Journal of Quantitative Economics (2022) 20 (Suppl 1):S115–S132
https://doi.org/10.1007/s40953-022-00322-x

ORIGINAL ARTICLE

Factor Analysis Regression for Predictive Modeling with High-Dimensional Data

Randy Carter[1] · **Netsanet Michael**[2]

Accepted: 8 June 2022 / Published online: 23 August 2022

Abstract

Factor analysis regression (FAR) of y_i on $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$, $i = 1,2,...,n$, has been studied only in the low-dimensional case ($p < n$), using maximum likelihood (ML) factor extraction. The ML method breaks down in high-dimensional cases ($p > n$). In this paper, we develop a high-dimensional version of FAR based on a computationally efficient method of factor extraction. We compare the performance of our high-dimensional FAR with partial least squares regression (PLSR) and principal component regression (PCR) under three underlying correlation structures: arbitrary correlation, factor model correlation structure, and when y is independent of $\boldsymbol{x}$. Under each structure, we generated Monte Carlo training samples of sizes $n < p$ from a multivariate normal distribution with each structure. Parameters were fixed at estimates obtained from analyses of real data sets. Given the independence structure, we observed severe over-fitting by PLSR compared to FAR and PCR. Under the two dependent structures, FAR had a notably better average mean square error of prediction than PCR. The performance of FAR and PLSR were not notably different given the dependent structures. Thus, overall, FAR performed better than either PLSR or PCR.

Keywords Bilinear factor model · Principal component analysis · Principal component regression · Partial least squares · Factor structure covariance matrix · Factor analysis regression · Mean square error of prediction · Monte Carlo studies · Cross-validation

Randy Carter and Netsanet Michael contributed equally to this work.

Prepared for the special issue of Journal of Quantitative Economics in honor of Prof. C. R. Rao.

✉ Randy Carter
rcarter@buffalo.edu

Extended author information available on the last page of the article

Introduction

In today's data-rich world, analysts often are faced with *high-dimensional data*, where the number of variables is greater than the sample size n. The focus of this paper is *high-dimension regression* and prediction with p predictors, where $p > n$ and special methods are required. The need to estimate such models occurs often in biomedical science, where high volumes of data are generated by 'omics' and imaging studies. High-dimension regression problems in economics and finance are emerging (Kalina 2017) as data sets with large numbers of variables are becoming more prevalent in advertising, insurance, portfolio optimization, risk management, labor market dynamics, customer analytics, finance, the automotive industry, and stock market dynamics (Kalina 2017; Belloni et al. 2014; Fan et al. 2015). Econometricians working in these areas will need to apply or adapt existing methods or develop new ones as the demand for high-dimension analyses increases. The most common solutions to high-dimensional regression problems involve:

(1) **Dimension reduction and component regression**. Solutions of this type include Principal Components Regression (PCR) (see, for example, Rao (1996)), using unsupervised component selection, and Partial Least Squares regression (PLSR) (Wold 1966), using supervised component selection. Such methods reduce the dimension of the predictor space by replacing the high-dimensional vector of predictors, $\boldsymbol{x}$, with k orthogonal linear combinations (called components), where $k << n$, and then use OLS to regress y on the k components.
(2) **Sparse model estimation**. These solutions include Lasso regression (Tibshirani 1996) and related methods (e.g., Elastic Net regression (Zou and Hastie 2005)). Lasso-like methods shrink estimates toward zero and fix estimates of coefficients on "non-predictive" variables at zero. They are, therefore, ideally suited for fitting sparse models (models where $p - k$ coefficients are 0, $k << n$) and for variable selection.
(3) **Latent variable models**. Factor analysis regression (FAR) was proposed by Scott (1966), based on the assumption that associations among the $\boldsymbol{x}$ and y variables are sometimes driven by an underlying FA structure. Given the factor model with $k << n$ factors, high-dimensional regression problems are reduced to the low-dimensional problem of estimating the factor model and associated structured covariance/correlation matrix. PCR and PLSR, usually thought of (correctly so) as dimension reduction/component regression methods, can also be viewed as latent variable model-based methods. Wold, in fact, first proposed PLSR based on a latent variable model conceptualization (Wold 1966). (The method, however, amounts to a forward selection component regression model building procedure). Rao presented a latent variable model problem that is solved by PCA and offered a solution to the prediction problem using PCR (Rao 1996). FAR, instead, focuses on estimating the structured covariance/correlation matrix given the latent variable model and then estimates regression parameters from it.

It seems self-evident that sparse model estimation methods, such as Lasso, are preferred when the high-dimensional regression model is sparse and that latent variable models are preferred when the correlations among $(y, \boldsymbol{x})$ are driven by these observable variables' common relationships with underlying latent variables. Subject area experts are likely to know whether sparseness or latent factor structure is the more reasonable assumption in any given application. The choice between these two general approaches, therefore, can usually be made by study investigators. Additional research, however, is needed for guidance on the choice between FAR, PCR, and PLSR when a latent model is assumed.

FAR has received relatively little attention in the literature and in practice. PCR and PLSR are much more favored, in spite of their shortcomings. It is well known that PCR suffers, at least in some samples, from the fact that it is an unsupervised method of component selection (Hadi and Ling 1998). Thus, the resulting predictors may not retain all relevant information for predicting y. PLSR was developed to solve this problem (Frank and Friedman 1993; Garthwaite 1994; Helland 2010; Wold 1966; Wold et al. 2001) but is a forward selection method of model building and likely suffers from the well-known deficiency of forward selection methods when $p >> n$ (Efron et al. 2004).

To our knowledge, all previous studies of FAR used maximum likelihood (ML) factor extraction, which is only possible in low-dimensional settings. PC-based factor extraction is possible with high-dimensional data but requires computation of eigenvalues and vectors of a $p \times p$ matrix, which can be burdensome when p is very large.

In this paper we present a high-dimensional version of FAR and derive a computationally efficient method of factor extraction from high-dimensional data. The resulting high-dimensional FAR is compared to PCR and PLSR in two simulation studies.

Model and Coefficient Definition

The high-dimensional regression model written in terms of observations of y and $\boldsymbol{x}$ is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\gamma} + \mathcal{E} \tag{1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations on response variable y, $\mathbf{X}$ is the $n \times p$ matrix of observations of $\boldsymbol{x}$, $\boldsymbol{\gamma}$ is a $p \times 1$ vector of parameters, and $\mathcal{E}$ is an $n \times 1$ vector of model errors with expectation zero and variance σ^2. We assume throughout (without loss of generality) that the observations in $\mathbf{Y}$ and each column of $\mathbf{X}$ are standardized to have sample mean zero and variance one.

