Available online at www.sciencedirect.com

Journal of Multivariate Analysis 97 (2006) 925–945

www.elsevier.com/locate/jmva

Who’s afraid of reduced-rank parameterizations of multivariate models? Theory and example

Scott Gilbert[a,*], Petr Zemčík[b]

[a] *Economics Department, Southern Illinois University, Carbondale, IL 62901-4515, USA*
[b] *CERGE-EI,[1] P.O. Box 882, Politických vězňu 7, Prague, Czech Republic*

Received 5 March 2004
Available online 1 December 2005

Abstract

Reduced-rank restrictions can add useful parsimony to coefficient matrices of multivariate models, but their use is limited by the daunting complexity of the methods and their theory. The present work takes the easy road, focusing on unifying themes and simplified methods. For Gaussian and non-Gaussian (GLM, GAM, mixed normal, etc.) multivariate models, the present work gives a unified, explicit theory for the general asymptotic (normal) distribution of maximum likelihood estimators (MLE). MLE can be complex and computationally hard, but we show a strong asymptotic equivalence between MLE and a relatively simple minimum (Mahalanobis) distance estimator. The latter method yields particularly simple tests of rank, and we describe its asymptotic behavior in detail. We also examine the method’s performance in simulation and via analytical and empirical examples.

AMS 1991 subject classification: 62H10; 62H12; 62H15; 62F05; 62F12

Keywords: Multivariate model; Regression; Coefficient matrix; Reduced-rank; Estimation; Test; Asymptotic theory

1. Introduction

Reduced-rank restrictions can add useful parsimony to coefficient matrices of multivariate models, but their use is limited by the daunting complexity of the methods and their theory. In particular, reduced rank regression [4,29], which has been extensively researched, is not yet included in most statistics textbooks, even at the graduate level, nor in most statistical software

* Corresponding author. Fax: +1 618 453 27 17.
E-mail address: gilberts@siu.edu (S. Gilbert).

[1] CERGE-EI is a joint workplace of the Center for Economic Research and Graduate Education, Charles University, and the Economics Institute of the Academy of Sciences of the Czech Republic.

0047-259X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2005.10.002

packages. This dearth of technical training and support exists in vicious cycle with the limited number of applications attempted so far.

In an attempt to make reduced-rank methods more accessible to the average multivariate modeller, the present work takes the easy road, focusing on unifying themes and simplified methods. For Gaussian and non-Gaussian (generalized linear models—GLM, generalized additive models—GAM, etc.) types of multivariate models, the present work gives a unified, explicit theory for the general asymptotic (normal) distribution of maximum likelihood estimators, and also studies some simpler methods. To set the context of this theory, for a random variable y and a k-vector x let $F(y|x)$ be the conditional (cumulative) distribution function of y given x. Let $\phi(x) = \Phi(F(\cdot|x))$ describe some feature of the conditional distribution of y, via a function(al) Φ that maps conditional distribution functions to functions of x (alone). For each of g groups $i = 1, 2, \dots, g$, with $g \leqslant k$, let $F_i(y|x)$ be the conditional distribution of y in that group, and let $\phi_i(x) = \Phi(F_i(\cdot|x))$.

Let the general feature ϕ be linear in parameters θ:

$$\phi_i(x) = \theta_i' x \tag{1}$$

for $i = 1, 2, \dots, g$, with coefficient k-vectors θ_i subject to reduced rank, meaning that the $g \times k$ coefficient matrix $\Theta = (\theta_1, \dots, \theta_g)'$ has rank $r < g$. For simplicity we suppose further that the user has arranged the data so that the first r rows of Θ form a basis for all rows. The model then has three important ingredients:

(i) a dependent variable for each of two or more groups,
(ii) linear linkage between dependent variable and independent variables,
(iii) limitations on links' degrees of freedom, due to a rank condition.

In the Gaussian multivariate linear model, the feature $\phi(x)$ is the conditional mean $\mu(x) = \int y \, dF(y|x)$. Here reduced-rank (iii) can be applied ad hoc, as an interesting model simplification, or can be motivated by some scientific theory. For example, a literature in financial economics (see [35, Chapter 8], for an excellent summary) takes the latter approach when modelling asset returns, as in [23]. Related to reduced-rank regression models are factor analysis, growth curve models, MIMIC (Multiple Indicator Multiple Cause) models, error-in-variables models, latent variables models, index models, common trends, error correction models and co-integration models, and for relevant discussion and applications we refer the reader to [4–6,8–11,47,32,25,43,21,22,39,2,3,41, 12,38,1,35, 48].

Reduced-rank parameterization has also been developed for some non-Gaussian multivariate models. These include the multinomial logit model [6,7], the vector generalized linear model (GLM) and vector generalized additive model (GAM), see [46] for recent discussion. Typically, in these models the matrix Θ parameterizes a feature ϕ which is not itself a (conditional) mean, but is related to mean of some (transformed) variable.

For many non-Gaussian multivariate models, reduced-rank methods are rarely (if ever) attempted. For example, as a measure of the center or location of a continuous distribution, an alternative to the conditional mean is the conditional median $m(x) = F^{-1}\left(\frac{1}{2}|x\right)$, this being the median of y conditional on x, for which $P(y \leqslant m(x)|x) = \frac{1}{2}$. When data have an asymmetric (hence non-Gaussian) distribution, the median typically differs from the mean. Linear models of conditional median date back at least to [16], and [26] provide a review of theory and some applications of such models (see also [28] for linear models of other location measures, and [33] for models of conditional quantiles including the median). Any time that reduced-rank MANOVA or multivariate linear regression models are employed, one can imagine trying out also reduced-rank *median*-based models (without normality assumptions). However, we know of no such attempt, perhaps due to the task's perceived difficulty. As we show, there is a reasonably easy way to approach such problems.

As another example, consider multivariate models of variability or scale, via the conditional standard deviation:

$$\sigma(x) = \sqrt{\int (y - \mu(x))^2 \, dF(y|x)}.$$

A linear model of variability is then $\sigma_i(x) = \theta_i' x$, $i = 1, 2, \ldots, g$, in which case the coefficient vectors θ_i describe a conditional variability/heteroskedasticity feature, rather than a conditional location feature. We are not aware of linear models of conditional standard deviation in the literature, but the example in Section 2 derives such a model from a form of stochastic dominance. The linear model of $\sigma(x)$ has the ingredients (i)–(iii), with a linear form (ii) of conditional standard deviation, and reduced-rank (iii) applied to the matrix Θ of conditional variability coefficients. We can similarly apply reduced-rank structure to linear models of conditional variance $\sigma^2(x)$ (these being common in economics/econometrics) and other features of the conditional distribution.

Maximum likelihood is the usual method for multivariate analysis, and we provide a unified theory for the general asymptotic (normal) distribution of maximum likelihood estimators (MLE) of reduced-rank multivariate models. However, maximum likelihood is often not the simplest method, and it may be computationally burdensome. By comparison, a relatively simple "minimum (Mahalanobis) distance" estimator, which we interpret as a "maximum approximate density" (MAD) estimator, is typically available, as in [23]. This sort of estimator has, under standard conditions, an asymptotic normal distribution which is fairly easy to establish (via the Delta Method) in broad form. We go further, describing the MAD estimator's behavior in more detail.

We show a strong asymptotic equivalence between MAD and MLE estimators, these two being perfectly correlated as sample size approaches infinity. To further interpret the MAD estimator, we note that it maximizes a particular (asymptotically valid) density function associated with a plug-in unrestricted (full-rank) estimator $\hat{\Theta}$. The MAD approach is intuitive and quite general, and we describe further similarities between it and the maximum likelihood estimator.

For MAD estimation, we assume that the plug-in $\hat{\Theta}$ is asymptotically normal, and this covers many cases of interest but time series models with unit root dynamics, where $\hat{\Theta}$ can be asymptotically non-normal (see for example [30,31,3,35, Chapter 5]), this being the subject of the project's sequel (in progress). The proposed MAD estimator takes as input an available full-rank estimator and plug-in variance–covariance estimate, and is consistent with an asymptotically normal distribution that we describe in detail (via explicit formulas for the relevant variance/covariance matrix). The estimator does not require a fully specified probability model, yet mimics some special behavior of maximum likelihood estimators. Also, the proposed estimator is identical, asymptotically, to constrained MLE when $\hat{\Theta}$ is (unconstrained) MLE. An advantage of the proposed method is its general practicality, whereas constrained MLE (for reduced-rank multivariate conditional variability, etc.) may be hard to compute (when available) for non-Gaussian models. We illustrate this advantage in the case of a mixed normal probability model.

We also propose a rank test, based on the ratio of asymptotic densities (RAD) for constrained and unconstrained estimators. This testing principle is intuitive and general. Since we assume that the unconstrained estimator $\hat{\Theta}$ is asymptotically normal, we report here test theory for this case only. Our approach tests whether the first r rows of coefficient matrix Θ span the rest, and hence is consistent against two (overlapping) alternatives: (a) that Θ has rank $> r$, and (b) that the first r rows are not a basis of Θ. Hence, our test allows us to check for misspecification of the posited row basis. By comparison, other general rank tests (including [24,17,18,36]) are consistent against (a) but not (b), because they test for the existence of reduced-rank regardless of which rows form a basis. Further, we show that our test is equivalent, asymptotically, to a likelihood ratio test (which may be hard to compute) when the plug-in $\hat{\Theta}$ is (unconstrained) MLE.

The remainder of the paper is organized as follows. Section 2 gives an economic example, Section 3 defines the proposed estimator and test, and Section 4 provides asymptotic theory for the methods. Section 5 continues the economic example, Section 6 studies performance through an analytical example and simulation, Section 7 concludes, and an Appendix contains mathematical proofs.

