Who's Afraid of Reduced-Rank Parameterizations of Multivariate Models? Theory and Example

Scott Gilbert
Petr Zemcik

CERGE-EI

Charles University
Center for Economic Research and Graduate Education
Academy of Sciences of the Czech Republic
Economics Institute

WORKING PAPER SERIES (ISSN 1211-3298)
Electronic Version

223

Who's Afraid of Reduced-Rank Parameterizations of Multivariate Models? Theory and Example

Scott Gilbert, Southern Illinois University Carbondale
Petr Zemčík, CERGE-EI, Prague

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, 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 difficult, 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.

Kdo se bojí parametrizace omezení hodnosti v modelech s více proměnnými? Teorie a příklad

Abstrakt

Omezení hodnosti matice mohou podstatně zjednodušit matici koeficientů v modelech s více promennými, ale jejich použití limituje složitost metod a jejich teorie. Náš článek se vydává jednodušší cestou se zaměřením na metodologické zobecnění a zároveň zjednodušení. Pro gaussovské a negaussovské modely více proměnných (v anglické literatuře označované GLM, GAM, atd.) poskytujeme jednotnou, explicitní teorii pro obecné asymptotické (normální) rozdělení estimátorů metody maximální věrohodnosti (EMMV). EMMV může mít složitou formu a nemusí být snadné jej spočítat, nicméně tuto překážku řešíme pomocí důkazu asymptotické ekvivalence mezi EMMV a relativně jednoduchým (Mahalanobis) estimátorem nejmenší vzdálenosti. Tato metoda je vhodná obzvláště pro testy omezení hodnosti matice a my popíšeme detailně její asymptotické vlastnosti v tomto kontextu. Navíc zahrneme studii metody v simulacích a analytických i empirických příkladech.

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 (Anderson 1951, Izenman 1975), which has been extensively researched (see below), is not yet included in most statistics textbooks, even at the graduate level, nor in most statistical software packages. A vicious cycle exists: a dearth of technical training and support leads to 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 (mle), 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 Φ that maps conditional distribution functions to functions of x. For each of g groups $i = 1, 2, ..., g$, with $g \leq k$, 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 $i = 1, 2, ..., g$, with coefficient k-vectors θ_i subject to reduced rank, meaning that the

$g \times k$ coefficient matrix $\Theta = (\theta_1, ..., \theta_g)'$ 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 the 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, the literature in financial economics (see Reinsel and Velu 1998, Ch. 8, for an excellent summary) takes the latter approach when modelling asset returns. Related to reduced-rank regression models are factor analysis, growth curve models, MIMIC 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 Anderson (1951, 1976, 1984a, 1991, 1999a,b), Anderson and Rubin (1956), Zellner (1970), Jöreskog and Goldberger (1975), Gleser (1981), Villegas (1982), Engle and Granger (1987), Fuller (1980,1987), Stock and Watson (1988), Ahn and Reinsel (1988,1990), van der Leeden (1990), Banks (1994), Schmidli (1995), Ahn (1997), and Reinsel and Velu (1998).

Reduced-rank parameterization has also been developed for some non-Gaussian multivariate models. These include the multinomial logit model (Anderson 1984), the vector generalized linear model (GLM) and vector generalized additive model (GAM), see Yee and Hastie (2000) for a recent discussion. Typically, in these models the matrix Θ pa-

rameterizes a feature ϕ which is not itself a (conditional) mean, but is related to the 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 $P(y \leq m(x)|x) = \frac{1}{2}$. 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 Boscovich (1757), and Gonin and Money (1989) provide a review of theory and some applications of such models (see also Huber 1981 for linear models of other location measures, and Koenker 2002 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 $\sigma_i(x) = \theta_i' x$, $i = 1, 2, ..., g$, 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), (ii) and (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). However, maximum likelihood is often not the simplest method, and it may be computationally burdensome. By comparison, a relatively simple "minimum (Mahalanobis) distance" estimator, or "maximum approximate density" (MAD) estimator, is typically available. 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 the MAD and mle estimators, these 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.

We assume that the plug-in $\hat{\Theta}$ is asymptotically normal, and this covers many cases of interest but not time series models with unit root dynamics, where $\hat{\Theta}$ can be asymptotically non-normal (see for example Johansen 1988, 1991, Ahn and Reinsel 1990 and Reinsel and Velu 1998, Ch. 5). 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 (mle). Also, the proposed estimator is identical, asymptotically, to constrained mle when $\hat{\Theta}$ is (unconstrained) mle. An advantage of the proposed method is its practicality, whereas constrained mle (for reduced-rank multivariate conditional variability, etc.) may be hard to compute (when available).