Define the $(p+1) \times 1$ vector $\boldsymbol{z} = (y, \boldsymbol{x})'$. The associated population correlation matrix is denoted by:

$$\mathbf{R}_z = \begin{pmatrix} 1 & \mathbf{R}'_{xy} \\ \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{pmatrix} \tag{2}$$

where $\mathbf{R}_{xx}$ is the $p \times p$ population correlation matrix of $\boldsymbol{x}$, and $\mathbf{R}_{xy}$ is the $p \times 1$ vector of population correlations between $\boldsymbol{x}$ and y. The coefficient parameter $\boldsymbol{\gamma}$ in (1) is defined by

$$\boldsymbol{\gamma} = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}. \tag{3}$$

Given a sample of size n, the sample correlation matrix, denoted by $\hat{\mathbf{R}}_z$, is partitioned identical to $\mathbf{R}_z$, where $\hat{\mathbf{R}}_{xy} = \mathbf{X}'\mathbf{Y}/n - 1$ and $\hat{\mathbf{R}}_{xx} = \mathbf{X}'\mathbf{X}/n - 1$. If $\mathbf{X}$ is full column rank, then $\hat{\boldsymbol{\gamma}}^{ols} = \hat{\mathbf{R}}_{xx}^{-1}\hat{\mathbf{R}}_{xy}$ is the best linear unbiased estimate (BLUE) of $\boldsymbol{\gamma}$. With high-dimensional data ($p > n$), $\mathbf{X}$ is not full rank and the BLUE doesn't exist.

Principal Components Regression

PCA dimension reduction amounts to projecting the columns of $\mathbf{X}$ onto a reduced dimension subspace of the column space of $\mathbf{X}$ that is spanned by the observation vectors of, say, k uncorrelated linear combinations of the covariates $\boldsymbol{x}$, called Principal Components (PCs) (Jolliffe 2005; Zhang et al. 2003; Friedman et al. 2009). In PCR the predictor variable space is reduced from p to k dimensions by replacing $\boldsymbol{x}$ in the regression model with the first $k < p$ PCs. The PCR estimate of $\boldsymbol{\gamma}$ is calculated from the retained components' coefficient vectors and the estimated coefficient vector from the regression of $\mathbf{Y}$ on the k retained components.

The first PC is defined as $c_1 = \boldsymbol{x}w_1$, where $\boldsymbol{w}_1'\boldsymbol{w}_1 = 1$ and $\boldsymbol{w}_1 = \arg\max_{\boldsymbol{w}'\boldsymbol{w}=1}(\boldsymbol{w}'\hat{\mathbf{R}}_{xx}\boldsymbol{w})$. The remaining $p - 1$ PCs are then defined as

$$\begin{aligned} \boldsymbol{w}_r = \quad & \operatorname{argmax} \quad (\boldsymbol{w}'\hat{\mathbf{R}}_{xx}\boldsymbol{w}) \\ & \boldsymbol{w}'\boldsymbol{w} = 1 \\ & \boldsymbol{w}'\hat{\mathbf{R}}_{xx}\boldsymbol{w}_l = 0 \end{aligned} \tag{4}$$

$l = 1, 2, \ldots, r-1$, $r = 2, 3, \ldots, p$. This produces components $\boldsymbol{c} = (c_1, c_2, \ldots, c_p)$ that are uncorrelated, with $Var(c_1) \le Var(c_2) \le Var(c_3) \le \cdots \le Var(c_p)$. The $\boldsymbol{w}_r$ vectors, $r = 1, 2, \ldots, p$ are computed as the p orthonormal eigenvectors of $\hat{\mathbf{R}}_{xx}$ associated with the largest to smallest eigenvalues, respectively. Varances of the PCs are the associated eigenvalues. Dimension reduction from the p dimensional space spanned by the columns of $\mathbf{X}$ to a $k < p$ dimensional space is achieved by retaining only the first k PCs. An $n \times p$ matrix of observed values of the full set of PCs is calculated as $\mathbf{C} = (\mathbf{C}_1, \mathbf{C}_2) = \mathbf{X}(\mathbf{W}_1, \mathbf{W}_2)$, where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the $p \times k$ and $p \times (p - k)$ matrices formed from the first k and last $p - k$ eigenvectors. A PCR is then completed by choosing k, regressing y on the retained PCs, writing $\mathbf{C}_1\hat{\boldsymbol{\beta}}^{pcr}$ as $\mathbf{X}\mathbf{W}_1\hat{\boldsymbol{\beta}}^{pcr}$, and calculating estimates of $\boldsymbol{\gamma}$ as $\hat{\boldsymbol{\gamma}}^{pcr} = \mathbf{W}_1\hat{\boldsymbol{\beta}}^{pcr}$.

Because $\mathbf{W}$ is an orthonormal matrix, the observation equations of the model fit in the regression step of PCR, as derived from (1), are

$$\mathbf{Y} = \mathbf{C}_1\boldsymbol{\beta}^{pcr} + \mathcal{E}^*, \tag{5}$$

where $\boldsymbol{\beta}^{pcr} = \mathbf{W}_1'\boldsymbol{\gamma}$ and $\mathcal{E}^* = \boldsymbol{X}\mathbf{W}_2\mathbf{W}_2'\boldsymbol{\gamma} + \mathcal{E}$. The PCR estimate of $\boldsymbol{\beta}^{pcr}$ is $\hat{\boldsymbol{\beta}}^{pcr} = \left(\mathbf{C}_1'\mathbf{C}_1\right)^{-1}\mathbf{C}_1'\mathbf{Y}$, which is biased when the first term of $\mathcal{E}^*$ is not $\mathbf{0}$. When $\boldsymbol{\gamma}$

is in the column space of $\mathbf{W}_1$, the first term is $\mathbf{0}$, the PCR estimator is the BLUE of $\boldsymbol{\beta}^{pcr}$, and $\hat{\boldsymbol{\gamma}}^{pcr} = \mathbf{W}_1\hat{\boldsymbol{\beta}}^{pcr}$ is the optimal estimator of $\boldsymbol{\gamma}$ subject to the restriction that $\boldsymbol{\gamma}$ is in the column space of $\mathbf{W}_1$. A second condition under which $\hat{\boldsymbol{\beta}}^{pcr}$ is BLUE is that the rank of $\boldsymbol{X}$ is k. In this case, the eigenvalues associated with the eigenvectors in $\mathbf{W}_2$ are all 0 and the first term in $\mathcal{E}^*$ vanishes. When neither of these conditions hold, the PCR estimator of $\boldsymbol{\gamma}$ is biased and loses the BLUE optimality properties. Intuitively speaking, bias and MSE increase as $\boldsymbol{\gamma}$ moves away from the column space of $\mathbf{W}_1$ and as one or more of the last $p-k$ eigenvalues increase from 0, respectively.

The predicted value of a new y, denoted by y_{n+1}, given a new $\boldsymbol{x}$, denoted by $\boldsymbol{x}_{n+1}$, is calculated as

$$\hat{y}_{n+1} = \boldsymbol{x}_{n+1}\hat{\boldsymbol{\gamma}}^{pcr} = \boldsymbol{x}_{n+1}\mathbf{W}_1\boldsymbol{\alpha}_1^{-1}\mathbf{W}_1'\mathbf{X}'\mathbf{Y}, \tag{6}$$

where $\boldsymbol{\alpha}_1$ is the diagonal matrix with the k eigenvalues associated with $\mathbf{W}_1$ on the diagonal.

Ideally, one would build the prediction equation from selected components that retain a highest percentage of variance possible in k dimensions, as do the first k PCs, but that also retain most of the information in the data about the association of $\boldsymbol{x}$ with y. PCR fails the latter, because it selects components of $\boldsymbol{x}$ ignoring their correlations with y. That is, PCR employs an unsupervised selection of components. This is the major drawback of PCR (Hadi and Ling 1998).

