Mathematical Statistics
Stockholm University

A two-parametric class of predictors in multivariate regression

Anders Björkström
Rolf Sundberg

Research Report 2007:8

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2007:8**,
http://www.math.su.se/matstat

A two-parametric class of predictors in multivariate regression

Anders Björkström *
Rolf Sundberg *

April 2007

Abstract

We demonstrate that a number of well-established multivariate regression methods for prediction are related, in that they are special cases of basically one general procedure. We try a more general method based on this procedure, with two metaparameters. In a simulation study, based on a latent structure model, we compare this method to ridge regression, multivariate PLSR and repeated univariate PLSR. For most types of data sets studied, all methods do approximately equally well. There are some cases where RR and LSRR yield larger errors than the other methods, and we conclude that one-factor methods are not adequate for situations where more than one latent variable are needed to describe the data. Among those based on latent variables, none of the methods tried is superior to the others in any obvious way.

Key words: Joint continuum regression, multivariate prediction, multivariate regression, PCR, PLSR, reduced rank regression, ridge regression, SIMPLS, total least squares

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: bjorks@math.su.se.

1 Introduction

1.1 Connections between regression methods

In regression, a substantial number of alternatives to the ordinary least squares method (OLSR) have been developed with the aim of reducing the variance that stems from near-collinearities among the explanatory variables. Well-known methods, at least for univariate response variables, are principal components regression (PCR), partial least squares regression (PLSR), ridge regression (RR), and continuum regression (CR); see Stone & Brooks (1990), Frank & Friedman (1993), Brown (1993), Sundberg (1999), Kalivas (1999) or de Jong et al (2001) for reviews. Naturally, different methods will yield different answers to a specific problem, and in order to avoid unnecessary confusion, it is desirable to understand why results differ, and to be able to explain why one alternative is likely to be preferable to another one, in a given situation. In addition to simulation studies, a number of theoretical results have been deduced that compare methods to each other. We know, for example, that PLSR is a shrinkage regressor (de Jong, 1995; Goutis, 1996), and also that PLSR yields larger correlation than PCR with the same number of factors (de Jong, 1993a). In this context we may also mention the early work by James & Stein (1961) concerning the inadmissibility of OLSR in some situations, and the proof by Hoerl & Kennard (1970) that ridge regression with a ridge parameter small enough will have smaller mean squared error than OLSR. Despite these and other results, many questions in the field of method comparison remain open, particularly with regard to the case of multivariate response variables.

One way to shed light on the relation between methods is to construct an indexed class of regressors, $\{B_\alpha\,;\,\alpha\in\mathcal{A}\}$, where each B_α denotes a regression procedure (cf. Eq. 1 below). A well-known example is Stone & Brook's (1990) continuum regression, CR, where B_α stands for a regressor defined as

$$B_\alpha(X,y)\propto \arg\max_{|c|=1}\{|c'X'y|\,|Xc|^{\alpha/(\alpha-1)-1}\}.$$

The index set $\mathcal{A}$ is the interval $[0,1]$. The set of regressors includes OLSR ($\alpha=0$), PLSR ($\alpha=1/2$), and PCR ($\alpha=1$). Another example is ridge regression (RR):

$B_\alpha(X, y) = b_{RR} = (X'X + \alpha I_p)^{-1}X'y$, where $\mathcal{A} = [0, \infty]$, and its modified form LSRR (Björkström & Sundberg, 1999). There is a close connection between the two constructions CR and RR (Sundberg, 1993; de Jong & Farebrother, 1994). Other examples are continuum power regression, CPR (Wise & Ricker, 1993), and principal covariates regression, PCovR (de Jong & Kiers, 1992), the latter particularly designed for the case where the response is multivariate.

In addition to enabling comparisons, frameworks with a continuous "metaparameter" α can open new methods for regression. By optimizing over α, *i.e*, using the value that is best in the light of the available data, one may achieve a regression method that is better than the traditional ones. However, the freedom introduced by a continuous parameter implies a risk for overfitting.

In the present paper, we define a class of regression procedures indexed by a two-dimensional metaparameter $(\alpha_x\,,\,\alpha_y)$. In principle, this two-parametric set can be used to define a new regression method by optimizing over $\mathcal{A} = \mathbf{R}^2$. However, we emphasize that we are not convinced about the superiority of this method as a tool for multivariate prediction – rather, our simulations below indicate that SIMPLS does at least as well. To say anything more conclusively about the prediction ability requires more case studies than will be presented here. What we claim to have proved is an optimality result: Under a natural definition of "best", the best regressor can be found within the quite limited class that we have defined.

