q-r)r$ counts the degrees of freedom of the obtained model, and $J\log(en/J)$ characterizes the risk inflation. The benefits of the criterion include that no noise scale parameter needs to be estimated, and minimizing (23) achieves the minimax optimal error rate when the true model is parsimonious, as shown below.

THEOREM 9. *Let* $P(B, C) = Jm + (m+q-r)r + J\log(en/J)$, *where* $r = r(B)$ *and* $J = \|C\|_{2,0}$. *Suppose that the true model is parsimonious in the sense that* $P(B^*, C^*) < mn/A_0$ *for some constant* $A_0 > 0$. *Let* $\delta(B, C) = AP(B, C)/(mn)$ *where* A *is a positive constant satisfying* $A < A_0$, *and so* $\delta(B^*, C^*) < 1$. *Then for sufficiently large values of* A_0 *and* A, *any* $(\hat{B}, \hat{C})$ *that minimizes* $\log\|Y - XB - C\|_{\mathrm{F}}^2 + \delta(B, C)$ *subject to* $\delta(B, C) < 1$ *satisfies* $M(\hat{B} - B^*, \hat{C} - C^*) \lesssim \sigma^2\{J^* m + (m+q-r^*)r^* + J^*\log(en/J^*)\}$ *with probability at least* $1 - c_1' n^{-c_1} - c_2' \exp(-c_2 mn)$ *for some constants* $c_1, c_1', c_2, c_2' > 0$.

Based on computer experiments, we set $A_1 = 7$ and $A_2 = 2$.

5. ARABIDOPSIS THALIANA DATA

We performed extensive simulation studies to compare our method with some classical robust multivariate regression approaches and several reduced-rank methods (Reinsel & Velu, 1998; Tatsuoka & Tyler, 2000; Aelst & Willems, 2005; Roelant et al., 2009; Bunea et al., 2011; Mukherjee & Zhu, 2011) in both low and high dimensions. The results are reported in the Supplementary Material; our robust reduced-rank regression shows performance comparable or superior to the other methods in terms of both prediction and outlier detection.

Isoprenoids are diverse and abundant compounds in plants, where they serve many important biochemical functions and play roles in respiration, photosynthesis and the regulation of growth and development. To examine the regulatory control mechanisms in gene networks for isoprenoids in *Arabidopsis thaliana*, a genetic association study was conducted, with $n = 118$ GeneChip microarray experiments performed to monitor gene expression levels under various experimental conditions (Wille et al., 2004). It was experimentally verified that strong connections exist between some downstream pathways and two isoprenoid biosynthesis pathways. We therefore considered a multivariate regression set-up, with the expression levels of $p = 39$ genes from the two isoprenoid biosynthesis pathways serving as predictors, and the expression levels of $m = 62$ genes from four downstream pathways, namely plastoquinone, carotenoid, phytosterol and chlorophyll, serving as the responses.

Because of the small sample size relative to the number of unknowns, we applied robust reduced-rank regression with the predictive information criterion for parameter tuning. The final

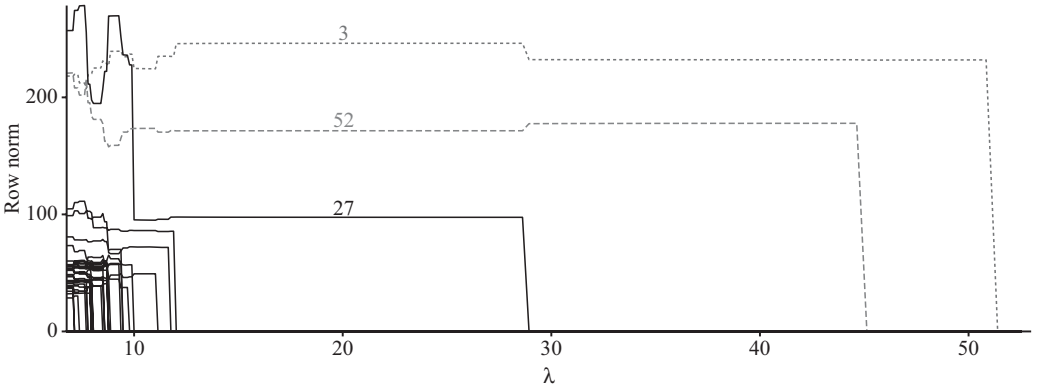

Fig. 1. *Arabidopsis thaliana* data: outlier detection paths obtained by the robust reduced-rank regression. Sample 3 and sample 52 are captured as outliers, whose paths are shown as a dotted line and a dashed line, respectively. The path plot also suggests sample 27 as a potential outlier.