We also propose a rank test, based on the a 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 Gill and Lewbel 1992, Cragg and Donald 1996, 1997, Robin and Smith 2000) 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 ingredients (i), (ii) and (iii), all of which are applied to conditional mean, median and standard deviation.

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 i-th gender group and j-th 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 Ruggles and Sobek 1997 for description). This data is a random sample, from the year 1990, of U.S. 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 Becker 1993, Borjas 2000 and Blau and Kahn 2000), 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 $\times$ education pairing.

The data in Table 1 are consistent with the idea that women in 1990 tended to earn about half of what men did, in each education category. Formally,

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

where $\stackrel{d}{=}$ means equality in distribution, and c a constant close to $1/2$. This characterization, which (with $x > 0$) 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 model (1), define 2×1 vectors $x_i = (x_{i1}, x_{i2})$, 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.

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 $\Omega^{1/2}$: $\Omega^{1/2}(\Omega^{1/2})' = \Omega$. 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 *maximum asymptotic density* (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 (Part I of a two-part project), 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:

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

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 Szroeter 1983).

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 (mle) 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 Magnus and Neudecker (1999, Theorem 3) to Reinsel and Velu (1998, 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 Fuller 1980 and Villegas 1982).

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 $v_1 = \mathrm{vec}\, Q'$ and $v_2 = \mathrm{vec} N'$, via $z = \mathrm{vec}\,([I_r, N']'Q)'$. Let $v = (v_1', v_2')'$ and $\psi = ((\mathrm{vec}\, B')', (\mathrm{vec}\, C')')'$, 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:

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

Because $\tilde{\Theta}$ minimizes $d^2(\hat{\Theta}, LQ; \hat{\Omega}^{-1})$ over matrices M of the form LQ, we can deduce (readily) that the rows of $\tilde{\Theta}$ are each linear combinations of the rows of $\hat{\Theta}$, as are the rows of the 'basis' matrix estimator $\tilde{B}$, and the proposed estimation problem is equivalent to finding an optimal linearly dependent set of vectors each of which are linear combinations of $\hat{\Theta}$ rows. 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 $\tilde{\psi} = ((\mathrm{vec}\,\tilde{B}')', (\mathrm{vec}\,\tilde{C}')')'$, 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 $d(\hat{\Theta}, \tilde{\Theta}; \hat{\Omega}^{-1}) = 1$. More generally, when the matrix $\hat{\Theta}$ is such that the first r rows are orthogonal to the last $g-r$ rows, there can be multiple mad estimators $\tilde{\Theta}$, but this form of $\hat{\Theta}$ must fail to hold (with probability approaching 1 in large samples, under (8)) if Θ satisfies H_0.

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

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

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

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

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

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

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 Magnus and Neudecker 1999, Ch.'s 3, 5), for which $\operatorname{vec} U' = K_{kr} \operatorname{vec} U$ for each $k \times r$ matrix U. The result is:

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

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 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:

$$V_{\tilde{\psi}11} = [\,(A \otimes I_k)'\Omega^{-1}(A \otimes I_k) - \\ ((A \otimes I_k)'\Omega^{-1}(I_g \otimes B')R)\;(R'(I_g \otimes B')'\Omega^{-1}(I_g \otimes B')R))^{-1}\;((A \otimes I_k)'\Omega^{-1}(I_g \otimes B')R)'\,]^{-1},$$

$$V_{\tilde{\psi}22} = [\,R'(I_g \otimes B')'\Omega^{-1}(I_g \otimes B')R - \\ ((A \otimes I_k)'\Omega^{-1}(I_g \otimes B')R)'\;((A \otimes I_k)'\Omega^{-1}(A \otimes I_k))^{-1}\;((A \otimes I_k)'\Omega^{-1}(I_g \otimes B')R)\,]^{-1}.$$

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, $\tilde{\psi}-\psi \approx (P'\Omega^{-1}P)^{-1}P'\Omega^{-1}\operatorname{vec}(\hat{\Theta}'-\Theta')$ and $V_{\tilde{\psi}}^{-1/2}(\tilde{\psi}-\psi)$ converges in distribution to $N(0, I_{rk+(g-r)r})$, hence:

(i) $V_{\tilde{B}}^{-1/2}\operatorname{vec}(\tilde{B}'-B') \xrightarrow{d} N(0, I_{rk})$,

(ii) $V_{\tilde{C}}^{-1/2}\operatorname{vec}(\tilde{C}'-C') \xrightarrow{d} N(0, I_{(g-r)r})$.