It should be noted that there are versions of PCR that involve either supervised selection of variables before the PCA step of PCR (Bair et al. 2006) or supervised cross validation determination of the number of components to keep in the regression step (James et al. 2013). The former mitigates the major drawback of PCR but does not eliminate it, unless the selection of components is also supervised. The later eliminates the concern but may result in an over specification of the PCR model. For example, suppose that a low variance component (say the 5th) is highly predictive. Then cross-validation component selection would include component 5, along with components 1-4 in the final model which may not be predictive. Inclusion of non-predictive components would introduce inefficiencies. So, the elimination of the major drawback of unsupervised PCR may help, but another drawback (inefficiency) replaces it.

PLS Regression

PLS regression was introduced by Wold (1966) as a solution to the major limitation of PCR mentioned above. The intention of PLSR is to form orthogonal component observation vectors ($\mathbf{C}_r$) that capture most of the information in the predictor variable data, $\mathbf{X}$, that is useful in predicting values of y (Frank and Friedman 1993; Garthwaite 1994; Helland 2010; Wold 1966; Wold et al. 2001). Component scores are determined sequentially through an iterative process that involves stepwise forward selection of components that represent factors in a k-dimensional factor model, $1 \leq k < rank(\mathbf{X})$, for $(y,\boldsymbol{x})'$.

Starting with standardized columns of $\mathbf{X}$ and $\mathbf{Y}$, given k, Wold et al. (1983, 2001) summarized his PLS1 algorithm for univariate y as a stepwise estimation of parameters in the following multivariate regression model data equations:

$$\begin{aligned}\mathbf{Y} &= \mathbf{C}_1\beta_y^{(1)} + \cdots + \mathbf{C}_k\beta_y^{(k)} + \mathbf{E}_y^{(k+1)} \\ \mathbf{X} &= \mathbf{C}_1\boldsymbol{\beta}_x^{(1)'} + \cdots + \mathbf{C}_k\boldsymbol{\beta}_x^{(k)'} + \mathbf{E}_x^{(k+1)},\end{aligned} \tag{7}$$

where $\mathbf{C}_r$, $r = 1, 2, \ldots, k$, are normalized component score vectors to be determined, $\mathbf{E}_y^{(k+1)}$ is a $n \times 1$ vector of residuals, and $\mathbf{E}_x^{(k+1)}$ is a $n \times p$ matrix of residuals. These equations are sample analogs of the bilinear factor model with values of latent factors ($f_r, r = 1, 2, \ldots, k$) replaced by corresponding component scores. The $\boldsymbol{\beta}$ coefficients in (7) are the loadings in the bilinear factor model. We wish to estimate both factor scores and loadings. which is accomplished by the following PLS1 algorithm involving 4 steps in each of k iterations, indexed by r, starting with $\mathbf{E}_y^{(0)} = \mathbf{Y}$ and $\mathbf{E}_x^{(0)} = \mathbf{X}$:

Step 1: Find the vector $\boldsymbol{w}_r$ that maximizes the squared correlation of $\mathbf{C}_r = \mathbf{E}_x^{(r-1)}\boldsymbol{w}$ with $\mathbf{E}_y^{(r-1)}$ over all unit length $\boldsymbol{w}$ vectors of dimension p that are orthogonal to the $\boldsymbol{w}_t$, $t = 1, 2, \ldots, r-1$, from previous iterations. This $\boldsymbol{w}_r$ is described by Frank and Friedman (1993) as

$$\boldsymbol{w}_r = \underset{\substack{\boldsymbol{w}'\boldsymbol{w} = 1 \\ \boldsymbol{w}'\hat{\mathbf{R}}_{xx}^{(r-1)}\boldsymbol{w} = 1 \\ \boldsymbol{w}'\hat{\mathbf{R}}_{xx}^{(r-1)}\boldsymbol{w}_t = 0}}{\text{argmax}} \left(\boldsymbol{w}'\hat{\mathbf{R}}_{xy}^{(r-1)}\hat{\mathbf{R}}_{xy}^{(r-1)'}\boldsymbol{w}\right), \tag{8}$$

where $t = 1, 2, \ldots, r-1$, and $\hat{\mathbf{R}}_{xx}^{(r-1)} = \mathbf{E}_x^{(r-1)'}\mathbf{E}_x^{(r-1)}/n - 1$ and $\hat{\mathbf{R}}_{xy}^{(r-1)} = \mathbf{E}_x^{(r-1)'}\mathbf{E}_y^{(r-1)}/n - 1$ are the sample partial covariance matrices of $\mathbf{X}$ and of $\mathbf{X}$ with $\mathbf{Y}$, respectively, controlling for $c_1, c_2, \ldots, c_{r-2}$, and c_{r-1}. For computations, note that $\boldsymbol{w}_r$ is the eigenvector of $\hat{\mathbf{R}}_{xy}^{(r-1)}\hat{\mathbf{R}}_{xy}^{(r-1)'}$ corresponding to the largest eigenvalue.

Step 2: Calculate $\mathbf{C}_r = \mathbf{E}_x^{(r-1)}\boldsymbol{w}_r$, the rth component' scores.

Step 3: Regress $\mathbf{E}_x^{(r-1)}$ on $\mathbf{C}_r$ and $\mathbf{E}_y^{(r-1)}$ on $\mathbf{C}_r$ to get estimates of the loadings of $\boldsymbol{x}$ on $\mathbf{C}_r$ (i.e., $\boldsymbol{\beta}_x^{(r)'}$) and of y on $\mathbf{C}_r$ (i.e. $\beta_y^{(r)}$), respectively. The estimates are calculated as $\hat{\boldsymbol{\beta}}_x^{(r)'} = \boldsymbol{w}_r'\mathbf{E}_x^{(r-1)'}\mathbf{E}_x^{(r-1)}$ and $\hat{\beta}_y^{(r)} = \boldsymbol{w}_r'\mathbf{E}_x^{(r-1)'}\mathbf{E}_y^{(r-1)}$, respectively.

Step 4: Calculate the residuals from the regression of $\mathbf{X}$ on $\mathbf{C}^{(r)}$ and the residuals from the regression of $\mathbf{Y}$ on $\mathbf{C}^{(r)}$, where $\mathbf{C}^{(r)} = (\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_r)$. This gives $\mathbf{E}_x^{(r)} = (\mathbf{X} - \mathbf{C}^{(r)}\hat{\boldsymbol{\beta}}_x^{(r)})$ and $\mathbf{E}_y^{(r)} = (\mathbf{Y} - \mathbf{C}^{(r)}\hat{\boldsymbol{\beta}}_y^{(r)})$, where $\hat{\boldsymbol{\beta}}_x^{(r)}$ is the row by row concatenation of the loading estimates $\hat{\boldsymbol{\beta}}_x^{(t)'}$, $t = 1, 2, \ldots, r$, and $\hat{\boldsymbol{\beta}}_y^{(r)} = (\hat{\beta}_y^{(1)}, \hat{\beta}_y^{(2)}, \ldots, \hat{\beta}_y^{(r)})'$ calculated in Step 3.