model is of rank five, which means that the effective number of unknowns is reduced by about 80% compared with the least-squares model. Interestingly, our method also identified two outliers, samples 3 and 52. Figure 1 shows the detection paths by plotting the ℓ_2-norm of each row in the C estimates for a sequence of λ values. The two unusual samples are distinctive. The outlyingness could be caused by different experimental conditions. In particular, sample 3 was the only sample with *Arabidopsis* tissue culture in a baseline experiment. The two outliers have a surprisingly large impact on both coefficient estimation and model prediction. This can be seen from $\|\hat{B} - \tilde{B}\|_{\mathrm{F}}/\|\tilde{B}\|_{\mathrm{F}} \approx 50\%$ and $\|X\hat{B} - X\tilde{B}\|_{\mathrm{F}}/\|X\tilde{B}\|_{\mathrm{F}} \approx 26\%$, where $\hat{B}$ and $\tilde{B}$ denote, respectively, the robust reduced-rank regression and the plain reduced-rank regression estimates. In addition, Fig. 1 reveals that sample 27 could be a potential outlier meriting further investigation.

The low-rank model obtained reveals robust score variables, or factors, constructed from isoprenoid biosynthesis pathways, in response to the 62 genes in the four downstream pathways. Let $\tilde{X}$ denote the design matrix after removing the two detected outliers, and let $\hat{U}\hat{D}\hat{V}^{\mathrm{T}}$ be the singular value decomposition of $\tilde{X}\hat{B}$. Then $\hat{U}$ delivers five orthogonal factors, and $\hat{V}\hat{D}$ gives the associated factor coefficients. Figure 2 plots the coefficients of the first three leading factors for all 62 response variables. Given the sth factor ($s = 1, 2, 3$), the genes are grouped into the four pathways separated by vertical lines, and two horizontal lines are placed at heights $\pm\sigma_s^{\tilde{X}\hat{B}} m^{-1/2}$. Therefore, the genes located beyond those two horizontal lines have relatively large-magnitude coefficients on the corresponding factor.

We also tested the significance of the factors in response to each of the 62 genes; see Table 1. Plastoquinone was excluded since it has only two genes and its behaviour couples with that of carotenoid most of the time. Even with the familywise error rate controlled at 0·01, the factors obtained are predictive overall according to the significance percentages, although they play very different roles in different pathways. In fact, according to Fig. 2 and Table 1, the genes that are correlated with the first factor are mainly from carotenoid and chlorophyll, and almost all the coefficients there are negative. It seems that the first factor interprets some joint characteristics of carotenoid and chlorophyll; the second factor differentiates phytosterol genes from carotenoid genes; and the third factor appears to contribute mainly to the phytosterol pathway. Therefore, by projecting the data onto a proper low-dimensional subspace in a supervised and robust manner, distinct behaviours of the downstream pathways and their potential subgroup structures can be revealed. Further biological insights could be gained by closely examining the experimental and background conditions.

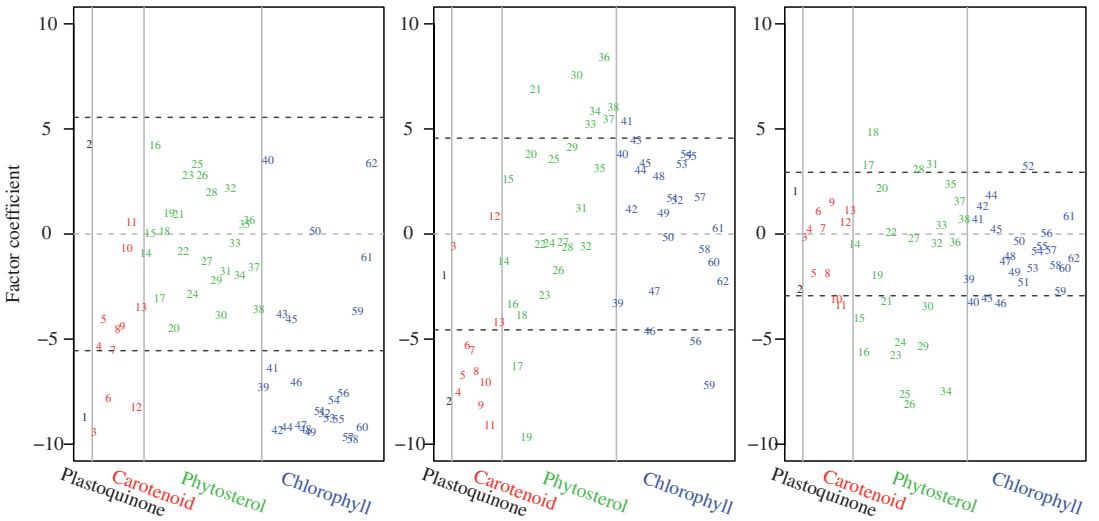

Fig. 2. *Arabidopsis thaliana* data: factor coefficients of the 62 response genes from plastoquinone, carotenoid, phytosterol, and chlorophyll pathways. From left to right the panels correspond to the top three factors estimated by the robust reduced-rank regression. For the sth factor ($s = 1, 2, 3$), two horizontal lines are plotted at heights $\pm\sigma_s^{\tilde{Y}\tilde{B}} m^{-1/2}$, and three vertical lines separate the genes into four different pathways.