The asymptotic variance matrices for $\operatorname{vec}\tilde{B}'$ and $\operatorname{vec}\tilde{C}'$ 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. Wilks (1938) exploits this sort of resemblance in his study of the likelihood ratio statistic (see also van der Vaart 1998, 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 van der Vaart 1998, 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, $\text{vec}\,(\tilde{\Theta}' - \Theta') \approx P(P'\Omega^{-1}P)^{-1}P'\Omega^{-1}\text{vec}\,(\hat{\Theta}' - \Theta')$, and hence asymptotically $\text{vec}\,(\tilde{\Theta}' - \Theta')$ 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:

Assumption 1: 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 $\pi = q(\nu)$ for some $b \times 1$ vector ν, $b < a$, and differentiable function q. Let $\pi^\dagger$ and $\hat{\pi}$ be the maximum likelihood estimators (mle's) with and without the restriction, respectively, and let $\nu^\dagger$ be the mle estimator of ν. Suppose that $\hat{\pi} - \pi \approx -(E\mathcal{L}_{\pi\pi'})^{-1}\mathcal{L}'_\pi$ and $\nu^\dagger - \nu \approx -(q'_\nu E\mathcal{L}_{\pi\pi'}q_\nu)^{-1}q'_\nu\mathcal{L}'_\pi$, where $\mathcal{L}_\pi$ is the $1 \times a$ vector of partial derivatives of $\mathcal{L}$ with respect to $\pi_1, ..., \pi_a$, and $\mathcal{L}_{\pi\pi'}$ is the $a \times a$ second derivative matrix of $\mathcal{L}$, each evaluated at π, and q_ν is the $a \times b$ derivative matrix of q, evaluated at ν. Also, suppose that $V_{\hat{\pi}}^{-1/2}(\hat{\pi} - \pi) \xrightarrow{d} N(0, I_a)$, with $a \times a$ matrix $V_{\hat{\pi}} = (-E\mathcal{L}_{\pi\pi'})^{-1}$ converging to zero (element-wise) in large samples. Let $\hat{V}_{\hat{\pi}}$ be an invertible estimate of $V_{\hat{\pi}}$ such that $\hat{V}_{\hat{\pi}}^{-1}V_{\hat{\pi}}$ converges (in probability) to the identity matrix, and with $f_{\hat{\pi}}(\xi;\pi,V_{\hat{\pi}})$ the normal density function with mean vector π and variance matrix $V_{\hat{\pi}}$ let $\tilde{\nu}$ be the 'mad' estimator of ν, maximizing the asymptotic density $f_{\hat{\pi}}(\hat{\pi};q(u),\hat{V}_{\hat{\pi}})$ over u, and let $\tilde{\pi} = q(\tilde{\nu})$.

The conditions on the likelihood imposed by Assumption 1 are standard (see for example van der Vaart 1998, Ch. 5.5).

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

To apply Theorem 3 to our case of reduced-rank matrix estimators, let ν be partitioned $\nu = (\nu_1', \nu_2')'$, with $\nu_1 = \psi$, and let π be partitioned as $\pi = (\pi_1', \pi_2')'$, with $\pi_1 = \phi$. Also, let $q(\nu) = (t(\nu_1)', \nu_2')'$, with t: $\phi = t(\psi)$. The mad estimator of π contains components $\tilde{\pi}_1$ and $\tilde{\pi}_2$, and because the specification $\pi_1 = t(\nu_1)$ and $\pi_2 = \nu_2$ allows π_1 and ν_2 (likewise π_2 and ν_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 sub-matrix (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 $\tilde{\nu}_1$ 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: $\tilde{\psi}^{(j+1)} = \tilde{\psi}^{(j)} - H^{-1}(\tilde{\psi}^{(j)})\, s(\tilde{\psi}^{(j)})$, $j = 1, 2, ...$, 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 covergence of the computational routines, but recommend the first of these routines (which we have used exensively, with real data and in simulations, with no problems). An easy-to-use computer program (in Microsoft Windows format), for implementing the first routine, is available from the first author upon request.

