Journal of Multivariate Analysis 140 (2015) 377–394

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On estimation in the reduced-rank regression with a large number of responses and predictors

Vladislav Kargin *

Cambridge University, Department of Pure Mathematics and Mathematical Statistics, Wilberforce Rd, Cambridge CB3 0WB, United Kingdom

ARTICLE INFO

Article history:
Received 25 September 2014
Available online 12 June 2015

AMS subject classifications:
primary 62H20
secondary 62H15
60B20
15B52

Keywords:
Random matrices
High-dimensional data
Reduced-rank regression
Rank selection
Rank estimation
Tracy–Widom distribution
Factor model

ABSTRACT

We consider a multivariate linear response regression in which the number of responses and predictors is large and comparable with the number of observations, and the rank of the matrix of regression coefficients is assumed to be small. We study the distribution of singular values for the matrix of regression coefficients and for the matrix of predicted responses. For both matrices, it is found that the limit distribution of the largest singular value is a rescaling of the Tracy–Widom distribution. Based on this result, we suggest algorithms for the model rank selection and compare them with the algorithm suggested by Bunea, She and Wegkamp. Next, we design two consistent estimators for the singular values of the coefficient matrix, compare them, and derive the asymptotic distribution for one of these estimators.

1. Introduction

In this paper we are concerned with the reduced rank variant of the multivariate response regression model. We are given N observations of the predictors $X_i \in \mathbb{R}^p$ and responses $Y_i \in \mathbb{R}^r$, which are assumed to be related by the linear regression model:

$$Y = XA + U, \tag{1}$$

where A is an unknown p-by-r matrix and U is a noise matrix. This model is ubiquitous in statistics, signal processing, and numerical analysis.

On methodological grounds one often postulates that the responses depend only on a small number of factors which are linear combinations of the predictors. This postulate leads to a model, in which A is assumed to be a low-rank matrix:

$$A = \sum_{j=1}^{s} \theta_j u_j v_j^*, \tag{2}$$

where $\{u_j \in \mathbb{R}^p\}$ and $\{v_i \in \mathbb{R}^r\}$ are two fixed orthonormal vector systems. This model appeared already in Anderson [1], and it was named *reduced-rank regression* in Izenman [17]. In some contexts, this model is also known under the names

* Correspondence to: 282 Mosher Way, Palo Alto, CA 94304, USA.
E-mail address: vladislav.kargin@gmail.com.

http://dx.doi.org/10.1016/j.jmva.2015.06.004

simultaneous linear prediction (Fortier [13]) and *redundancy analysis* (van den Wollenberg [34]), both of which assume that U has the covariance matrix equal to $\sigma^2 I$. The reduced-rank model has been intensively studied, and many results are collected in the monograph by Reinsel and Velu [30].

In this paper, we assume that U has the covariance matrix equal to $\sigma^2 I$, and we are interested in the situation in which all three variables, p, r, and N, grow at the same rate.

Assumption A1. It is assumed that as $N \to \infty$, $\frac{N}{p} \to 1+\lambda \geq 1$ and $\frac{N}{r} \to \mu > 0$.

It is also useful to define $\beta \overset{\text{def}}{=} \lim_{N\to\infty} \frac{p}{r} = \mu/(1+\lambda)$.

The studies devoted to the reduced-rank regression in this setup are relatively recent and include Bunea, She, and Wegkamp [9] and Giraud [14].

We address the following questions. First, is it possible to detect that the true matrix A is not zero? If yes, then how do we estimate the rank and singular values of A?

Our approach to these questions is based on the study of the statistical properties of the standard least squares estimator

$$\widehat{A} \overset{\text{def}}{=} X \setminus Y \equiv \left(X^*X\right)^{-1} X^*Y$$

and the matrix of fitted responses:

$$\widehat{Y} \overset{\text{def}}{=} X\widehat{A}.$$

By using this approach, we will develop a rank-selection algorithm which performs better than the algorithm from [9] in a certain range of parameters and is simpler than the algorithm in [38]. In addition, we will develop tools for consistent estimation of singular values θ_i. The paper [9] does not address this issue, since its focus is on minimizing the prediction error, in particular on bounds for $\mathbb{E}\left\|XA - X\widetilde{A}\right\|$, where $\widetilde{A}$ is an estimator of A and the expectation is over randomness in U.