Steps 1-4 are repeated k times to build the k dimensional factor model analog (7). At each iteration of steps, a new term is added to Model (7). Thus, PLSR involves stepwise forward model building. The prediction equation for $\mathbf{y}_{n+1}$ as a

function of an observed $\boldsymbol{x}_{n+1}$ is $\hat{\boldsymbol{y}}_{n+1} = \mathbf{C}_{n+1}^{(k)}\hat{\boldsymbol{\beta}}_y^{(k)} = \boldsymbol{x}_{n+1}\hat{\boldsymbol{\gamma}}^{pls}$. It has been shown that $\hat{\boldsymbol{\gamma}}^{pls}$ can be written as

$$\hat{\boldsymbol{\gamma}}^{pls} = \mathbf{W}^{(k)}(\hat{\boldsymbol{\beta}}_x^{(k)}\mathbf{W}^{(k)})^{-1}\hat{\boldsymbol{\beta}}_y^{(k)}, \tag{9}$$

where $\mathbf{W}^{(k)}$ is the concatenation of the vectors $\mathbf{w}_t$, $t = 1, 2, \ldots, k$ (see Wold et al. 2009).

In practice, the dimension of the actual underlying factor model is not known and one must rely on the data to choose k. In Sect. 3.7, we discuss how to estimate k using a K-fold cross-validation (CV(K)) approach. See, also, Wold et al. (2001).

Because PLSR builds a prediction equation through forward selection of components, we hypothesize that it suffers from the same potentially severe over-fitting as forward selection regression when $p >> n$ (Efron et al. 2004). This criticism applies to forward selection model building in any context, be it usual regression, PCR or PLSR.

The problem is illustrated intuitively as follows in the context of PLSR. Suppose that $n = 100$ and $p = 120$ and that the population correlation between y and all linear combinations of $\boldsymbol{x}$'s are zero. Some sample correlations calculated from the full calibration data set, nevertheless, will be large simply by chance. Suppose that CV(5) cross-validation is used to determine the number of components to include in the model. It is likely that some of the holdout samples of 20 observations and their corresponding retained samples of 80 are both representative of the full calibration sample. In these representative pairs the high correlation components are likely to be highly correlated in both the holdout and retained samples. They will, therefore, reduce the PRESS in the holdout sample and overall when added to the model. They will be added due to their chance correlation with y. The chance of such superfluous components entering the model increases as p increases relative to n. Since a new validation sample is likely to "misrepresent" the population by chance in different ways, the highly correlated (with y) components will be mostly different, and the model built from the calibration may result in poor predictions in the validation sample.

Factor Analysis Regression

Factor Analysis

Given a sample of n observations from a vector of observable random variables$\boldsymbol{z} = (z_1, z_2, \ldots, z_{p+1})'$, the common factor model (CFM) for $\boldsymbol{z}$ is written in matrix form as:

$$\boldsymbol{z} = \boldsymbol{\lambda}'\boldsymbol{f} + \mathbf{U}, \tag{10}$$

where $\boldsymbol{f}$ is the vector $(f_1, f_2, \ldots, f_k)'$ of latent variables called common factors, $\mathbf{U}$ is a $(p+1) \times 1$ vector of unique factors (i.e., U_j is related to z_j but independent of the other observable variables), and $\boldsymbol{\lambda}$ is the $k \times (p+1)$ matrix of coefficient parameters to be estimated. The usual model assumptions are:

(1) $\mathrm{E}(\boldsymbol{f}) = \mathbf{0}_k$, $\mathrm{E}(\boldsymbol{U})=\mathbf{0}_{p+1}$
(2) $\mathrm{Cov}(\boldsymbol{f}) = \mathbf{I}_k$
(3) $\mathrm{Cov}(\boldsymbol{U}) = \boldsymbol{\Psi}$ a $(p+1) \times (p+1)$ diagonal matrix and
(4) $\boldsymbol{f}$ and $\boldsymbol{U}$ are independent.

The assumption of zero means for all variables in the model is made without loss of generality.

In general, the parameters in $\boldsymbol{\lambda}$ are called factor pattern coefficients. When the observable variables are standardized, however, special interpretations are possible (e.g., the (i, j)th entry in $\boldsymbol{\lambda}$ is the population correlation between z_i and f_j) and they are given the name of *factor loadings*.

Note that there is an indeterminacy in Model (10). That is, there are multiple sets of parameters, $\boldsymbol{\lambda}$, and underlying factor vectors, $\boldsymbol{f}$, that produce the same $\boldsymbol{\lambda}'\boldsymbol{f}$. $\boldsymbol{\lambda}$ and $\boldsymbol{f}$ can only be identified up to an orthonormal rotation. The indeterminency can be removed by applying a rotation of particular interest. For example, one may want to rotate toward a simple model that leads to easily named factors.

Bilinear Common Factor Model

In the context of regression analysis, we define $\boldsymbol{z} = (y, \boldsymbol{x})'$. Then, the CFM in equation (10) can be partitioned as

$$\begin{aligned} y &= \boldsymbol{\lambda}_y'\boldsymbol{f} + \mathrm{U}_y \\ \boldsymbol{x}' &= \boldsymbol{\lambda}_x'\boldsymbol{f} + \mathbf{U}_x. \end{aligned} \tag{11}$$

A partitioned form of the CFM in (10) is called a Bilinear Common Factor Model (BCFM).

Model (11) is the basis for FAR estimation of $\boldsymbol{\gamma}$ in Model (1) and for building a high-dimensional prediction equation.

Covariance Matrix

The covariance matrix of $\boldsymbol{z}$ under model (10) with standardized $\boldsymbol{z}$ and Assumptions 1-4 above is the correlation matrix

$$\mathbf{R}_z = \boldsymbol{\lambda}'\boldsymbol{\lambda} + \boldsymbol{\Psi}. \tag{12}$$

The diagonal elements of $\boldsymbol{\lambda}'\boldsymbol{\lambda}$ and $\boldsymbol{\Psi}$ are called communalities and unique factor variances, respectively. The ith communality, h_i^2 say, is the proportion of the variance of z_i that is explained by its relationship with the latent factors $\boldsymbol{f}$ and the ith unique variance, ψ_i say, is the proportion left unexplained.

Now partition $\boldsymbol{\lambda}$ and $\boldsymbol{\Psi}$ in (12) such that $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_y, \boldsymbol{\lambda}_x)$ and $\boldsymbol{\Psi} = diagonal(\psi_y, \boldsymbol{\Psi}_x)$ where $\boldsymbol{\Psi}_x$ is $p \times p$ diagonal matrix. Then, for $\boldsymbol{z} = (\mathrm{y}, \boldsymbol{x})'$, we have

$$\mathbf{R}_z = \begin{pmatrix} \lambda_y'\lambda_y + \psi_y & \lambda_y'\lambda_x \\ \lambda_x'\lambda_y & \lambda_x'\lambda_x + \mathbf{\Psi}_x \end{pmatrix} \tag{13}$$

We see from (3) and (13) that

$$\boldsymbol{\gamma} = (\lambda_x'\lambda_x + \mathbf{\Psi}_x)^{-1}\lambda_x'\lambda_y, \tag{14}$$

if the bilinear factor model holds. An estimate obtained by substituting FA estimates of the parameters in (11) into (14) is a FAR estimate. FAR estimates will vary depending on the factor extraction method used. We chose PC factor extraction, because it is amenable to a computationally efficient modification that is helpful (sometimes necessary) in high-dimensional settings.