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 condition on LR is standard (as in van der Vaart 1998, Ch. 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 $V_{\hat{\pi}}^{-1/2}(\hat{\pi} - \pi_0) \xrightarrow{d} N(\delta, I_a)$, for some $\pi_0 = q(\nu_0)$, some ν_0, and a vector δ. Also, in the Appendix setup for Lemmas 1 - 3 let $V^{-1/2}(\hat{\mu} - \mu_0) \xrightarrow{d} N(\epsilon, I_m)$, 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

Applying the proposed methods to the income data, 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 $m = 1, ..., 4$ 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 i-th sex $\times$ j-th 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 $y_{(1)}, ..., y_{(n_{ij})}$ the (i, j)-th cell's data in ascending order (see Wilcox 2003, p. 134), (III) for standard deviations let $\hat{\Omega}_{mm} = (4n_{ij}s_{ij}^2)^{-1}((n_{ij} - 1)^{-1} \sum_{k=1}^{n_{ij}} \left(y_{ijk} - \bar{y}_{ij}\right)^4 - (s_{ij}^2)^2\big)$.

Table 2 reports point estimates of parameters, and their standard errors, as well as tests of reduced-rank in the 2×2 matrix Θ. To obtain standard errors for the mad estimator, 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 fail to reject H_0, with p-values ≥ 0.20 in each case.

In the case where coefficients are mean values we can interpret Θ as a matrix of regression coefficients (with regressors z_{ij} being dummy variables indicating the (i, j)-th classification), and here the mad estimates coincide with Gaussian maximum likelihood (reduced-rank) estimates.

To interpret the results, in terms of our earlier discussion (Section 2) the proposed methods suggest that in 1990 men's income strongly stochastically dominated that of women. Specifically, at each education level men's income appears to dominate that of women, in terms of central tendency (measured by mean or median) and variability

(measured by standard deviation). To check whether this "gender gap" has narrowed since 1990, the proposed methods could be applied to recent Census data.

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.

For asymptotics we allow the coefficient concept θ to be generic, and 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:

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

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:

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

which agrees with the 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:

$$M_{\tilde{\phi}} = \tfrac{1}{(1+c^2)(b_1^2+b_2^2)} \times$$

$$\begin{bmatrix} b_1^2(1+c^2)+b_2^2 & b_1b_2c^2 & b_2^2c & -b_1b_2c \\ b_1b_2c^2 & b_1^2+b_2^2(1+c^2) & -b_1b_2c & b_1^2c \\ b_2^2c & -b_1b_2c & b_1^2(1+c^2)+b_2^2c^2 & b_1b_2 \\ -b_1b_2c & b_1^2c & b_1b_2 & b_1^2c^2+b_2^2(1+c^2) \end{bmatrix}.$$

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.

Turning to finite-sample estimator performance, Table 3 reports the simulated sample mean and standard deviation of the proposed reduced-rank estimators of c, b_1, b_2, with $\hat{\Omega}$ given by the methods used in Section 5. The simulated data for the i-th group, $i = 1, 2$, is a pseudo-sample of sample size n, mutually independent realizations distributed as:

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

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. In this simulation model Θ has rank $r = 1$, and when the coefficient concept is mean or median we have $\Theta_1 = (\mu_{11}, \mu_{12})$ and $\Theta_2 = c\,\Theta_1$, while for the standard deviation coefficient concept we have $\Theta_1 = (\sigma_{11}, \sigma_{12})$ and $\Theta_2 = c\,\Theta_1$. The values of c, μ and σ, and the general setup, are such that the model matches reasonably the empirical results (Table 1) for our economic data, except that here we keep sample size fixed across categories (i, j).

With 10,000 simulation rounds, Table 3 reports on the performance of the proposed mad estimators 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 the 5% significance level. We notice some upward bias in the c estimator, for each coefficient concept, diminishing in larger samples. By comparison, consider the estimator of (the matrix) C minimizing $d(\hat{\Theta}_2, [I_r, N']'\hat{\Theta}_1; \hat{V}_{2|1})$, over $(g-r)\times r$ matrices N, where $V_{2|1}$ is the variance-covariance matrix of $\operatorname{vec}\hat{\Theta}_2'$ conditional on $\operatorname{vec}\hat{\Theta}_1'$, and $\hat{V}_{2|1}$ is a consistent estimate of $V_{2|1}$. In simulation we have found versions of this C estimator to be near-perfectly correlated with the mad estimator $\tilde{C}$ in larger samples, and to frequently have less bias but greater variance than $\tilde{C}$. For the proposed test, rejection rates are close to the nominal 5% rate, with some under-rejection (in the case of the median coefficient concept) that diminishes in larger samples.