2. Example

We give a simple example that illustrates reduced-rank multivariate linear modelling of both conditional location (via mean and median) and conditional variability. The model, which posits a form of stochastic dominance between groups, has aforementioned ingredients (i), (ii) and (iii), applied to conditional mean, median and standard deviation, respectively. Estimating these models, subject to reduced-rank on the matrix of interest, is sometimes difficult via traditional maximum likelihood methods, but can be more convenient using approximate density (MAD and RAD) methods.

Let there be $g = 2$ groups of workers, the first group male and the second female. For a random sample of workers, with n_1 males and n_2 females, let y_{ij} be the income of a worker in the ith gender group and jth education level, with $j = 1$ indicating at most a high school degree, and $j = 2$ indicating some college education.

We use data from the Integrated Public Use Micro-data Samples database (available at www.ipums.umn.edu, see [37] for description). This data is a random sample, from the year 1990, of US persons 16 years and older who earn a positive amount of income and have at most a bachelor's degree. The sample has features typically observed in income data (see [13,15,14]), including higher incomes for the more educated workers, and higher incomes for men. From Table 1, both income and log-income show high kurtosis (fat tails), and there is positive skew for income and negative skew for log-income, in each gender × education pairing.

Table 1
Income sample statistics, by sex and education

		Male		Female	
		Low ed	High ed	Low ed	High ed
Income	n	1556	403	1632	306
	Mean	20,871.82	47,767.38	11,570.22	24,185.74
	Median	17,000.00	36,000.00	8,344.00	20,057.50
	SD	18,711.83	43,395.45	10,729.81	19,994.79
	Skewness	3.28	2.49	2.35	2.93
	Kurtosis	25.32	10.39	14.44	21.70
Log-income	Skewness	−1.63	−0.85	−1.38	−1.44
	Kurtosis	8.81	5.99	7.03	5.60

Table 2
Income, mixture model

	Male		Female	
	Low ed	High ed	Low ed	High ed
Parameters				
π	0.88	0.84	0.62	0.96
μ_1	17,056.30	33,999.09	6,159.57	22,058.32
μ_2	50,024.76	119,998.56	20,453.47	77,368.69
σ_1	11,452.10	17,944.54	3,923.93	14,520.58
σ_2	32,560.96	61,905.36	12,316.92	46,342.79
Descriptives				
Mean	20,871.82	47,767.38	11,570.22	24,185.74
Median	18,380.37	37,685.91	8,250.88	22,617.45
SD	18,705.82	43,341.57	10,726.53	19,962.09
Skewness	1.99	2.32	1.275	2.46
Kurtosis	10.24	9.38	4.432	16.79

To model the skewed and fat-tailed income distribution for women and men at different education levels, we apply a Mixed Normal (abbreviated MN) probability model, separately to each of the four groups: male—low ed, male—high ed, female—low ed, female—high ed. The Mixed Normal, with suitably many mixture components, provides a flexible generalization of the normal (Gaussian) distribution. For each group, we model income as the probability mixture of two normal distributions. With gender labels $i = 1, 2$ for categories (male, female) and education labels $j = 1, 2$ for categories (low, high), we specify the probability density for gender i and education level j as

$$f_{ij}(y) = \pi f_{\mathcal{N}}(y; \mu_{ij1}, \sigma^2_{ij1}) + (1 - \pi) f_{\mathcal{N}}(y; \mu_{ij2}, \sigma^2_{ij2}),$$

where $f_{\mathcal{N}}(y; \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2. For our random sample, income y is assumed independent across genders i and education levels j.

For each group we estimate the Mixed Normal model via maximum likelihood, using the EM algorithm of [19]. We report parameter estimates in the top part of Table 2, and from the fitted density functions $\hat{f}_{ij}(y)$ we compute the fitted group-wise mean, median, and standard deviation:

$$\hat{\mu}_{ij} = \int y \hat{f}_{ij}(y)\, dy, \quad \hat{m}_{ij} = \hat{F}_{ij}^{-1}\left(\frac{1}{2}\right), \quad \hat{\sigma}_{ij} = \sqrt{\int (y - \hat{\mu})^2 \hat{f}_{ij}(y)\, dy},$$

where F is the cumulative distribution function. The mean and standard deviation have known formulas, based on the MN model (see for example [27,20]), and for the median we use simulation of MN with a pseudo-sample of size 100,000. We report these results in the bottom part of Table 2, as well as the skewness and kurtosis for each MN model, obtained via simulation.

The income descriptives in Tables 1 and 2 are consistent with the idea that women in 1990 tended to earn about half of what men did, in each education category. Formally,

$$y_{2j} \overset{d}{=} c y_{1j}, \quad j = 1, 2, \tag{2}$$

where $\overset{d}{=}$ means equality in distribution, and c a constant close to 1/2. This characterization, which is a form of (first-order) *stochastic dominance*, allows a general income distribution for men at each education level, and restricts only the relative performance of women versus men.

To put this form of stochastic dominance in the context of the feature model (1), define 2×1 vectors $x_i = (x_{i1}, x_{i2})'$, $i = 1, 2$, with dummy variables x_{ij}, $j = 1, 2$, indicating education level (low and high). Then, with y_1 and y_2 the incomes of males and females (irrespective of education level), stochastic dominance (2) implies reduced-rank multivariate linear models of conditional location, when specified in terms of either mean or median, and also implies a model of conditional variability, specified in terms of standard deviation. That is

$$\mu_i(y_i|x_i) = \theta'_{\mu i}\, x_i, \qquad m_i(y_i|x_i) = \theta'_{mi}\, x_i, \qquad \sigma_i(y_i|x_i) = \theta'_{\sigma i}\, x_i$$

for some 2×1 vectors $\theta_{\mu i}, \theta_{mi}, \theta_{\sigma i}$, $i = 1, 2$, which yield 2×2 matrices Θ_μ, Θ_m, Θ_σ having typical rows $\theta'_{\mu i}$, θ'_{mi}, $\theta'_{\sigma i}$, respectively. More generally, (2) implies a model (1) of conditional quantiles (including the median) and of higher-order (standardized) moments. In all of these models, linearity (ii) is not a strong assumption since x_i consists of dummy variables, and reduced-rank (iii) is implied by the stochastic dominance condition.

Suppose we want to compare incomes of males and females, by education, in terms of a basic descriptive such as mean, median, or standard deviation. Population descriptives can be estimated by their sample counterparts, or by maximizing the likelihood of some flexible parametric probability model like MN. In either case we can arrange mean values for the four gender × education groups into a matrix, and use a reduced-rank matrix restriction to state the idea that female mean income is equal to a constant times male mean income, with the constant being the same for each education category. To maximize MN likelihood subject to this restriction, we would need something like an extended EM algorithm, with alternating least squares woven into each EM pass. By comparison, the MAD estimator requires only simple EM followed by simple alternating least squares, both computable with existing software.

If we use median income, rather than mean income, as our income descriptive, and consider the matrix of median incomes (by gender and income), we can again use a reduced-rank matrix restriction to compare male and female incomes. Here, maximum likelihood estimation of the MN model, subject to reduced-rank on the matrix of group median values, is much more difficult computationally since the median of mixed normal distributions is generally not a closed-form function of model parameters. It is feasible via Monte Carlo or simulation methods, perhaps combined with grid search, but is not supported by existing software, and would impose quite a programming/computation burden on statisticians attempting it themselves. By comparison, the MAD estimator requires only simple EM followed by a single Monte Carlo (to compute median values, a simple exercise) and then simple alternating least squares.

3. Definitions

We define here the proposed estimator and test, and later explore their properties and performance. When reduced-rank holds there is a factorization of the coefficient matrix:

$$\Theta = AB, \tag{3}$$

with A and B being $g \times r$ and $r \times k$ full-rank matrices, respectively. With I_r the $r \times r$ identity matrix, we specify

$$A = \begin{bmatrix} I_r \\ C \end{bmatrix}, \tag{4}$$

with C some $(g - r) \times r$ matrix which we will call the multiplier matrix. The first r rows of Θ then form a basis, spanning the remaining rows, and we partition Θ as

$$\Theta = \begin{bmatrix} \Theta_1 \\ \Theta_2 \end{bmatrix}, \tag{5}$$

with Θ_1 the 'basis' sub-matrix consisting of the first r rows of Θ, and Θ_2 consisting of the last $g - r$ rows. Then, under (4), for the factorization $\Theta = AB$ we have

$$\Theta_1 = B, \tag{6}$$

$$\Theta_2 = C\Theta_1. \tag{7}$$

Let S^* be the set of $g \times k$ matrices whose first r rows are linearly independent and span the remaining rows. The reduced-rank form of interest is then the hypothesis
H_0: $\Theta \in S^*$.

To introduce the proposed methods, let $\phi = \operatorname{vec}\Theta'$ and $\hat{\phi} = \operatorname{vec}\hat{\Theta}'$ (with full-rank plug-in $\hat{\Theta}$), each $gk \times 1$ vectors, and let $f_*(\zeta; \mu, \Sigma)$ be a known family of probability density functions for $gk \times 1$ vectors ζ, with density parametrized by its $gk \times 1$ mean vector μ and $gk \times gk$ variance–covariance matrix Σ. Suppose that

$$\Omega^{-1/2}(\hat{\phi} - \phi) \xrightarrow{d} f_*(\cdot; 0, I),$$

for some $gk \times gk$ invertible variance–covariance matrix Ω which depends on sample size, with each element $\Omega_{ij} \to 0$ in large samples, and where $\Omega^{-1/2} = (\Omega^{1/2})^{-1}$ with Cholesky root $\Omega^{1/2}$: $\Omega^{1/2}(\Omega^{1/2})' = \Omega$. Note that we assume invertibility and hence full rank of covariance matrix Ω for the unrestricted estimator $\operatorname{vec}\hat{\Theta}'$, while at the same time entertaining reduced-rank in Θ. This is not a contradiction, provided that particular values of Θ elements are unrelated to the correlation among the elements of estimator vector $\operatorname{vec}\hat{\Theta}'$, as is the usual case in statistical estimation of multiple parameters, like in ANOVA, regression, etc. For example, if we the parameters of interest be the population mean income values for four groups (male—low education, male—high ed., female—low ed., female—high ed.), then these mean values, when arranged into a 2×2 matrix (by gender and education), may or may not satisfy a reduced-rank matrix condition. Regardless, provided that incomes are sampled at random from the population, when estimating population mean income by sample mean income—for each group, the resulting group means are uncorrelated, and the 4×1 vector $\operatorname{vec}\hat{\Theta}'$ has a 4×4 covariance matrix which is diagonal, with positive diagonal elements of the form σ_{ij}^2/n_{ij}, and hence is full-rank.