The rest of the paper is organized as follows. Section 2 describes the major results. Section 3 provides the details of the proofs. Section 4 recapitulates the results. Appendix provides a proof for the theorem about the limiting distribution of singular values of $\widehat{A}$.

2. Major results

2.1. Tests of the null hypothesis

Let X be a p-by-r real Gaussian matrix: each row is an independent observation from $\mathcal{N}(0, \Sigma)$. Then, an r-by-r matrix X^*X is said to be a *Wishart matrix* with distribution $W_r(\Sigma, p)$.

A random m-by-m matrix X is said to belong to the *(real) Jacobi ensemble* with parameters α_1 and α_2, if its distribution is invariant with respect to orthogonal transformations and the distribution of its eigenvalues is given by

$$f^{(\alpha_1,\alpha_2)}(\lambda_1, \ldots, \lambda_m) = \frac{1}{c}\prod_{j=1}^{m} \lambda_j^{\alpha_1}(1-\lambda_j)^{\alpha_2} \prod_{1\leq j<k\leq m} \left|\lambda_j - \lambda_k\right|. \tag{3}$$

The following result is fundamental for the analysis of matrices $\widehat{A}$ and $\widehat{Y}$.

Theorem 2.1. (i) *Suppose that U is an N-by-r matrix with i.i.d standard real Gaussian entries, and X is an N-by-p full-rank matrix ($N \geq p$) independent of U. Then the squared singular values of $\widehat{Y} \overset{\text{def}}{=} X(X \setminus U)$ are distributed as the eigenvalues of the Wishart matrix with distribution $W_r(I, p)$.*

(ii) *In addition, suppose that X has i.i.d standard real Gaussian entries. Let s_i^2 be the squared singular values of $\widehat{A} \overset{\text{def}}{=} X \setminus U$ and $f_i = s_i^2/(1+s_i^2)$. Then, the positive f_i are distributed as eigenvalues of the Jacobi ensemble with parameters $m = \min\{p, r\}$, $\alpha_1 = (|r-p|-1)/2$ and $\alpha_2 = (N-p-1)/2$.*

Proof. The matrix $\widehat{Y} = X(X \setminus U)$ is the orthogonal projection of r column vectors of U on the p-dimensional column span of X. Hence, in an appropriate basis, $\widehat{Y}$ is a block matrix with one block given by a p-by-r matrix with i.i.d. standard Gaussian entries and another block of $(N-p)$-by-r matrix of zeros. This proves the first part of the theorem. For the second part, note that positive eigenvalues of $\widehat{A}^*\widehat{A} = U^*X\,(X^*X)^{-2}\,X^*U$ have the same distribution as positive eigenvalues of $B^{-1}C$, where B and C are independent Wishart matrices.

Indeed, the rank of matrices $U^*X(X^*X)^{-2}X^*U$ and $X(X^*X)^{-2}X^*UU^*$ is $\min\{p, r\}$, and their positive eigenvalues are the same. Let W be an orthogonal N-by-p matrix formed by the eigenvectors of $X(X^*X)^{-2}X^*$ and such that the matrix $W^*X(X^*X)^{-2}X^*W$ is diagonal with positive eigenvalues on the diagonal. These eigenvalues coincide with positive eigenvalues of the inverse of a Wishart matrix, $(X^*X)^{-1}$, where the Wishart matrix has the distribution $W_p(I, N)$. The matrix W^*UU^*W is Wishart with distribution $W_p(I, r)$.

In addition, matrices $W^*X(X^*X)^{-2}X^*W$ and W^*UU^*W are independent because the eigenvalues and eigenvectors of $X(X^*X)^{-2}X^*$ are independent. Finally, since similarity transformations do not change eigenvalues, the distribution of