An analogous simulation (omitted, for brevity) with standard normal u_{ij} yields similar

results, as does a non-parametric bootstrap method where (I) n values y_{1j}, for each $j = 1, 2$, are drawn at random (with replacement) from the available sample, creating the 'male' pseudo-sample, then (II) another n values of y_{1j} are drawn, then each multiplied by c, creating a 'female' pseudo-sample.

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. 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

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 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 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 1 then yields the desired result.

Lemma 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 1.

Proof of Theorem 2: It suffices to apply Lemma 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 $\nu^\dagger - \nu \approx -(q'_\nu E\mathcal{L}_{\pi\pi'}q_\nu)^{-1}q'_\nu\mathcal{L}'_\pi$, so with $V_{\hat{\pi}} = (-E\mathcal{L}_{\pi\pi'})^{-1}$ we obtain:

$$\nu^\dagger - \nu \approx (q'_\nu V_{\hat{\pi}}^{-1}q_\nu)^{-1}q'_\nu V_{\hat{\pi}}^{-1}(\hat{\pi} - \pi) \ .$$

Applying Lemma 1 with $\mu = \pi$ and $\lambda = \nu$ and $\bar{\lambda} = \tilde{\nu}$, we get:

$$\tilde{\nu} - \nu \approx (q'_\nu V_{\hat{\pi}}^{-1}q_\nu)^{-1}q'_\nu V_{\hat{\pi}}^{-1}(\hat{\pi} - \pi) \ ,$$

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

$$\pi^\dagger - \pi \approx q_\nu(q'_\nu V_{\hat{\pi}}^{-1}q_\nu)^{-1}q'_\nu V_{\hat{\pi}}^{-1}(\hat{\pi} - \pi) \ .$$

Applying Lemma 2 with $\mu = \pi$ and $\lambda = \nu$ and $\bar{\lambda} = \tilde{\nu}$, we conclude that $\tilde{\pi} - \pi \approx q_\nu(\tilde{\nu} - \nu)$. Hence $\tilde{\pi} \approx \pi^\dagger$.

Lemma 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 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 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

Ahn, S. K. (1997), "Inference for vector autoregressive models with cointegration and scalar components," *Journal of the American Statistical Association* 92, 350-356.

Ahn, S. K. and G. C. Reinsel (1988), "Nested reduced-rank autoregressive models for multiple time series," *Journal of the American Statistical Association* 83, 849-856.

Ahn, S. K. and G. C. Reinsel (1990), "Estimation of partially nonstationary multivariate autoregressive models," *Journal of the American Statistical Association* 85, 813-823.

Anderson, T. W. (1951), "Estimating linear restrictions on regression coefficients for multivariate normal distributions," *Annals of Mathematical Statistics* 22, 327-351.

Anderson, T. W. (1976), "Estimation of linear functional relationships: Approximate distributions and connections with simultaneous equations in econometrics (with discussion)," *Journal of the Royal Statistical Society*, Series B, 38, 1-36.

Anderson, T. W. (1984a), "Estimating linear statistical relationships," *Annals of Statistics* 12, 1-45.

Anderson, T. W. (1984b), "Regression and ordered categorical variables (with discussion), *Journal of the Royal Statistical Society*, Series B, 29-34.

Anderson, T. W. (1991), "Trygve Haavelmo and simultaneous equations models," *Scandanavian Journal of Statistics* 18, 1-19.

Anderson, T. W. (1999a), "Asymptotic theory for canonical correlation analysis," *Journal of Multivariate Analysis* 70, 1-29.

Anderson, T. W. (1999b), "Asymptotic distribution of the reduced-rank regression estimator under general conditions," *The Annals of Statistics* 27, 1141-1154.

Anderson, T. W. and H. Rubin (1956), "Statistical inference in factor analysis," In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 5, ed. J. Neyman, pp. 111-150.

Banks, R. B. (1994), *Growth and diffusion phenomena: Mathematical frameworks and applications*. New York: Springer-Verlag.

Becker, G. S. (1993), *Human Capital*, 3rd edition. Chicago: University of Chicago Press.

Blau, F. D. and L. M. Kahn (2000), "Gender differences in pay," *Journal of Economic Perspectives* 14, 75-99.

Borjas, G. J. (2000), *Labor Economics*, 2nd edition. New York: Irwin McGraw Hill.