Principal Component Factor Extraction

Given a calibration sample, let $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$ denote the $n \times (p+1)$ matrix of standardized observed values of the random vector $\boldsymbol{z}$. Then, by the spectral decomposition theorem,

$$\hat{\mathbf{R}}_z = \left(\frac{1}{n-1}\right)\mathbf{Z}'\mathbf{Z} = \mathbf{W}\boldsymbol{\alpha}\mathbf{W}', \tag{15}$$

where the orthogonal columns of $\mathbf{W}$ and diagonal matrix $\boldsymbol{\alpha}$ are the eigenvectors and associated eigenvalues of $\hat{\mathbf{R}}_z$, respectively. First, choose the number of factors, $k < rank(\mathbf{X})$. (We used CV for this in Sect. 3.7) Then, partition $\mathbf{W}$ into matrices of order $(p+1) \times k$ and $(p+1) \times (p+1-k)$ and $\boldsymbol{\alpha}$ correspondingly; i.e.,

$$\begin{aligned} \mathbf{W} &= [\mathbf{W}_1 \,:\, \mathbf{W}_2] \\ \boldsymbol{\alpha} &= \text{Block diagonal}[\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2] \end{aligned} \tag{16}$$

The principal component factor extraction method yields

$$\hat{\lambda} = \hat{\boldsymbol{\alpha}}_1^{\frac{1}{2}}\mathbf{W}_1', \quad \text{and} \quad \hat{\mathbf{F}} = \mathbf{Z}\mathbf{W}_1\hat{\boldsymbol{\alpha}}_1^{-\frac{1}{2}} \tag{17}$$

(See Johnson and Wichern 2007, page 387; and Rao (1996), Equation 3.5, which is standardized here) $\hat{\mathbf{F}}$ is the $n \times k$ matrix of standardized PC scores. Communality estimates are calculated as $\hat{h}_j^2 = \sum_{r=1}^{k} \hat{\lambda}_{rj}^2$ and unique varance estimates as $\hat{\psi}_j = 1 - \hat{h}_j^2$, $j = 1, 2, \ldots, p+1$. $\hat{h}_j^2$ is interpreted as the proportion of sample variance of z_j explained by the k common factors and ψ_j as the proportion unexplained.

Several comments are in order before leaving this section:

Comment 1. The PC estimates in (17) are optimal under the restricted version of Model (10), with restrictions $\psi_1 = \psi_2 = \cdots = \psi_{p+1}$, in the sense that $\hat{\lambda}'$ and $\hat{\boldsymbol{f}}_1, \hat{\boldsymbol{f}}_2, \ldots, \hat{\boldsymbol{f}}_n$, the rows of $\hat{\mathbf{F}}$ transposed, together are a solution that minimizes

$$\sum_{i=1}^{n}(z_i - \boldsymbol{\lambda}'\boldsymbol{f}_i)'(z_i - \boldsymbol{\lambda}'\boldsymbol{f}_i) \tag{18}$$

in that case. (See Equation 3.5 and surrounding discussion in Rao (1996)).

Comment 2. Although the PC estimates estimate "a" factor model, they do not estimate "the" factor model of interest (10).

Comment 3. In practice, PC extraction often estimates population parameters that are "close" to the parameters in Model (10) (Schneeweiss and Mathes 1995). Schneeweiss and Mathes (1995) derived the conditions under which this occurs. The conditions are likely to be satisfied when a large number of variables are included, all with something in common with at least some of the other variables, and the number of factors is correctly identified. (k must be large enough so that unique factor variances are small but small enough to exclude factors that explain little of the total variance in observed variables). So, one can reasonably expect PC extraction to provide nearly valid estimates of Model (10) in carefully designed high-dimensional studies with a good method for selecting the number of factors (e.g., the CV method).

Comment 4. The MLE method of factor extraction cannot be used in high-dimensional cases and principal axis factor analysis cannot be applied with the most commonly chosen option for prior communalities (i.e., SMC). While other choices of priors are workable, repeated computation of the eigenvalues and vectors of $(p+1)\times(p+1)$ matrices when p is very large may be burdensome. We can reduce the computational burden significantly with PC extraction, as shown in the next section.

Comment 5. As already mentioned, PCR employs only $\hat{\mathbf{R}}_{xx}$, ignoring $\hat{\mathbf{R}}_{xy}$. PLS solves that problem but utilizes a forward variable (i.e., component) selection process that is expected to over-fit the sample. We hypothesize that FAR shares neither of these deficiencies and, therefore, will outperform both PCR and PLSR in our Monte Carlo studies

High-Dimensional Factor Extraction

For very large values of $p >> n$, it is computationally expensive to obtain factor model parameter estimates, which requires computation of the eigenvalues and eigenvectors of the $(p+1)\times(p+1)$ matrix $\hat{\mathbf{R}}_z$. Instead, we achieve PC extraction from the eigenvalues and eigenvectors of the $n \times n$ Gram matrix, $\mathbf{ZZ}'$.

Let $\mathbf{G}_1$ denote the $n \times k$ matrix of the first k orthonormal eigenvectors of $\mathbf{ZZ}'$ and let $\boldsymbol{\vartheta}$ denote the diagonal matrix of associated ordered eigenvalues. That is, $\mathbf{ZZ}'\mathbf{G}_1 = \mathbf{G}_1\boldsymbol{\vartheta}_1$. It follows that $1/(n-1)\mathbf{Z}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{G}_1\boldsymbol{\vartheta}_1^{-\frac{1}{2}}\right) = 1/(n-1)\left(\mathbf{Z}'\mathbf{G}_1\boldsymbol{\vartheta}_1^{-\frac{1}{2}}\right)\boldsymbol{\vartheta}_1$ and, thus, that $\mathbf{V}_1^* = \left(\mathbf{Z}'\mathbf{G}_1\boldsymbol{\vartheta}_1^{-\frac{1}{2}}\right)$ is a $p \times k$ matrix of the first k eigenvectors of $\hat{\mathbf{R}}_z$ with associated eigenvalues $\frac{\boldsymbol{\vartheta}_1}{n-1}$. Furthermore, $\mathbf{V}_1^{*\prime}\mathbf{V}_1^*=\mathbf{I}$.

Therefore, by the uniqueness of orthonormal eigenvectors and associated eigenvalues, we have

$$\begin{aligned} \mathbf{W}_1 &= \mathbf{Z}'\mathbf{G}_1\boldsymbol{\vartheta}_1^{-\frac{1}{2}} \\ \boldsymbol{\alpha}_1 &= \frac{\boldsymbol{\vartheta}_1}{n-1}, \end{aligned} \tag{19}$$

By substituting (19) into (17) we obtain the more computationally efficient estimates

$$\hat{\boldsymbol{\lambda}} = \frac{\mathbf{G}_1'\mathbf{Z}}{\sqrt{n-1}} \tag{20}$$

$$\hat{\mathbf{F}} = \mathbf{Z}\mathbf{Z}'\mathbf{G}_1\boldsymbol{\vartheta}_1^{-1}\sqrt{n-1}. \tag{21}$$

Estimates of $\boldsymbol{\Psi}$ and h_i^2 follow, as calculated above.