We define $f_{\hat{\phi}}(\zeta; \phi, \Omega) = f_*(\zeta; \phi, \Omega)$ as the asymptotic density function of $\hat{\phi}$. Let $\tilde{\phi}$ maximize the asymptotic density value $f_{\hat{\phi}}(\hat{\phi}; z, \hat{\Omega})$ over $z = \operatorname{vec} M'$ such that M lies in the set S^*, where $\hat{\Omega}$ is a plug-in (invertible) estimator of Ω, for which we assume that $\hat{\Omega}^{-1}\Omega \to I$ (in probability). We then call $\tilde{\phi}$ a MAD estimator, and call $\tilde{\Theta} = \tilde{A}\tilde{B}$ the MAD estimator of Θ, such that $\operatorname{vec}\tilde{\Theta}' = \tilde{\phi}$, with component estimators $\tilde{A} = [I_r, \tilde{C}']'$ and $\tilde{B}$.

To test H_0 we introduce a *Ratio of Asymptotic Densities* (RAD) test statistic:

$$W = -2\left(\ln\left(\frac{f_{\hat{\phi}}(\hat{\phi};\tilde{\phi},\hat{\Omega})}{f_{\hat{\phi}}(\hat{\phi};\hat{\phi},\hat{\Omega})}\right)\right),$$

which is based on the ratio $f_{\hat{\phi}}(\hat{\phi};\tilde{\phi},\hat{\Omega})/f_{\hat{\phi}}(\hat{\phi};\hat{\phi},\hat{\Omega})$ of restricted (via H_0) and unrestricted (asymptotic) density values.

In the remainder of this paper, we suppose that $\hat{\Theta}$ is asymptotically normal:

$$\Omega^{-1/2}\operatorname{vec}\left(\hat{\Theta}' - \Theta'\right) \to N(0, I). \tag{8}$$

Let $\mathcal{M}_{pq}$ be the set of $p \times q$ matrices, for some given p and q, and define the Mahalanobis metric

$$d(a, b; \Delta) \;=\; \left[\operatorname{vec}'(a' - b')\,\Delta\operatorname{vec}(\,a' - b')\right]^{1/2}$$

for each a and b in $\mathcal{M}_{pq}$ and some symmetric positive definite $pq \times pq$ matrix Δ. Then, under (8), the MAD estimator $\tilde{\Theta}$ minimizes $d(\hat{\Theta}, M; \hat{\Omega}^{-1})$ over $M \in S^*$, and hence is a "minimum distance" estimator, while RAD test statistic $W = d^2(\hat{\Theta}, \tilde{\Theta}; \hat{\Omega}^{-1})$. The decision rule for the proposed test is to reject H_0 if W exceeds the relevant critical value from the chi-square distribution with $(g-r)(k-r)$ degrees of freedom, in which case the test is a "minimum chi-square" test (alternatively called a "generalized Wald" test by [40]).

When suitably applied to multivariate models of conditional *mean* (as in MANOVA, regression, and errors-in-variables models), the proposed methods reduce to well-known maximum likelihood estimators and likelihood ratio (LR) tests. For example, in the context of Gaussian reduced-rank regression, if $\hat{\Theta}$ is the unconstrained MLE estimator, and $\hat{\Omega}$ is its maximum likelihood variance/covariance estimate, then $\tilde{\Theta}$ is a reduced-rank MLE and W is a likelihood ratio test statistic for H_0, as can be seen by applying [34, Theorem 3]) to [35, line 14 of p. 31]. Similarly, W can take the form of a Rao/score/Lagrange multiplier test when $\hat{\Omega}$ is obtained from constrained maximum likelihood. For models of conditional mean in which the errors can be non-normally distributed, the proposed estimator is not necessarily maximum likelihood but can take the form of "generalized least squares" (as in [21,43]).

4. Theory

To proceed, for each reduced-rank matrix $M \in S^*$ write $M = LQ$ for some $g \times r$ matrix $L = [I_r, N']'$, $r \times k$ matrix Q, and $(g-r) \times r$ matrix N. Then we can view $f_{\hat{\phi}}(\hat{\phi}; z, \hat{\Omega})$ as a function of vectors $v_1 = \operatorname{vec} Q'$ and $v_2 = \operatorname{vec} N'$, via $z = \operatorname{vec}([I_r, N']'Q)'$. Let $v = (v_1', v_2')'$ and $\psi = ((\operatorname{vec} B')', (\operatorname{vec} C')')'$, each an $(rk + (g-r)r) \times 1$ vector. Recalling the connection between $f_{\hat{\phi}}$ and distance $d(\hat{\Theta}, M; \hat{\Omega}^{-1})$, it is useful to write

$$d^2(\hat{\Theta}, LQ; \hat{\Omega}^{-1}) = \operatorname{vec}'(\hat{\Theta}' - Q'L')\,\hat{\Omega}^{-1}\operatorname{vec}(\hat{\Theta}' - Q'L').$$

There can occasionally be multiple MAD estimators $\tilde{\Theta}$, as when $g = 2 = k$, $\hat{\Theta} = I_2$ and $\hat{\Omega} = I_4$, where there are two (readily obtained) candidates for $\tilde{\Theta}$ and for $\tilde{\psi} = ((\operatorname{vec}\tilde{B}')', (\operatorname{vec}\tilde{C}')')'$, namely: $\tilde{\psi} = (1/2, 1/2, 1)'$, $\tilde{\Theta} = ((1/2, 1/2)', (1/2, 1/2)')$; and $\tilde{\psi} = (1, 0, 0)'$, $\tilde{\Theta} = ((1, 0)', (0, 0)')$, each of which yield $d(\hat{\Theta}, \tilde{\Theta}; \hat{\Omega}^{-1}) = 1$. In this case, the matrix $\hat{\Theta}$ is such that the first r

rows are orthogonal to the last $g-r$ rows, but such orthogonality must fail to hold (with probability approaching 1 in large samples, under (8)) if Θ satisfies H_0.

Noting that $\operatorname{vec} Q'L' = (L \otimes I_k) \operatorname{vec} Q'$, using the chain rule we have the $1 \times rk$ vector of partial derivatives of $\ln(f_{\hat{\phi}})$ with respect to v_1:

$$\frac{\partial \ln(f_{\hat{\phi}})}{\partial v_1} = \frac{\partial \ln(f_{\hat{\phi}})}{\partial z} \frac{\partial z}{\partial v_1} = \operatorname{vec}'(\hat{\Theta}' - Q'L')\,\hat{\Omega}^{-1}(L \otimes I_k). \tag{9}$$

Likewise, using the fact that $\operatorname{vec} Q'L' = (I_g \otimes Q') \operatorname{vec} L'$ we get the $1 \times (g-r)r$ vector:

$$\frac{\partial \ln(f_{\hat{\phi}})}{\partial v_2} = \frac{\partial \ln(f_{\hat{\phi}})}{\partial z} \frac{\partial z}{\partial v_2} = \operatorname{vec}'(\hat{\Theta}' - Q'L')\hat{\Omega}^{-1}(I_g \otimes Q')R, \tag{10}$$

where R is the $gr \times (g-r)r$ matrix:

$$R = \begin{bmatrix} 0_{r^2,(g-r)r} \\ I_{(g-r)r} \end{bmatrix} = \frac{\partial \operatorname{vec} L'}{\partial v_2},$$

with $0_{r^2,(g-r)r}$ the $r^2 \times (g-r)r$ matrix with all entries $= 0$.

Setting derivatives equal to zero, we obtain partial solutions for $\tilde{B}$ and $\tilde{C}$:

$$\operatorname{vec} \tilde{B}' = \left[(\tilde{A} \otimes I_k)'\,\hat{\Omega}^{-1}\,(\tilde{A} \otimes I_k)\right]^{-1}(\tilde{A} \otimes I_k)\,\hat{\Omega}^{-1} \operatorname{vec} \hat{\Theta}', \tag{11}$$

$$\operatorname{vec} \tilde{C}' = \left[((I_g \otimes \tilde{B}')R)'\,\hat{\Omega}^{-1}\,(I_g \otimes \tilde{B}')R\right]^{-1}((I_g \otimes \tilde{B}')R)'\,\hat{\Omega}^{-1} \operatorname{vec} \hat{\Theta}'. \tag{12}$$