Boscovitch, R. J. (1757), "De litteraria expeditione per pontficiam dictionem et synopsis amplioris operis, ac habentur plura ejus ex exemplaria atiam sensorum impressa," ini *Bononiensi Scientiarum et Artum Instituto atque Academia Commentarii* 4, 353-396.

Cragg, J. G. and S. G. Donald (1996), "On the asymptotic properties of LDU-based tests of the rank of a matrix", *Journal of the American Statistical Association* 91, 1301-1309.

Cragg, J. G. and S. G. Donald (1997), "Inferring the rank of a matrix," *Journal of Econometrics* 76, 223-250.

Fuller, W. (1980), "Properties of some estimators for the errors-in-variables model," *The Annals of Statistics*, 407-422.

Fuller, W. (1987), *Measurement Error Models*. New York: Wiley.

Gill, L. and A. Lewbel (1992), "Testing the rank and definiteness of estimated matrices with applications to factor, state space, and ARMA models," *Journal of the American Statistical Association* 87, 766-776.

Gleser, L. J. (1981), "Estimation in a multivariate 'errors in variables' regression model," *The Annals of Statistics* 9, 24-44.

Gonin, R. and A. H. Money (1989), *Nonlinear L_p-norm Estimation.* New York: Dekker.

Huber, P. J. (1981), *Robust Statistics.* New York: Wiley.

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

Johansen, S. (1988), "Statistical analysis of cointegration vectors," *Journal of Economic Dynamics and Control* 12, 231-254.

Johansen S. (1991), "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models," *Econometrica* 59, 1551-1580.

Jöreskog, K. G. and A. S. Goldberger (1975), "Estimation of a model with multiple indicators and multiple causes of a single latent variable," *Journal of the American Statistical Association* 70, 631-639.

Koenker, R. (2002), "Quantile regression," in the *International Encyclopedia of the Social Sciences,* Statistics Volume, eds. S. Fienberg and J. Kadane.

Magnus, J. R. and H. Neudecker (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, revised edition. New York: Wiley.

Reinsel, G. C. and R. P. Velu (1998), *Multivariate Reduced-Rank Regression.* New York: Springer.

Robin, J.-M. and R. J. Smith (2000), "Tests of rank," *Journal of Econometric Theory*, 151-175.

Ruggles, S. and M. Sobek (1997), Integrated Public Use Microdata Series: Version 2.0. Historical Census Projects, University of Minnesota, Minneapolis.

Schmidli, H. (1995), *Reduced Rank Regression*. Heidelberg: Physica.

Stock, J. H. and M. W. Watson (1988), "Testing for common trends," *Journal of the American Statistical Association* 83, 1097-1107.

Szroeter, J. (1983), "Generalized Wald methods for testing nonlinear implicit and overidentifying restrictions," *Econometrica* 51, 335-353

van der Leeden, R. (1990), *Reduced Rank Regression with Structured Residuals*. Leiden: DSWO Press.

van der Vaart, A. W. (1998), *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Villegas, C. (1982), "Maximum likelihood and least squares estimation in linear and affine functional models," *The Annals of Statistics* 10, 256-265.

Wilcox, R. R. (2003), *Applying Contemporary Statistical Techniques*. New York: Academic Press.

Wilks, S. S. (1938), "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Annals of Mathematical Statistics* 19, 60-62.

Yee, T. and T. Hastie (2000) "Reduced-rank Vector Generalized Linear Models," unpublished manuscript, Standford University.

Zellner, A. (1970), "Estimation of regression relationships containing unobservable variables," *International Economic Review* 11, 441-454.

TABLE 1. Income statistics, by Sex and Education

		male		female	
		low ed	high ed	low ed	high ed
income	n	1,556	403	1,632	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
	std dev	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 coefficients, reduced-rank method

coefficient	c		b_1		b_2		test	
concept	est.	s.d.	est.	s.d.	est.	s.d.	stat.	p
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
st.dev.	0.547	0.04	19288.92	1106.27	41091.85	2987.49	1.67	0.20

TABLE 3. Estimator Simulation Results

		estimator						
sample	coefficient	c		b_1		b_2		test rej.
size	concept	mean	s.d.	mean	s.d.	mean	s.d.	rate
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
	s.d.	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
	s.d.	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
	s.d.	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
	s.d.	0.547	0.05	18143.43	1597.39	42057.02	3650.23	0.045

CERGE-EI
P.O.BOX 882
Politických vezòù 7
111 21 Prague 1
Czech Republic
http://www.cerge-ei.cz