Estimation of the High-Dimensional Regression Model and Prediction

Our primary goals are to estimate $\boldsymbol{\gamma}$ in (1) and to build a prediction equation in high-dimensional settings, assuming the low- dimensional factor model (10), or equivalently (11), holds. Factor analyzing $\hat{\mathbf{R}}_z = (\frac{1}{n-1})\mathbf{Z}'\mathbf{Z}$, using computationally efficient high-dimensional factor extraction, yields the estimates in (20) and (21). Substituting these and the associated $\hat{\boldsymbol{\Psi}}_x$ into (14), we obtain the FAR estimate of $\boldsymbol{\gamma}$,

$$\hat{\boldsymbol{\gamma}}^{FA} = (\hat{\boldsymbol{\lambda}}_x'\hat{\boldsymbol{\lambda}}_x + \hat{\boldsymbol{\Psi}}_x)^{-1}\hat{\boldsymbol{\lambda}}_x'\hat{\boldsymbol{\lambda}}_y. \tag{22}$$

If $\hat{\boldsymbol{\Psi}}_x$ is nonsingular (i.e., $\hat{\boldsymbol{\psi}}_j > 0$, $\forall j = 1,2,...,p$), then we can write $(\hat{\boldsymbol{\lambda}}_x'\hat{\boldsymbol{\lambda}}_x + \hat{\boldsymbol{\Psi}}_x)^{-1}$ as

$$\hat{\boldsymbol{\Psi}}_x^{-1} - \hat{\boldsymbol{\Psi}}_x^{-1}\hat{\boldsymbol{\lambda}}_x'(\mathbf{I}_k + \hat{\boldsymbol{\lambda}}_x\hat{\boldsymbol{\Psi}}_x^{-1}\hat{\boldsymbol{\lambda}}_x')^{-1}\hat{\boldsymbol{\lambda}}_x\hat{\boldsymbol{\Psi}}_x^{-1}$$

(See Theorem 18.2.8 in Harville (1998)). Hence, the inverse in (22) can be computed by inverting only a $p \times p$ diagonal matrix and a $k \times k$ matrix. The FAR estimator of $\boldsymbol{\gamma}$ in computationally efficient form is

$$\hat{\boldsymbol{\gamma}}^{FA} = (\hat{\boldsymbol{\Psi}}_x^{-1} - \hat{\boldsymbol{\Psi}}_x^{-1}\hat{\boldsymbol{\lambda}}_x'(\mathbf{I}_k + \hat{\boldsymbol{\lambda}}_x\hat{\boldsymbol{\Psi}}_x^{-1}\hat{\boldsymbol{\lambda}}_x')^{-1}\hat{\boldsymbol{\lambda}}_x\hat{\boldsymbol{\Psi}}_x^{-1})\hat{\boldsymbol{\lambda}}_x'\hat{\boldsymbol{\lambda}}_y, \tag{23}$$

which is used to form a prediction equation for new values of the response variable, y, as a function of the associated new value of $\boldsymbol{x}$. That is,

$$\hat{\mathbf{y}}_{(n+1)}^{FA} = \boldsymbol{x}_{(n+1)}\hat{\boldsymbol{\gamma}}^{FA}. \tag{24}$$

Choosing the Number of Factors

All regression methods described above were applied for a given k. Choosing k is an important but difficult problem. We prefer the K-fold cross-validation method (CV(K)) described by Wold (1978). We present details of its application here in the context of FAR, with K chosen to be 5.

We applied CV(5) for each possible value of k up to q,

$$q = \min\left(\mathrm{k}_{kr}, \frac{n}{5}\right), \tag{25}$$

where k_{kr} is the number of factors chosen by the Kaiser criterion (i.e., retain factors with eigenvalues greater than 1.0). Our rationale for limiting k as in (25) was based on preliminary studies that showed the Kaiser criterion tended to over-estimate the true k and the rule of thumb that we need at least 5 observations for each factor.

We split the calibration sample into five groups, as dictated by our choice of the CV(5) method. One group was omitted and the dataset of the other four groups combined was analyzed to compute $\hat{\boldsymbol{\gamma}}^{FA}$ in Eq. (22), or in more computationally efficient form in (23). This was repeated with each of the other groups omitted and predicted values of y were calculated from (24) in each of the held out validation samples. The 5 prediction error sum of squares (PRESS) were then pooled to obtain an overall PRESS statistic. This was done for each $r = 1, 2, \ldots, q$ and k was chosen to be the r with the minimum MSEP=PRESS/$(n - r - 1)$.

Monte Carlo Simulation

Simulation studies were performed to investigate the relative performance of FAR, PCR, and PLS regression/prediction. Calibration and validation samples were generated. For each calibration sample, the above methods to retain the number of factors, estimate the coefficients, and build a prediction equation from PCR, FAR and PLSR were applied. The results were applied to predict values of y in each validation sample.

Data Structures

Two Monte Carlo (MC) studies were performed (MC study-1 and MC study-2) with study parameters taken to be the estimated parameters from two real datasets. In each of these studies, data were generated under three underlying correlation structures:

- Arbitrary correlation structure—The $\mathbf{R}_z$ for each study was taken to be that estimated from the real sample.
- Factor model structure—Factor analysis was applied to each study's data to estimate the loading matrix $\boldsymbol{\lambda}$ and $\boldsymbol{\Psi}$ in (12). Then data were generated from a distri-

bution with covariance matrix $\hat{\mathbf{R}}_z = \hat{\lambda}'\hat{\lambda} + \hat{\mathbf{\Psi}}$ with parameters estimated from the data.

- Independence structure—y was generated independent of $\boldsymbol{x}$; i.e., $\mathbf{R}_z = \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{R}_{xx} \end{pmatrix}$, where $\mathbf{R}_{xx}$ is an arbitrary symmetric positive definite by $p \times p$ sample correlation matrix from the real data.

Design of Monte Carlo Studies

MC Study-1

The parameters for this study were estimated from kidney disease registry data where a complete set of 26 lab scores were observed at least once on each of 57 patients. For each lab test j, $j = 1, 2, \ldots, 26$, the data record of each patient, x_{ij}^*, $i = 1, 2, \ldots, 57$, was transformed to scores, x_{ij}, measuring the patient's deviation from normal, based on a range of values in a normal population. Then, standardized deviation scores were taken as the predictor variables, $\boldsymbol{x}$, and standardized average monthly cost of medical services over the previous 13 months was the response variable, y. R-square from the regression analysis of $\mathbf{Y}$ on $\mathbf{X}$ was 0.60 with p value of 0.06 (and p value = 0.001 for the model derived from backward selection). As discussed above, $\mathbf{R}_z$ were specified to satisfy the restrictions of each underlying correlation structure. For the factor model structure, the number of factors was chosen ($k = 8$), using Kaiser criterion. Using these parameters, 500 MC calibration samples of each size $n = 10$, 15, 20, 25, and 30 were generated from a MVN($\mathbf{0}$, $\mathbf{R}_z$) distribution. For each calibration sample, a validation sample of size 400 was independently generated from the same distribution.