The $(rk + (g-r)r) \times (rk + (g-r)r)$ Hessian matrix of second partial derivatives for $\ln(f_{\hat{\phi}})$ with respect to v is

$$H = \begin{bmatrix} \frac{\partial}{\partial v}\,(L \otimes I_k)'\,\hat{\Omega}^{-1} \operatorname{vec}(\hat{\Theta}' - Q'L') \\ \frac{\partial}{\partial v}\,R'(I_g \otimes Q')'\,\hat{\Omega}^{-1} \operatorname{vec}(\hat{\Theta}' - Q'L') \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}' & H_{22} \end{bmatrix},$$

with H_{11} the upper-left $rk \times rk$ sub-matrix of H, H_{12} the upper-right $rk \times (g-r)r$ sub-matrix, etc. Evaluating Q and N at $\tilde{B}$ and $\tilde{C}$, respectively, yields the result $\tilde{H}$ for H. Using the above-mentioned formulas relating $\operatorname{vec} Q'L'$ to $\operatorname{vec} Q'$ and $\operatorname{vec} L'$, respectively, we obtain

$$\tilde{H}_{11} = -(\tilde{A} \otimes I_k)'\,\hat{\Omega}^{-1}(\tilde{A} \otimes I_k), \tag{13}$$

$$\tilde{H}_{22} = -R'(I_g \otimes \tilde{B}')'\,\hat{\Omega}^{-1}(I_g \otimes \tilde{B}')R. \tag{14}$$

For the cross-derivative term $\tilde{H}_{12}$, we repeatedly make use of the chain rule and the fact that $\operatorname{vec}(L \otimes I_k)' = \operatorname{vec}(L' \otimes I_k) = (I_g \otimes G) \operatorname{vec} L'$ where G is the $k^2 r \times r$ matrix $(K_{kr} \otimes I_k)(I_r \otimes \operatorname{vec} I_k)$ and K_{kr} is the $kr \times kr$ commutation matrix (as discussed in [34, Chapters 3, 5]), for which $\operatorname{vec} U' = K_{kr} \operatorname{vec} U$ for each $k \times r$ matrix U. The result is

$$\tilde{H}_{12} = Z\left[I_g \otimes G\right]R, \tag{15}$$

with Z the $kr \times gk^2r$ matrix:

$$Z = -(\mathrm{vec}\,\hat{\Theta}')'\,\hat{\Omega}^{-1} \otimes I_{kr} + (\mathrm{vec}\,\tilde{B}')'\,(\tilde{A} \otimes I_k)'\hat{\Omega}^{-1} \otimes I_{rk} + (\tilde{A} \otimes I_k)'\,\hat{\Omega}^{-1} \otimes (\mathrm{vec}\,\tilde{B}')'.$$

Using the fact that $\tilde{\Theta} = \tilde{A}\tilde{B}$ is a (weakly) consistent estimator of Θ under H_0 and (8) (as is readily shown, and can be obtained from Lemma A.1 in the Appendix), we get a convenient asymptotic approximation $\tilde{H}_{12} \approx -(\tilde{A} \otimes I_k)'\,\hat{\Omega}^{-1}(I_g \otimes \tilde{B}')R$, where for sample-specific random matrices a and b, $a \approx b$ means that $a = b(1 + o_p(1))$, with $o_p(1)$ a term vanishing in probability in large samples. From this we obtain $-\tilde{H}V_{\tilde{\psi}} \xrightarrow{p} I_{rk+(g-r)r}$, where

$$V_{\tilde{\psi}} = [P'\Omega^{-1}P]^{-1},$$

with P the $gk \times (rk + (g-r)r)$ matrix:

$$P = (A \otimes I_k, (I_g \otimes B')R).$$

Partition $V_{\tilde{\psi}}$ as we did H, yielding upper-left $rk \times rk$ sub-matrix $V_{\tilde{\psi}11}$, etc. in which case (using the partitioned inverse formula) we have:

$$\begin{aligned} V_{\tilde{\psi}11} = \Big[& (A \otimes I_k)'\Omega^{-1}(A \otimes I_k) - \Big((A \otimes I_k)'\Omega^{-1}(I_g \otimes B')R\Big) \\ & \times \Big(R'(I_g \otimes B')'\Omega^{-1}(I_g \otimes B')R)\Big)^{-1} \Big((A \otimes I_k)'\Omega^{-1}(I_g \otimes B')R\Big)'\Big]^{-1}, \\ V_{\tilde{\psi}22} = \Big[& R'(I_g \otimes B')'\Omega^{-1}(I_g \otimes B')R - \Big((A \otimes I_k)'\Omega^{-1}(I_g \otimes B')R\Big)' \\ & \times \Big((A \otimes I_k)'\Omega^{-1}(A \otimes I_k)\Big)^{-1} \Big((A \otimes I_k)'\Omega^{-1}(I_g \otimes B')R\Big)\Big]^{-1}. \end{aligned}$$

Defining $V_{\tilde{B}} = V_{\tilde{\psi}_{11}}$ and $V_{\tilde{C}} = V_{\tilde{\psi}_{22}}$, we have:

Theorem 1. *Under* (8) *and* H_0, *each of the following holds*:

(i) $\tilde{\psi} - \psi \approx (P'\Omega^{-1}P)^{-1}P'\Omega^{-1}\,\mathrm{vec}(\hat{\Theta}' - \Theta')$,
(ii) $V_{\tilde{\psi}}^{-1/2}(\tilde{\psi} - \psi)$ *converges in distribution to* $N(0, I_{rk+(g-r)r})$,
(iii) $V_{\tilde{B}}^{-1/2}\,\mathrm{vec}(\tilde{B}' - B') \xrightarrow{d} N(0, I_{rk})$,
(iv) $V_{\tilde{C}}^{-1/2}\,\mathrm{vec}(\tilde{C}' - C') \xrightarrow{d} N(0, I_{(g-r)r})$.

The asymptotic variance matrices for vec $\tilde{B}'$ and vec $\tilde{C}'$ coincide (asymptotically) with $-\tilde{H}^{11}$ and $-\tilde{H}^{22}$, respectively, where $\tilde{H}^{ij}$ is the (i, j)th partitioned block of the inverse $\tilde{H}^{-1}$ of Hessian matrix $\tilde{H}$ (with partitioning as in H); hence the asymptotic theory of MAD estimators mimics classical asymptotics for maximum likelihood estimators. [45] exploits this sort of resemblance in his study of the likelihood ratio statistic (see also [42, p. 240]). We can further this resemblance by introducing the $(rk + (g-r)r) \times 1$ vector $\tilde{s} = \left(\frac{\partial \ln(f_{\hat{\phi}})}{\partial v_1}|_{M=\Theta}, \frac{\partial \ln(f_{\hat{\phi}})}{\partial v_2}|_{M=\Theta}\right)'$, consisting of partial derivatives (9) and (10) evaluated at $M = \Theta$, in which case, from Theorem 1 we conclude:

$$\tilde{\psi} - \psi \approx -\tilde{H}^{-1}\tilde{s},$$

mimicking the asymptotic behavior of maximum likelihood estimators (as described in [42], Section 5.5, for example).

It is interesting to interpret the asymptotic variance matrices $V_{\tilde{B}}$ and $V_{\tilde{C}}$ in light of formulas (11) and (12). If in (11) the value of A were known we could re-define $\tilde{A} = A$, in which case $\operatorname{vec} \tilde{B}'$ would be a linear function of $\operatorname{vec} \hat{\Theta}'$ and would have asymptotic variance matrix $[(A \otimes I_k)' \, \Omega^{-1} \, (A \otimes I_k)]^{-1}$, but with A unknown $V_{\tilde{A}}$ is larger (by a positive definite matrix) than this 'ideal' variance matrix. Similarly, $V_{\tilde{C}}$ is larger than the 'ideal' variance $[((I_g \otimes B')R)' \, \Omega^{-1} \, (I_g \otimes B')R]^{-1}$ that could be obtained for $\operatorname{vec} \tilde{C}'$ if B were known.

With $\tilde{\Theta} = \tilde{A}\tilde{B}$ we obtain the asymptotic distribution of $\tilde{\Theta}$ from that of its components:

Theorem 2. *Under* (8) *and* H_0, $\operatorname{vec}(\tilde{\Theta}' - \Theta') \approx P(P'\Omega^{-1}P)^{-1}P'\Omega^{-1}\operatorname{vec}(\hat{\Theta}' - \Theta')$, *and hence asymptotically* $\operatorname{vec}(\tilde{\Theta}' - \Theta')$ *is normal with zero mean and variance matrix* $V_{\tilde{\Theta}} = P\,(P'\Omega^{-1}P)^{-1}P'$.

To examine the proposed estimators in the context of probability models and likelihood functions, consider the following general situation. Let $\mathcal{L}(x; \pi)$ be a (generalized) log-likelihood function with some $a \times 1$ parameter vector π. Let the restricted form of the model have $\pi = q(v)$ for some $b \times 1$ vector v, $b < a$, and differentiable function q. Let $\pi^{\dagger}$ and $\hat{\pi}$ be the maximum likelihood estimators with and without the restriction, respectively, and let $v^{\dagger}$ be the MLE estimator of v. $\mathcal{L}_{\pi}$ is the $1 \times a$ vector of partial derivatives of $\mathcal{L}$ with respect to $\pi_1, \ldots, \pi_a$, and $\mathcal{L}_{\pi\pi'}$ is the $a \times a$ second derivative matrix of $\mathcal{L}$, each evaluated at π, and q_v is the $a \times b$ derivative matrix of q, evaluated at v. Define $a \times a$ matrix $V_{\hat{\pi}} = (-E\mathcal{L}_{\pi\pi'})^{-1}$. Let $\hat{V}_{\hat{\pi}}$ be an invertible estimate of $V_{\hat{\pi}}$. With $f_{\hat{\pi}}(\xi; \pi, V_{\hat{\pi}})$ the normal density function with mean vector π and variance matrix $V_{\hat{\pi}}$, let $\tilde{v}$ be the "MAD" estimator of v, maximizing the asymptotic density $f_{\hat{\pi}}(\hat{\pi}; q(u), \hat{V}_{\hat{\pi}})$ over u, and let $\tilde{\pi} = q(\tilde{v})$.