Table 1. *Arabidopsis thaliana data: percentage of genes on each response pathway that show significance of a given factor, with the familywise error rate controlled at level* 0·01

Pathway	Number of genes	Factor 1	Factor 2	Factor 3
Carotenoid	11	55%	73%	9%
Phytosterol	25	20%	48%	32%
Chlorophyl	24	75%	21%	0%

Acknowledgement

The authors thank the editor, associate editor and referees for suggestions that have significantly improved the paper. The first author thanks Art Owen for his valuable comments on an earlier version of the manuscript, and Elvezio Ronchetti for inspiration. This work was partially supported by the U.S. National Science Foundation, National Institutes of Health and Department of Defense.

Supplementary material

Supplementary material available at *Biometrika* online includes all technical details of the proofs, a financial example and the simulation studies.

References

Aelst, S. V. & Willems, G. (2005). Multivariate regression S-estimators for robust estimation and inference. *Statist. Sinica* **15**, 981–1001.

Agarwal, A., Negahban, S. & Wainwright, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.* **40**, 1171–97.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22**, 327–51.

Bunea, F., She, Y. & Wegkamp, M. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39**, 1282–309.
Candès, E. J., Li, X., Ma, Y. & Wright, J. (2011). Robust principal component analysis? *J. Assoc. Comp. Mach.* **58**, 1–37.
Chen, K., Dong, H. & Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika* **100**, 901–20.
Donoho, D. & Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, P. J. Bickel, K. Doksum & J. L. Hodges, eds., Wadsworth Statistics/Probability Series. Belmont, California: Wadsworth, pp. 157–84.
Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.
Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.
Foygel, R., Horrell, M. & Lafferty, M. D. J. (2012). Nonparametric reduced rank regression. *Adv. Neural Info. Proces. Syst.* **25**, 1637–45.
Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (2005). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.
Huber, P. (1981). *Robust Statistics*. New York: Wiley.
Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Mult. Anal.* **5**, 248–64.
Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. New York: Springer.
Koltchinskii, V., Lounici, K. & Tsybakov, A. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Ann. Statist.* **39**, 2302–29.
Lounici, K., Pontil, M., van de Geer, S. & Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39**, 2164–204.
Mukherjee, A. & Zhu, J. (2011). Reduced rank ridge regression and its kernel extensions. *Statist. Anal. Data Mining* **4**, 612–22.
Reinsel, G. C. & Velu, P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. New York: Springer.
Roelant, E., Aelst, S. V. & Croux, C. (2009). Multivariate generalized S-estimators. *J. Mult. Anal.* **100**, 876–87.
Rohde, A. & Tsybakov, A. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39**, 887–930.
Rousseeuw, P. J. & van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–23.
She, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Comp. Statist. Data Anal.* **56**, 2976–90.
She, Y. (2013). Reduced rank vector generalized linear models for feature extraction. *Statist. Interface* **6**, 197–209.
She, Y. & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *J. Am. Statist. Assoc.* **106**, 626–39.
She, Y., Wang, J., Li, H. & Wu, D. (2013). Group iterative spectrum thresholding for super-resolution sparse spectral selection. *IEEE Trans. Sig. Proces.* **61**, 6371–86.
Tatsuoka, K. & Tyler, D. (2000). The uniqueness of S-functionals and M-functionals under nonelliptical distributions. *Ann. Statist.* **28**, 1219–43.
Vounou, M., Nichols, T. E. & Montana, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage* **53**, 1147–59.
Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L. et al. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.* **5**, R92.
Wright, J., Ganesh, A., Rao, S., Peng, Y. & 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*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams & A. Culotta, eds. Neural Information Processing Systems (NIPS), pp. 2080–8.
Ye, F. & Zhang, C.-H. (2010). Rate minimaxity of the lasso and Dantzig selector for the l_q loss in l_r balls. *J. Mach. Learn. Res.* **11**, 3519–40.
Yee, T. & Hastie, T. J. (2003). Reduced rank vector generalized linear models. *Statist. Mod.* **3**, 367–78.
Yuan, M., Ekici, A., Lu, Z. & Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Statist. Soc.* B **69**, 329–46.
Zhang, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.
Zhang, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11**, 1081–107.
Zhou, Z., Li, X., Wright, J., Candès, E. & Ma, Y. (2010). Stable principal component pursuit. In *Proc. 2010 IEEE Int. Symp. Info. Theory*. Institute of Electrical and Electronics Engineers, pp. 1518–22.

[*Received on* 13 *September* 2015. *Editorial decision on* 9 *April* 2017]