MC Study-2

In this MC study, we generated 500 MC calibration samples of size n = 10, 20, 30, 40, 50, 60 and 70 and a 400 observation validation sample for each calibration sample for each correlation structure. All observations were drawn from a MVN($\mathbf{0}$, $\mathbf{R}_z$) distribution. The correlation matrix parameters were estimated from microarray gene expression observations from 65 probe-set measures on 120 twelve-week-old male F2 rats (Stone et al. 2006) to identify genes whose mutations are associated with Bardet-Biedl syndrome (BBS, one of the rare genetic disorders collectively called ciliopathies). For this MC sampling, we added 54 randomly selected probesets to the 10 probesets for known BBS genes in Stone et al. (2006) and chose a newly discovered ciliopathy gene TOPORS (probe id X1392610_at) as a response variable to see whether TOPORS is associated with the known BBS genes. The R-square from the standardized regression analysis of $\mathbf{Y}$ on $\mathbf{X}$ was 0.9012 with p value < 0.0001. Using the Kaiser criterion, $k = 11$ factors were selected for the factor model structure. The factor analysis results were used to form a structured correlation matrix in the form of (13). For the independence case the off-diagonal

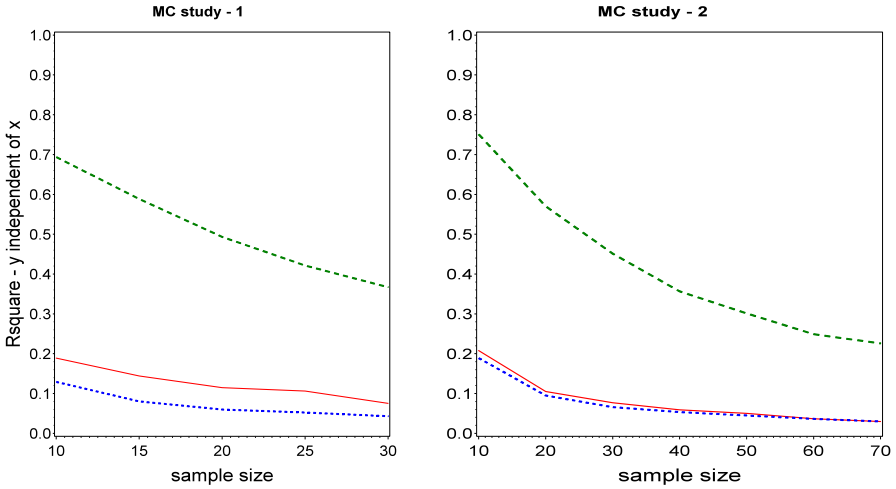


Fig. 1 Estimated R-square averaged across MC calibration samples when $\boldsymbol{x}$ and y are independent. FA regression(solid red curve), PLS regression(dashed green curve), and PCR (dotted blue curve)

elements of the first row and first column of the sample correlation matrix were set to zeros.

Criteria of Evaluation

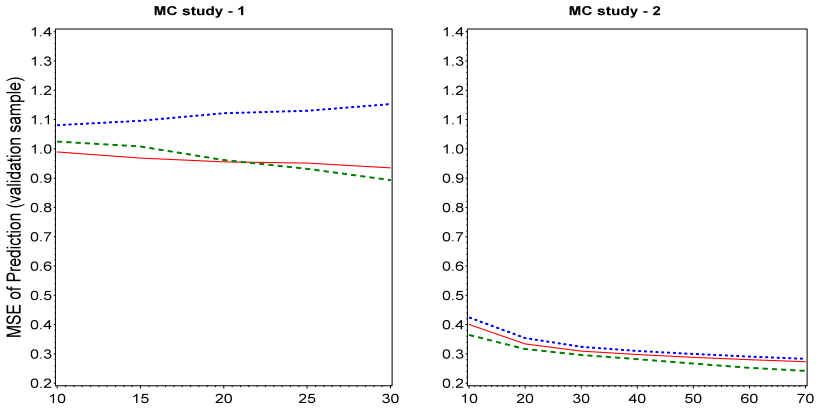
Initially, the CV(5) methods were used separately to choose the number of factors, k, in the FAR, the number of components in the PCR, and the number of scores in the PLSR from the calibration datasets.

After selecting k, using CV(5), each calibration sample was reanalyzed by FAR, PCR and PLSR, respectively and a prediction equation obtained for each method. The y value in each observation in each corresponding validation sample was predicted by each method. The performance metric used to compare methods was the MSE of prediction averaged across the 500 MC calibration samples by sample size, n,

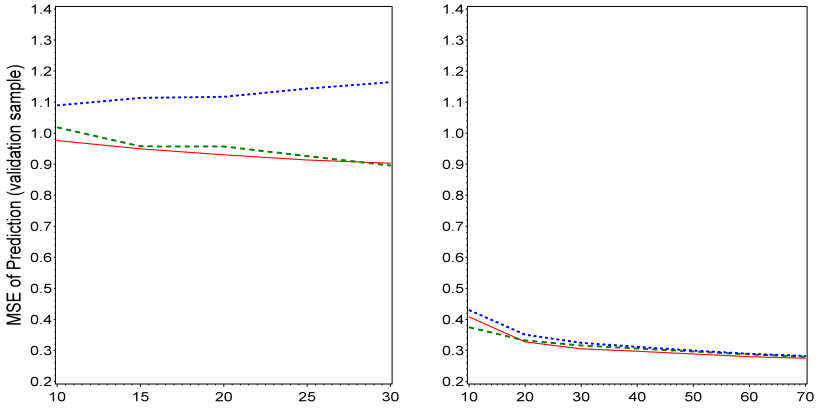
$$\mathrm{MSE}(\hat{Y}_v) = \frac{1}{500}\sum_{m=1}^{500}\frac{(\mathbf{Y}_{v,m}-\hat{\mathbf{Y}}_{v,m})'(\mathbf{Y}_{v,m}-\hat{\mathbf{Y}}_{v,m})}{400-k-1}. \tag{26}$$

In samples generated from the independence structure (i.e. $\mathbf{R}_{xy}=\mathbf{0}$), we calculated the average R-square from the regression of y on $\boldsymbol{x}$ across calibration samples, i.e.,

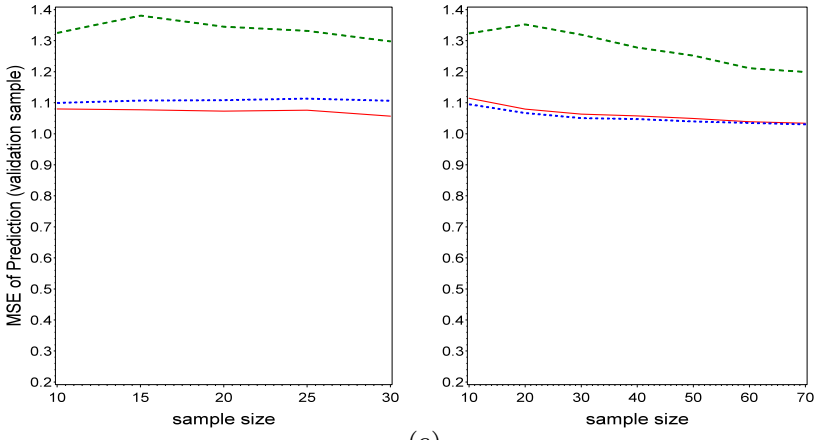
Fig. 2 MSE of Prediction averaged across the 500 MC validation samples. **a** Arbitrary correlation structure; **b** factor model structure; **c** $\boldsymbol{x}$ and y are independent. FA regression (solid curve), PLS regression (dashed curve), and PCR (dotted curve) ▶


MC study - 1
MC study - 2
MSE of Prediction (validation sample)

(a)


MSE of Prediction (validation sample)

(b)


MSE of Prediction (validation sample)
sample size
sample size

(c)

$$R^2 = \frac{1}{500}\sum_{m=1}^{500} \text{corr}^2(\mathbf{Y}_m, \hat{\mathbf{Y}}_m) \tag{27}$$

to compare FAR to PCR and PLSR. This would reveal over-fitting if it occurred.