Assumption 1. Suppose that

(i) $\hat{\pi} - \pi \approx -(E\mathcal{L}_{\pi\pi'})^{-1}\mathcal{L}'_{\pi}$ and
(ii) $v^{\dagger} - v \approx -(q'_v E\mathcal{L}_{\pi\pi'} q_v)^{-1} q'_v \mathcal{L}'_{\pi}$,
(iii) $V_{\hat{\pi}}^{-1/2}(\hat{\pi} - \pi) \xrightarrow{d} N(0, I_a)$,
(iv) $V_{\hat{\pi}}$ converges to zero (element-wise) in large samples,
(v) $\hat{V}_{\hat{\pi}}^{-1} V_{\hat{\pi}}$ converges (in probability) to the identity matrix.

The conditions on the likelihood imposed by Assumption 1 are standard (see for example [42, Chapter 5.5]).

Theorem 3. *Under Assumption* 1, *MAD estimators are asymptotically equivalent to maximum likelihood estimators of the restricted model*: $\tilde{v} \approx v^{\dagger}$ *and* $\tilde{\pi} \approx \pi^{\dagger}$.

To apply Theorem 3 to our case of reduced-rank matrix estimators, let v be partitioned $v = (v'_1, v'_2)'$, with $v_1 = \psi$, and let π be partitioned as $\pi = (\pi'_1, \pi'_2)'$, with $\pi_1 = \phi$. Also, let $q(v) = (t(v_1)', v'_2)'$, with t: $\phi = t(\psi)$. The MAD estimator of π contains components $\tilde{\pi}_1$ and $\tilde{\pi}_2$, and because the specification $\pi_1 = t(v_1)$ and $\pi_2 = v_2$ allows π_1 and v_2 (likewise π_2 and v_1) to freely vary with respect to each other, $\tilde{\pi}_1$ minimizes $d(\hat{\pi}_1, t(u); \hat{V}_{\pi_1}^{-1})$ over u, with $V_{\hat{\pi}_1}$ the upper-left submatrix (corresponding to π_1) of $V_{\hat{\pi}}$. Setting $\hat{V}_{\pi_1} = \hat{\Omega}$, we have $d(\hat{\pi}_1, \tilde{\pi}_1; \hat{V}_{\pi_1}^{-1}) = d(\hat{\Theta}, \tilde{\Theta}; \hat{\Omega}^{-1})$, hence $\tilde{\pi}_1$ is of the form $\tilde{\phi}$, and $\tilde{v}_1$ is of the form $\tilde{\psi}$.

To compute the MAD reduced-rank matrix estimator $\tilde{\Theta}$ and its component matrices $\tilde{B}$ and $\tilde{C}$, various numerical routines are possible. A simple method is to start with the estimator $\hat{B} = \hat{\Theta}_1$ of B, plug this into (11) to get an estimate of C, then plug this C estimate into (12) to get an updated estimate of B, etc., until convergence. Another approach is the Newton–Raphson sequence: $\tilde{\psi}^{(j+1)} = \tilde{\psi}^{(j)} - H^{-1}(\tilde{\psi}^{(j)})\, s(\tilde{\psi}^{(j)})$, $j = 1, 2, \ldots$, given some initial value $\tilde{\psi}^{(1)}$, with H as above and s the matrix of first partial derivatives given by (9) and (10) (forming the upper and lower rows of s, respectively), each evaluated at $\tilde{\psi}^{(j)}$. Note that we do not here prove convergence of the computational routines, but recommend the first of these routines (which we have used extensively, with real data and in simulations, with no problems).

Regarding the proposed RAD test of reduced-rank we have:

Theorem 4. *Under* (8) *and H_0 the RAD test statistic W converges in distribution to chi-square, with $(g-r)(k-r)$ degrees of freedom.*

Further, writing $V_{\hat{\phi}} = \Omega$ we have

$$W \approx (\hat{\phi} - \tilde{\phi})' V_{\hat{\phi}}^{-1} (\hat{\phi} - \tilde{\phi}),$$

under (8) and H_0. This behavior of W imitates that of the likelihood ratio test, as we now explain. In the setting described in Assumption 1, define the likelihood ratio test statistic $LR = -2(\mathcal{L}^{(0)} - \mathcal{L}^{(1)})$, with $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(0)}$ the unconstrained and constrained log-likelihoods, respectively.

Assumption 2.

$$LR \approx (\hat{\pi} - \pi^{\dagger})' V_{\hat{\pi}}^{-1} (\hat{\pi} - \pi^{\dagger}).$$

This high-level assumption about LR is standard and valid under known low-level primitive conditions, such as smoothness of the data density function (see for example [42, Chapter 16]).

Theorem 5. *Under H_0 and Assumptions* 1 *and* 2, *the RAD test statistic W is* (*asymptotically*) *equivalent to the likelihood ratio test statistic LR.*

We can extend the test equivalence in Theorem 5 to local alternatives. For this, generalize Assumption 1 so that $V_{\hat{\pi}}^{-1/2}(\hat{\pi} - \pi_0) \overset{d}{\to} N(\delta, I_a)$, for some $\pi_0 = q(v_0)$, some v_0, and a vector δ. Also, in the Appendix setup for Lemmas A.1–A.3 let $V^{-1/2}(\hat{\mu} - \mu_0) \overset{d}{\to} N(\varepsilon, I_m)$, with μ_0 satisfying a hypothesized restriction on parameter vector μ, and a vector ε. Local alternatives arise when vectors δ and ε have non-zero elements. To cover this situation we can readily extend Theorem 3 under Assumption 2 and generalized Assumption 1, and from this find that the (local) power of the RAD test and likelihood ratio test are the same, given by the non-central chi square distribution $\chi^2_{(g-r)(k-r)}(\delta'\delta)$.

5. Example, continued

We apply the convenient approximate density (MAD and RAD) methods to income descriptives (mean, median, standard deviation) of males and females, in two ways. First, we use sample descriptives as the input to the approximate density methods. Second, we use as input descriptives obtained from the Mixture of Normals model.

To proceed with inputs given by sample descriptives, let the full-rank estimator $\hat{\Theta}$ consist of sample means, medians or standard deviations. For the estimated variance matrix $\hat{\Omega}$ of $\hat{\Theta}$, let all off-diagonal elements equal zero (since each two-way cell is sampled independently of the others) and, for diagonal elements $\hat{\Omega}_{mm}$ (with $m = 1, \ldots, 4$ corresponding to $(i, j) = (1, 1), (1, 2), (2, 1), (2, 2)$), (I) in the case of means let $\hat{\Omega}_{mm} = s_{ij}^2/n_{ij}$, where s_{ij}^2 and n_{ij} are the sample variance and sample size for ith sex × jth education level, (II) for medians let $\hat{\Omega}_{mm} = (y_{(n-k_{ij}+1)} - y_{(k_{ij})})^2/(4z_{0.995}^2)$, with $k_{ij} = (n_{ij}+1)/2 - z_{0.995}\sqrt{n_{ij}/4}$, $z_{0.995}$ the 0.995 quantile of the standard normal distribution, and $y_{(1)}, \ldots, y_{(n_{ij})}$ the (i, j)th cell's data in ascending order (see [44, p. 134]), (III) for standard deviations let $\hat{\Omega}_{mm} = (4n_{ij}s_{ij}^2)^{-1}((n_{ij}-1)^{-1}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij})^4 - (s_{ij}^2)^2)$.

To proceed with inputs given by descriptives of the Mixture of Normals model, let the full-rank estimator $\hat{\Theta}$ consist of means, medians or standard deviations associated with the MN models which in Section 2 we estimated via maximum likelihood. For the covariance $\hat{\Omega}$ of vec $\hat{\Theta}'$, we use $\hat{\Omega} = D_{\mathcal{E}}\hat{V}D'_{\mathcal{E}}$, with $\hat{V}$ the outer-product-of-scores estimator of covariance for the maximum likelihood parameter estimates, and with $D_{\mathcal{E}}$ the 1 × 5 vector of estimated partial derivatives for the MN model's descriptive $\mathcal{E}$—either mean, median, or standard deviation, with respect to the MN parameters:

$$D_\mu = \big(\mu_1 - \mu_2, \pi, 1 - \pi, 0, 0\big),$$

$$D_m = -\frac{1}{f(m)} \times \Big(F_\pi(m), F_{\mu_1}(m), F_{\mu_2}(m), F_{\sigma_1^2}(m), F_{\sigma_2^2}(m)\Big),$$

$$D_\sigma = q \times \Big(\sigma_1^2 + \mu_1^2 - (\sigma_2^2 + \mu_2^2) - 2(\pi\mu_1 + (1-\pi)\mu_2)(\mu_1 - \mu_2), 0, 0, \pi, (1-\pi)\Big),$$

with

$$q = \frac{1}{2\sqrt{\pi(\sigma_1^2 + \mu_1^2) + (1-\pi)(\sigma_2^2 + \mu_2^2) - (\pi\mu_1 + (1-\pi)\mu_2)^2}},$$

$f(m)$ the MN density function evaluated at the median m, and F_π, F_{μ_1}, F_{μ_2}, $F_{\sigma_1^2}$, $F_{\sigma_2^2}$ the partial derivatives of the MN cumulative distribution function, with respect to π, μ_1, etc., in which case $F_\pi(m) = \int_{-\infty}^{m} f_\pi(y)\,dy$, etc., with f_π the partial derivative of MN density f with respect to π, etc. For example, when $\mathcal{E}$ is given by population mean μ, in the MN model $\mu = \pi\mu_1 + (1-\pi)\mu_2$, a function of MN parameters, and D_μ contains the partial derivatives of this function $\mu(\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

For income effects measured by sample descriptives and, alternatively, Mixture Model descriptives, Table 3 reports estimates of reduced-rank matrix components, and their standard errors, and well as tests of reduced-rank in the 2 × 2 matrix Θ. To obtain standard errors for MAD estimators, we use the (asymptotically valid) variance matrix $V_{\tilde{\psi}}$ with unknown Ω, A, B replaced by $\hat{\Omega}$, $\tilde{A}$, $\tilde{B}$. With male and female income coefficients (by education level) given by the 1 × 2 row vectors Θ_1 and Θ_2, the reduced-rank ($r = 1$) restriction is $\Theta_2 = c\Theta_1$, and the proposed estimates of c are near 1/2 for each coefficient concept (mean, median, etc.), consistent with Table 1 and our earlier discussion. The proposed rank tests mostly fail to reject H_0, with p-values $\geqslant 0.20$ in all cases except for the test of median incomes, based on the MN model, where evidence is marginal—$p = 0.05$.

Table 3
Income MAD estimators and RAD test

Coefficient	c		b_1		b_2		Rank test	
concept	Est.	SD	Est.	SD	Est.	SD	Stat.	p
Based on sample descriptives								
Mean	0.545	0.02	21054.04	451.01	46040.11	1653.02	1.55	0.21
Median	0.500	0.02	16855.69	427.46	36851.49	1440.65	1.62	0.20
SD	0.547	0.04	19288.92	1106.27	41091.85	2987.49	1.67	0.20
Based on mixture model								
Mean	0.542	0.022	21172.61	728.10	46493.51	2018.04	0.91	0.34
Median	0.504	0.014	16793.28	325.18	37219.28	1067.16	3.81	0.05
SD	0.557	0.038	19057.44	1071.08	39694.13	3995.35	1.15	0.28

6. Performance

Let $g = 2$, $k = 2$ and $r = 1$, in which case $A = [1, c]'$, $B = (b_1, b_2)$ and $\Theta = [1, c]'(b_1, b_2)$, for some scalars b_1, b_2, c. Also, let each of the four (i, j) classifications have a sample of the same size n. To describe estimator performance we first obtain some asymptotic formulas, then report on some finite-sample simulations.

6.1. Asymptotics

For asymptotics we set $\Omega = \sigma^2 I_4/n$, for some $\sigma^2 > 0$ and sample size $n = 25, 50, 100, 200$. To analyze the proposed estimator $\tilde{\psi}$ of $\psi = (b_1, b_2, c)'$, we require the matrix P (defined earlier) which here takes the form

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ c & 0 & b_1 \\ 0 & c & b_2 \end{bmatrix}.$$

Applying Theorem 1 yields

$$\begin{bmatrix} \tilde{b}_1 - b_1 \\ \tilde{b}_2 - b_2 \\ \tilde{c} - c \end{bmatrix} \approx M_{\tilde{\psi}} \begin{bmatrix} \hat{\Theta}_{11} - \Theta_{11} \\ \hat{\Theta}_{12} - \Theta_{12} \\ \hat{\Theta}_{21} - \Theta_{21} \\ \hat{\Theta}_{22} - \Theta_{22} \end{bmatrix},$$

where $M_{\tilde{\psi}} = (P'\Omega^{-1}P)^{-1}P'\Omega^{-1}$ is the 3×4 matrix:

$$M_{\tilde{\psi}} = \frac{1}{(1+c^2)(b_1^2+b_2^2)} \times \begin{bmatrix} b_1^2(1+c^2)+b_2^2 & b_1 b_2 c^2 & b_2^2 c & -b_1 b_2 c \\ b_1 b_2 c^2 & b_1^2 + b_2^2(1+c^2) & -b_1 b_2 c & b_1^2 c \\ -b_1 c(1+c^2) & -b_2 c(1+c^2) & b_1(1+c^2) & b_2(1+c^2) \end{bmatrix}.$$

This ties the performance of $\tilde{\psi}$ explicitly to that of $\hat{\Theta}$. Further, we find by direct computation the Hessian matrix $\tilde{H}$ and the probability limit:

$$\operatorname{plim} n^{-1}\tilde{H} = -\frac{1}{\sigma^2}\begin{bmatrix} 1+c^2 & 0 & cb_1 \\ 0 & 1+c^2 & cb_2 \\ cb_1 & cb_2 & b_1^2+b_2^2 \end{bmatrix},$$

and using the fact that $-\tilde{H}V_{\tilde{\Phi}} \to I_3$ in probability, we compute $V_{\tilde{\psi}} = -n^{-1}(\operatorname{plim} n^{-1}\tilde{H})^{-1}$ to obtain

$$V_{\tilde{\psi}} = \frac{\sigma^2}{n(b_1^2+b_2^2)}\begin{bmatrix} b_1^2+b_2^2/(1+c^2) & b_1b_2c^2/(1+c^2) & -cb_1 \\ b_1b_2c^2/(1+c^2) & b_1^2/(1+c^2)+b_2^2 & -cb_2 \\ -cb_1 & -cb_2 & 1+c^2 \end{bmatrix},$$

which agrees with formula $V_{\tilde{\psi}} = (P'\Omega^{-1}P)^{-1}$ given in Section 4. With $\tilde{\psi} = (\tilde{b}_1, \tilde{b}_2, \tilde{c})'$, the asymptotic variance of $\tilde{b}_1$ and $\tilde{b}_2$ is falling in $|c|$, and the asymptotic variance of $\tilde{c}$ is falling in $|b_1|$ and $|b_2|$.

For reduced-rank estimation of Θ we have the proposed MAD estimator $\tilde{\Theta}$:

$$\tilde{\Theta} = \begin{bmatrix} \tilde{b}_1 & \tilde{b}_2 \\ \tilde{c}\,\tilde{b}_1 & \tilde{c}\,\tilde{b}_2 \end{bmatrix},$$

and applying Theorem 2 yields

$$\begin{bmatrix} \tilde{\Theta}_{11}-\Theta_{11} \\ \tilde{\Theta}_{12}-\Theta_{12} \\ \tilde{\Theta}_{21}-\Theta_{21} \\ \tilde{\Theta}_{22}-\Theta_{22} \end{bmatrix} \approx M_{\tilde{\phi}}\begin{bmatrix} \hat{\Theta}_{11}-\Theta_{11} \\ \hat{\Theta}_{12}-\Theta_{12} \\ \hat{\Theta}_{21}-\Theta_{21} \\ \hat{\Theta}_{22}-\Theta_{22} \end{bmatrix},$$

where $M_{\tilde{\phi}} = P(P'\Omega^{-1}P)^{-1}P'\Omega^{-1}$ is the 4 × 4 matrix:

$$M_{\tilde{\phi}} = \frac{1}{(1+c^2)(b_1^2+b_2^2)} \times \begin{bmatrix} b_1^2(1+c^2)+b_2^2 & b_1b_2c^2 & b_2^2c & -b_1b_2c \\ b_1b_2c^2 & b_1^2+b_2^2(1+c^2) & -b_1b_2c & b_1^2c \\ b_2^2c & -b_1b_2c & b_1^2(1+c^2)+b_2^2c^2 & b_1b_2 \\ -b_1b_2c & b_1^2c & b_1b_2 & b_1^2c^2+b_2^2(1+c^2) \end{bmatrix}.$$

This ties $\tilde{\Theta}$'s performance explicitly to that of $\hat{\Theta}$. Further, evaluating the asymptotic variance of $\tilde{\Theta}$ (as given in Theorem 2) yields

$$V_{\tilde{\Theta}} = M_{\tilde{\phi}}\frac{\sigma^2}{n},$$

in which case the elements of the MAD restricted estimator $\tilde{\Theta}$ have smaller asymptotic variance than those of the unrestricted estimator (which has asymptotic variance matrix $= \sigma^2 I_4/n$), to an extent that depends on the values of b and c.

6.2. Simulation—exponential model

Turning to finite-sample performance, we first simulate a model of random variables $y_{ij},\ i = 1, 2,\ j = 1, 2$, mutually independent with exponential distributions, having density functions:

$$f_{ij}(y) = \frac{1}{\theta_{ij}}\, e^{-y/\theta_{ij}}$$

for $y \geqslant 0$, $f_{ij}(y) = 0$ otherwise. In this model mean values are $E[y_{ij}] = \theta_{ij}$, and standard deviations are $\sigma[y_{ij}] = \theta_{ij}$. With Θ the 2×2 matrix having (i, j)th element θ_{ij}, we consider the case where matrix Θ has reduced rank = 1.

For the exponential multivariate model, subject to reduced rank, we study the approximate density (MAD and RAD) methods and maximum likelihood methods, in simulation. For the former, as plug-ins we use the unconstrained MLE $\hat{\theta}_{ij} = \bar{y}_{ij}$, with covariance matrix $\hat{\Omega}$, for vec $\hat{\Theta}'$, being diagonal with variance estimates of form $(\bar{y}_{ij})^2$ on the diagonal.

To obtain (constrained) reduced-rank MLE, we phrase the rank restriction as: $\Theta_{11} = b_1$, $\Theta_{12} = b_2, \Theta_{21} = cb_1, \Theta_{22} = cb_2$, for some b_1, b_2, c. With this form, we maximize the likelihood with respect to b_1, b_2, c. For this, we apply a repeated partial maximizations with respect to c and (b_1, b_2), iterating until convergence. Evaluating the restricted likelihood, and also the unrestricted likelihood, we then compute the likelihood ratio test statistic.

Using the above-described methods, we simulate the methods on pseudo-samples of size $n = 25, 50, 100$. For each sample size, we loop through 10,000 rounds of our sample-generating routine, each time-generating pseudo-data having the posited exponential distributions. The particular choice of parameter values is: $\theta_{11} = 1,\ \theta_{12} = 2,\ \theta_{21} = 2,\ \theta_{22} = 4$, in which case Θ has rank = 1, with $b_1 = 1,\ b_2 = 2,\ c = 2$. For each pseudo-sample we compute the desired statistics, recording the results. The result is a record of 10,000 trial values of the various statistics, and we summarize these in Table 4, with rej(10), rej(5), rej(1) being dummy variables that equal 1 if the relevant test statistic exceeds its critical value at significance level 10, 5, 1, respectively. The simulations are consistent with our earlier claim that the MAD estimator is asymptotically perfectly correlated with the (reduced-rank) maximum likelihood estimator. Also, the MAD estimator requires only unconstrained maximum likelihood estimation (trivial here) and a generically available minimum-distance algorithm, whereas maximum likelihood for the reduced-rank exponential model requires a tailor-made algorithm. The RAD test and LR test perform similarly, again consistent with theory.