Results

Results of our comparison of the three methods given independence of y and $\boldsymbol{x}$ are shown in Figure 1 below.

One would expect the curves shown to be close to zero in this setting. The curve for PLSR, however, are much greater than zero for all calibration sample sizes in both MC studies. This, we believe, is the result of PLSR over-fitting due to being a forward selection methodology. PCR and FAR show a much less severe bias in R-square estimates from the calibration sample. In essence, PLSR can lead us to believe the existence of a relationship between y and $\boldsymbol{x}$ when there is none.

The MSE of prediction averaged over the 500 MC validation samples under arbitrary, factor model, and independence structures is summarized by study in Figure 2 for the three underlying structures.

The over-fit PLSR prediction equation leads to poor predictions and high average prediction MSE, under the independence structure, as seen in Figure 2c. Under factor model and arbitrary correlation structures, we observed comparable performance by FAR and PLSR in both studies. PCR was not a viable competitor in Study-1 but performed as well as FAR and PLSR in Study-2. The poor performance by PCR in Study-1 was expected, potentially, and was likely due to the fact that it employs an unsupervised selection of components. The PCR results in Study-2 are attributed to the characteristics of the data, which happened to have expression profiles of some of the covariates (i.e. the BBS gene expressions) that are highly variable as well as highly correlated with the expression profile of the response variable, TOPORS.

Our MC study results can be roughly summarized as follows: FAR never performed notably worse than PLSR and performed much better when y and $\boldsymbol{x}$ were independent. Similarly, FAR never performed notably worse than PCR and performed better, overall, when y and $\boldsymbol{x}$ were related.

Conclusion

FAR outperformed both PLSR and PCR in our MC studies. It did so in a predictable fashion given the weaknesses of PLSR and PCR, known a priori. Because PLSR performs a forward selection of components, it was expected to over-fit the data by including components that are only predictive in the calibration sample. This is a problem that is most obviously manifest when there is no relationship between y and $\boldsymbol{x}$. Because PCR selects components in an unsupervised fashion, it was expected that it might leave low variance components that were predictive

of y out of the prediction equation and therefore suffer bias in some studies. Such was the case, apparently, in MC Study-2.

Because of these complementary deficiencies, neither PLSR nor PCR should be uniformly preferred over the other. PLSR performed better when there was a relationship between y and $\boldsymbol{x}$, but PCR performed better when there was not.

Ridge regression (RR) (Hoerl and Kennard 1970) is not considered in this article. It is left to future research to compare Lasso, FAR, and RR when underlying conditions are such that neither Lasso nor FAR are clearly preferred. RR is not expected to out-perform Lasso when the model is sparse nor to out-perform FAR when there is an underlying factor model correlation structure, conditions that often can be reasonably judged by investigators.

Supervised PCR methods were also not included in our Monte Carlo comparisons. The partially supervised method (Bair et al. 2006) without supervised selection of components, suffers potential bias due to non-inclusion of low variance but predictive components. When the selection of components is supervised as in James et al. (2013), there is potential for the inclusion of extraneous components in the final model and, hence, inefficient prediction. Theoretically, therefore, FAR is expected to out-perform both unsupervised and supervised versions of PCR.

References

Bair, E., T. Hastie, D. Paul, and R. Tibshirani. 2006. Prediction by supervised principal components. *Journal of the American Statistical Association* 101 (473): 119–137.

Belloni, A., V. Chernozhukov, and C. Hansen. 2014. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives* 28 (2): 29–50.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. *The Annals of Statistics* 32: 407–499.

Fan, J., Y. Liao, and J. Yao. 2015. Power enhancement in high-dimensional cross-sectional tests. *Econometrica* 83 (4): 1497–1541.

Frank, I.E., and J.H. Friedman. 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35 (2): 109–135.

Friedman, J., R. Tibshirani, and T. Hastie. 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Berlin: Springer.

Garthwaite, P.H. 1994. An interpretation of partial least squares. *Journal of the American Statistical Association* 89 (425): 122–127.

Hadi, A.S., and R.F. Ling. 1998. Some cautionary notes on the use of principal components regression. *The American Statistician* 52 (1): 15–19.

Harville, D.A. 1998. *Matrix Algebra from a Statistician's Perspective*. New York: Taylor & Francis.

Helland, I.S. 2010. *Steps Towards a Unified Basis for Scientific Models and Methods*. Singapore: World Scientific Pub. Co.

Hoerl, A.E., and R.W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12 (1): 55–67.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An Introduction to Statistical Learning vol. 112.

Johnson, R.A., and D.W. Wichern. 2007. *Applied Multivariate Statistical Analysis*. Hoboken: Pearson Prentice Hall.

Jolliffe, I. 2005. Principal component analysis. *Encyclopedia of Statistics in Behavioral Science* 20: 20.

Kalina, J. 2017. High-dimensional data in economics and their (robust) analysis. *Serbian Journal of Management* 12 (1): 157–169.

Rao, C.R. 1996. Principal component and factor analyses. *Handbook of Statistics* 14: 489–505.

Schneeweiss, H., and H. Mathes. 1995. Factor analysis and principal components. *Journal of Multivariate Analysis* 55 (1): 105–124.
Scott, J.T., Jr. 1966. Factor analysis and regression. *Econometrica: Journal of the Econometric Society* 20: 552–562.
Stone, E., A. Chiang, T. Scheetz, K. Kim, R. Swiderski, D. Nishimura, L. Affatigato, J. Huang, T. Casavant, and V. Sheffield. 2006. Analysis of correlated gene expression in a large cohort of rats assists the discovery of two new genes involved in bardet biedl syndrome (bbs). *Investigative Ophthalmology & Visual Science* 47 (13): 5919.
Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 58 (1): 267–288.
Wold, H. 1966. Estimation of principal components and related models by iterative least squares. *Multivariate Analysis* 20: 391–420.
Wold, S. 1978. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 20 (4): 397–405.
Wold, S., M. Høy, H. Martens, J. Trygg, F. Westad, J. MacGregor, and B.M. Wise. 2009. The pls model space revisited. *Journal of Chemometrics: A Journal of the Chemometrics Society* 23 (2): 67–68.
Wold, S., H. Martens, and H. Wold. 1983. The multivariate calibration-problem in chemistry solved by the pls method. *Lecture Notes in Mathematics* 973: 286–293.
Wold, S., M. Sjöström, and L. Eriksson. 2001. Pls-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58 (2): 109–130.
Zhang, M.H., Q.S. Xu, and D.L. Massart. 2003. Robust principal components regression based on principal sensitivity vectors. *Chemometrics and Intelligent Laboratory Systems* 67 (2): 175–185.
Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67 (2): 301–320.

Authors and Affiliations

Randy Carter[1] · Netsanet Michael[2]

Netsanet Michael
netsanet.t.michael@boeing.com

[1] Department of Biostatistics, State University of New York at Buffalo, 725 Kimball Tower, Buffalo, NY 14214, USA

[2] Boeing Commercial Airplanes, The Boeing Company, Mail Code 9U2-05, P.O. Box 3707, Seattle, WA 98124, USA