6.3. Simulation—Student's t model

To further describe the approximate density (MAD and RAD) methods, consider again incomes for genders $i = 1, 2$ (male, female), and education levels $j = 1, 2$ (low, high). For each two-way classification, we create a pseudo-sample of sample size n, mutually independent realizations distributed as

$$y_{1j} = \mu_{1j} + \frac{\sigma_{1j}u_{1j}}{1.42}, \quad y_{2j} = c\mu_{1j} + c\,\frac{\sigma_{1j}\,u_{2j}}{1.42}, \quad j = 1, 2,$$

with u_{ij} a Student's t random variable (degrees of freedom = 4, matching income kurtosis, Table 1), where $(\mu_{11}, \mu_{12}) = (20871.82, 47767.38)$, $(\sigma_{11}, \sigma_{12}) = (18711.83, 43395.45)$, $c = 0.545$ and the constant 1.42 is such that the variables $u_{1j}/1.42$ and $u_{2j}/1.42$ have unit variance. Numerical values are chosen to approximate the income features reported in Table 1.

Table 4
Simulation, under exponential distribution

n	Statistic	Statistical method				Method correl.
		Approx. density		Likelihood		
		Mean	SD	Mean	SD	
25	c	2.04	0.43	2.12	0.48	0.89
	b_1	0.98	0.17	1.00	0.20	0.84
	b_2	1.96	0.35	2.00	0.40	0.84
	rej(10)	0.11		0.12		0.84
	rej(5)	0.06		0.06		0.81
	rej(1)	0.01		0.01		0.69
50	c	2.02	0.29	2.06	0.32	0.92
	b_1	0.99	0.12	1.00	0.14	0.85
	b_2	1.98	0.25	2.00	0.28	0.85
	rej(10)	0.10		0.11		0.87
	rej(5)	0.05		0.06		0.86
	rej(1)	0.01		0.01		0.75
100	c	2.01	0.20	2.03	0.22	0.93
	b_1	1.00	0.09	1.00	0.10	0.85
	b_2	1.99	0.18	2.00	0.20	0.86
	rej(10)	0.10		0.11		0.91
	rej(5)	0.05		0.05		0.88
	rej(1)	0.01		0.01		0.78

As in Section 5, let the 2×2 coefficient matrix Θ consist of population mean values (by gender and education), or medians, or standard deviations. For MAD and RAD methods we use as plug-in inputs $\hat{\Theta}$ and $\hat{\Omega}$ the sample descriptives and their associated covariance, the same ones as in Section 5.

With 10,000 simulation rounds, Table 5 reports on the performance of the MAD estimator and RAD test of reduced rank. Reported are the (simulation pseudo-sample) mean and standard deviation of the estimators, and the rejection rate (under H_0) for the RAD test at 5% significance level. The results suggest reasonable accuracy of the MAD estimator, even in smaller samples, and reasonable fidelity between the RAD test rejection rates and the claimed significance level (5%). An analogous simulation (omitted, for brevity), with standard normal u_{ij}, yields similar results. If desired we could also examine the MAD and RAD methods based on a Mixed Normal MLE plug-in, as described in Section 5, with potentially greater efficiency; however, we have found such simulations to take much longer, hence do not attempt a large number of them here.

7. Conclusion

The present work proposes reduced-rank estimators, and a test, of 'coefficient' matrices, with coefficients for multivariate linear models of features (such as mean, median, standard deviation) of conditional distributions. We demonstrate the feasibility of the methods, and give a first-order asymptotic theory for the proposed estimator. It would be interesting to attempt some second-order analysis of bias and variance, and to conduct a simulation study of the power of the proposed test.

Table 5
Simulation, under t distribution

n	Coefficient	MAD estimator						RAD
	concept	c		b_1		b_2		rej. rate
		Mean	SD	Mean	SD	Mean	SD	
25	Mean	0.554	0.10	20867.01	3178.92	47794.65	7342.20	0.054
	Median	0.553	0.10	20862.08	3089.65	47846.53	7141.72	0.026
	SD	0.557	0.13	17119.06	3603.43	39856.26	8471.59	0.059
50	Mean	0.550	0.07	20861.13	2268.24	47762.44	5193.21	0.053
	Median	0.550	0.07	20864.86	2145.99	47758.13	5000.90	0.036
	SD	0.551	0.09	17610.16	2718.04	40867.94	6415.60	0.051
100	Mean	0.547	0.05	20865.35	1594.68	47785.38	3628.10	0.055
	Median	0.546	0.05	20866.05	1519.53	47795.84	3521.67	0.044
	SD	0.548	0.07	17937.34	2100.88	41593.80	4819.97	0.041
200	Mean	0.546	0.03	20863.92	1120.12	47738.62	2636.02	0.047
	Median	0.546	0.03	20855.67	1067.55	47713.45	2517.72	0.045
	SD	0.547	0.05	18143.43	1597.39	42057.02	3650.23	0.045

Also, while the proposed reduced-rank coefficients estimator and rank test rely on an asymptotic normal distribution for the unrestricted coefficients estimator, we are currently pursuing the case of non-normal distributions (as arise in unit root time series), including error correction models of conditional medians.

Appendix A.

For an $m \times 1$ vector μ let $\mu = q(\lambda)$ for an (unknown, unique) $l \times 1$ vector λ, with $l < m$, and a (known) continuously differentiable function q. Let $q_\lambda(v) = \partial q(v)/\partial v$ be the $m \times l$ matrix of partial derivatives, and suppose that $q_\lambda(\lambda)$ is full-rank. Let $\hat{\mu}$ be an estimator for which $V^{-1/2}(\hat{\mu} - \mu) \xrightarrow{d} N(0, I_m)$ in large samples, where V is the variance/covariance matrix of $\hat{\mu}$, with (the elements of) $V \to 0$ in large samples. Let $\bar{\lambda}$ minimize $(\hat{\mu} - q(v))'\hat{V}^{-1}(\hat{\mu} - q(v))$ over v, with $\hat{V}$ a (positive definite) estimator of V such that $\hat{V}^{-1}V \xrightarrow{p} I_m$, and let $\bar{\mu} = q(\bar{\lambda})$.

Lemma A.1. $\bar{\lambda} - \lambda \approx ((q'_\lambda(\lambda)V^{-1}q_\lambda(\lambda))^{-1}q'_\lambda(\lambda)V^{-1}(\hat{\mu} - \mu)$, *and hence*

$$((q'_\lambda(\lambda)V^{-1}q_\lambda(\lambda))^{-1})^{-1/2}(\bar{\lambda} - \lambda) \xrightarrow{d} N(0, I_m),$$

in large samples.

Proof. The (weak) consistency of $\bar{\lambda}$ follows from that of $\hat{\mu}$, and for minimizer $\bar{\lambda}$ the first-order condition is $(\hat{\mu} - q(\bar{\lambda}))'\hat{V}^{-1}q_\lambda(\bar{\lambda}) = 0$. Further, since q is continuously differentiable and $q_\lambda(\lambda)$ has full rank, with the approximation $q(\bar{\lambda}) \approx q(\lambda) + q_\lambda(\lambda)(\bar{\lambda} - \lambda)$ the first-order condition yields $\bar{\lambda} - \lambda \approx ((q'_\lambda(\lambda)V^{-1}q_\lambda(\lambda))^{-1}q'_\lambda(\lambda)V^{-1}(\hat{\mu} - \mu)$. Since $\hat{\mu} \approx N(\mu, V)$, the result follows. □

Proof of Theorem 1. We apply Lemma A.1 with $\mu = \phi = \operatorname{vec}\Theta'$, $\lambda = \psi$, $\phi = q(\lambda)$ given by the restriction $H_0 : \Theta = (I_r, C')'B$, and $V = \Omega$.

We have two equivalent forms of q: $\phi = (A \otimes I_k)\lambda_B$ and $\phi = (I_g \otimes B)R\,\lambda_C$, where λ_B, λ_C partition λ into its first rk and last $(g-r)r$ elements. To compute $q_\lambda(\lambda)$ we proceed component-by-component, using (respectively) the two forms of q, in which case we arrive at $q_\lambda(\lambda) = (A \otimes I_k, (I_q \otimes B)R)$. Lemma A.1 then yields the desired result. □

Lemma A.2. *$q(\bar\lambda) \approx q(\lambda) + q_\lambda(\lambda)(\bar\lambda - \lambda)$, and hence, asymptotically, $\bar\mu$ is normal with mean vector μ and variance matrix $q_\lambda(q'_\lambda(\lambda)V^{-1}q_\lambda(\lambda))^{-1}q'_\lambda$.*

Proof. With $\bar\mu = q(\bar\lambda)$ we obtain $\bar\mu \approx q(\lambda) + q_\lambda(\lambda)(\bar\lambda - \lambda)$, so the result follows from Lemma A.1. □

Proof of Theorem 2. It suffices to apply Lemma A.2, with the same notational conventions as in the proof of Theorem 1, and with the fact that $V_{\tilde\psi} = (P'\Omega^{-1}P)^{-1}$. □

Proof of Theorem 3. Under Assumption 1, $\hat\pi - \pi \approx -(E\mathcal{L}_{\pi\pi'})^{-1}\mathcal{L}'_\pi$ and $v^\dagger - v \approx -(q'_v E\mathcal{L}_{\pi\pi'} q_v)^{-1} q'_v\mathcal{L}'_\pi$, so with $V_{\hat\pi} = (-E\mathcal{L}_{\pi\pi'})^{-1}$ we obtain

$$v^\dagger - v \approx (q'_v V_{\hat\pi}^{-1} q_v)^{-1} q'_v V_{\hat\pi}^{-1}(\hat\pi - \pi).$$

Applying Lemma A.1 with $\mu = \pi$ and $\lambda = v$ and $\bar\lambda = \tilde v$, we get

$$\tilde v - v \approx (q'_v V_{\hat\pi}^{-1} q_v)^{-1} q'_v V_{\hat\pi}^{-1}(\hat\pi - \pi),$$

hence $\tilde v \approx v^\dagger$. Moreover, with $\pi^\dagger = q(v^\dagger)$, weak consistency of $\hat\pi$ (implied by convergence of $V_{\hat\pi}$ to zero element-wise) implies weak consistency of $v^\dagger$ and $\pi^\dagger$, in which case $\pi^\dagger - \pi \approx q_v(v^\dagger - v)$. Hence

$$\pi^\dagger - \pi \approx q_v(q'_v V_{\hat\pi}^{-1} q_v)^{-1} q'_v V_{\hat\pi}^{-1}(\hat\pi - \pi).$$

Applying Lemma A.2 with $\mu = \pi$ and $\lambda = v$ and $\bar\lambda = \tilde v$, we conclude that $\tilde\pi - \pi \approx q_v(\tilde v - v)$. Hence $\tilde\pi \approx \pi^\dagger$. □

Lemma A.3. *$(\hat\mu - \bar\mu)'V^{-1}(\hat\mu - \bar\mu) \xrightarrow{d} \chi^2_{m-l}$ in large samples.*

Proof. Write $\bar\mu - \mu = q(\bar\lambda) - q(\lambda)$. From the proof of Lemma A.1, $q(\bar\lambda) - q(\lambda) \approx q_\lambda(\lambda)(\bar\lambda - \lambda)$, with $\bar\lambda - \lambda \approx ((q'_\lambda(\lambda)V^{-1}q_\lambda(\lambda))^{-1}q'_\lambda(\lambda)V^{-1}(\hat\mu - \mu)$. Hence, $\bar\mu - \mu \approx JV^{-1}(\hat\mu - \mu)$, with J the $m \times m$ matrix $J = q_\lambda(\lambda)(q'_\lambda(\lambda)V^{-1}q_\lambda(\lambda))^{-1}q'_\lambda(\lambda)$. Hence $(\hat\mu - \bar\mu)'V^{-1}(\hat\mu - \bar\mu) \approx (\hat\mu - \mu)'(I_m - J)'V^{-1}(I_m - J)(\hat\mu - \mu)$, and since the matrix $(I_m - J)'V^{-1}(I_m - J)$, when multiplied by V, is an idempotent matrix of rank $m - l$, the result follows from the fact that $\hat\mu \approx N(\mu, V)$. □

Proof of Theorem 4. It suffices to apply Lemma A.3, with the same notational conventions as in the proof of Theorems 1 and 2. □

Proof of Theorem 5. Follows from the equivalence of RAD and reduced rank estimators (Theorem 3). □

References

[1] S.K. Ahn, Inference for vector autoregressive models with cointegration and scalar components, J. Amer. Statist. Assoc. 92 (1997) 350–356.

[2] S.K. Ahn, G.C. Reinsel, Nested reduced-rank autoregressive models for multiple time series, J. Amer. Statist. Assoc. 83 (1988) 849–856.

[3] S.K. Ahn, G.C. Reinsel, Estimation of partially nonstationary multivariate autoregressive models, J. Amer. Statist. Assoc. 85 (1990) 813–823.

[4] T.W. Anderson, Estimating linear restrictions on regression coefficients for multivariate normal distributions, Ann. Math. Statist. 22 (1951) 327–351.

[5] T.W. Anderson, Estimation of linear functional relationships: approximate distributions and connections with simultaneous equations in econometrics (with discussion), J. Roy. Statist. Soc. Ser. B 38 (1976) 1–36.

[6] T.W. Anderson, Estimating linear statistical relationships, Ann. Statist. 12 (1984) 1–45.

[7] T.W. Anderson, Regression and ordered categorical variables (with discussion), J. Roy. Statist. Soc. Ser. B (1984) 29–34.

[8] T.W. Anderson, Trygve Haavelmo and simultaneous equations models, Scandinavian J. Statist. 18 (1991) 1–19.

[9] T.W. Anderson, Asymptotic theory for canonical correlation analysis, J. Multivariate Anal. 70 (1999) 1–29.

[10] T.W. Anderson, Asymptotic distribution of the reduced-rank regression estimator under general conditions, Ann. Statist. 27 (1999) 1141–1154.

[11] T.W. Anderson, H. Rubin, Statistical inference in factor analysis, in: J. Neyman (Ed.), Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 5, 1956, pp. 111–150.

[12] R.B. Banks, Growth and Diffusion Phenomena: Mathematical Frameworks and Applications, Springer, New York, 1994.

[13] G.S. Becker, Human Capital, third ed., University of Chicago Press, Chicago, 1993.

[14] F.D. Blau, L.M. Kahn, Gender differences in pay, J. Econ. Perspect. 14 (2000) 75–99.

[15] G.J. Borjas, Labor Economics, second ed., Irwin McGraw Hill, New York, 2000.

[16] R.J. Boscovitch, De litteraria expeditione per pontficiam dictionem et synopsis amplioris operis, ac habentur plura ejus ex exemplaria atiam sensorum impressa, in: Bononiensi Scientiarum et Artum Instituto atque Academia Commentarii, vol. 4, 1757, pp. 353–396.

[17] J.G. Cragg, S.G. Donald, On the asymptotic properties of LDU-based tests of the rank of a matrix, J. Amer. Statist. Assoc. 91 (1996) 1301–1309.

[18] J.G. Cragg, S.G. Donald, Inferring the rank of a matrix, J. Econometr. 76 (1997) 223–250.

[19] A. Dempster, N. Laird, D. Rubin, Maximum likelihood for incomplete data via the EM algorithm, J. Roy. Statist. Soc. Ser. B 39 (1977) 1–38.

[20] B.S. Everitt, D.J. Hand, Finite Mixture Distributions, Chapman & Hall, London, 1981.

[21] W. Fuller, Properties of some estimators for the errors-in-variables model, Ann. Statist. (1980) 407–422.

[22] W. Fuller, Measurement Error Models, Wiley, New York, 1987.

[23] S. Gilbert, P. Zemčík, Testing for latent factors in models with autocorrelation and heteroskedasticity of unknown form, South. Econ. J. 72 (July 2005) forthcoming.

[24] L. Gill, A. Lewbel, Testing the rank and definiteness of estimated matrices with applications to factor, state space, and ARMA models, J. Amer. Statist. Assoc. 87 (1992) 766–776.

[25] L.J. Gleser, Estimation in a multivariate " errors in variables" regression model, Ann. Statist. 9 (1981) 24–44.

[26] R. Gonin, A.H. Money, Nonlinear L_p-norm Estimation, Dekker, New York, 1989.

[27] J. Hamilton, Time Series Analysis, Princeton University Press, Princeton, NJ, 1994.

[28] P.J. Huber, Robust Statistics, Wiley, New York, 1981.

[29] A.J. Izenman, Reduced-rank regression for the multivariate linear model, J. Multivariate Anal. 5 (1975) 248–264.

[30] S. Johansen, Statistical analysis of cointegration vectors, J. Econ. Dynam. Control 12 (1988) 231–254.

[31] S. Johansen, Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models, Econometrica 59 (1991) 1551–1580.

[32] K.G. Jöreskog, A.S. Goldberger, Estimation of a model with multiple indicators and multiple causes of a single latent variable, J. Amer. Statist. Assoc. 70 (1975) 631–639.

[33] R. Koenker, Quantile regression, in: S. Fienberg, J. Kadane (Eds.), International Encyclopedia of the Social Sciences, Statistics Volume, 2002.

[34] J.R. Magnus, H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, revised ed., Wiley, New York, 1999.

[35] G.C. Reinsel, R.P. Velu, Multivariate Reduced-Rank Regression, Springer, New York, 1998.

[36] J.-M. Robin, R.J. Smith, Tests of rank, J. Econometr. Theory (2000) 151–175.
[37] S. Ruggles, M. Sobek, Integrated Public Use Microdata Series: Version 2.0. Historical Census Projects, University of Minnesota, Minneapolis, 1997.
[38] H. Schmidli, Reduced Rank Regression, Physica, Heidelberg, 1995.
[39] J.H. Stock, M.W. Watson, Testing for common trends, J. Amer. Statist. Assoc. 83 (1988) 1097–1107.
[40] J. Szroeter, Generalized Wald methods for testing nonlinear implicit and overidentifying restrictions, Econometrica 51 (1983) 335–353.
[41] R. van der Leeden, Reduced Rank Regression with Structured Residuals, DSWO Press, Leiden, 1990.
[42] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 1998.
[43] C. Villegas, Maximum likelihood and least squares estimation in linear and affine functional models, Ann. Statist. 10 (1982) 256–265.
[44] R.R. Wilcox, Applying Contemporary Statistical Techniques, Academic Press, New York, 2003.
[45] S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Statist. 19 (1938) 60–62.
[46] T. Yee, T. Hastie, Reduced-rank Vector Generalized Linear Models, Standford University, 2000, unpublished manuscript.
[47] A. Zellner, Estimation of regression relationships containing unobservable variables, Internat. Econ. Rev. 11 (1970) 441–454.
[48] R.F. Engle, C.W.J. Granger, Co-Integration and Error Correction: Representation, Estimation and Testing, Econometrica 55 (1987) 251–276.