Nowadays, many regression problems involve more than one response variable. There is always the option to predict each of the responses without regard to the others. However, there are good reasons for avoiding this. A gross error in the training data may not be noticed unless the y-variables are analyzed simultaneously. See for example Breiman & Friedman (1997) for more arguments against mere concatenation of univariate predictions. The number of multivariate regression methods proposed and available has increased, and the question arises if they can also be tied together within a common framework by introducing metaparameters. An early attempt at this was Brooks & Stone (1994) with their *joint continuum regression* (JCR), being an extension of CR. Unfortunately, these authors concluded that jointness seldom pays. Other approaches towards the same end have been taken

by Breiman & Friedman (1997) with their "curds and whey" method, applying different shrinkage to different canonical coordinates, and by Burnham et al (1996), building frameworks that include canonical correlation regression and reduced rank regression. In the present paper, we suggest another way of bringing methods into a common frame, by generalizing the one-dimensional pathway of ridge regressors to a two-dimensional surface.

Our paper is organized as follows. After notations and terminology, we discuss the concept of factors, or latent variables, implicit in several regression methods. From a computational point of view, latent variables can be identified with linear combinations of the x- or y-variables. In Section 2 we show that most regression methods based on latent variables fall within a set that can be parameterized with two real numbers. Some well-known methods are shown to be special cases. We define a new regression method based on this principle, and in Sections 3 and 4 we compare it empirically with other methods in terms of their predictive ability. This spreads light also on the other methods.

1.2 Notation and terminology

Training data consist of n pairs, $(x_i\,, y_i)$, $1 \leq\ i\ \leq n$, where x_i is a p-vector, $x_i \in \mathbf{R}^p$, and y_i is a q-vector, $y_i \in \mathbf{R}^q$. The task is to use training data to specify an algorithm or function f such that $f(x_0)$ is likely to be a good predictor of y_0, where $(x_0\,, y_0)$ represents a new pair with known x_0 but unknown y_0. The training data are most conveniently represented by a p-vector $\bar{x}\ = \Sigma x_i/n$ and a q-vector $\bar{y}\ = \Sigma y_i/n$ of sample means, together with two matrices X $(n \times p)$ and Y $(n \times q)$ of centered data, where each row corresponds to a pair $(x_i\,, y_i)$. We restrict our search to functions $f(x_0)$ that are linear in x_0 for fixed X and Y, and centered at $(\bar{x}\,, \bar{y})$, that is, we can write

$$f(x_0) = \bar{y} + \ B(X,Y)'\,(x_0 - \bar{x})\,. \tag{1}$$

We use the term *method* for a function B intended for use in equation (1). A method is thus a matrix-valued function of two matrix arguments, $B = B(X,Y)$. For instance, we speak of the "method of ordinary least-squares" $B(X,Y) = (X'X)^{-1}X'Y$.

Alternatives to OLSR involve metaparameters. By a metaparameter we mean any variable other than X and Y that has to be specified in order to evaluate the function $B(X, Y)$. A metaparameter can be discrete, for example the number of factors to include in PCR or PLSR, or continuous, such as the parameter in RR or CR, and it may be vector-valued.

The term "regression method" is often used vaguely, ignoring specification of how to set the metaparameter. For our present discussion, it is necessary to maintain a distinction between methods and "procedures": A *procedure* describes how to obtain the matrix B, given X and Y *and* given the value for any metaparameter involved. A procedure may be denoted $\mathcal{B}_\alpha(X, Y)$ or $\mathcal{B}(X, Y; \alpha)$ where α denotes the metaparameter. A procedure with a metaparameter gives rise to a method only when it is specified how to choose the metaparameter. A "selector function" yields α as a function of the training data:

$$\alpha = \alpha_{\text{best}}(X, Y) \tag{2}$$

The expression

$$B(X, Y) = \mathcal{B}(X, Y\,;\, \alpha_{\text{best}}(X, Y)). \tag{3}$$

depends only on X and Y and thus defines a method. We use the index "best" in equation (2), since the rule for $\alpha_{\text{best}}(X, Y)$ normally is based on optimizing some function of X, Y, with respect to α. The most usual choice is cross-validation based on leaving out one observation at a time (Stone, 1974):

$$\alpha_{\text{best}}(X, Y) = \text{argmin}_\alpha \,\text{PRESS}(\alpha) \;=\; \Sigma_{i=1}^{n} |y_i - \widehat{y}_{\setminus i}|^2, \tag{4}$$

where

$$\widehat{y}_{\setminus i} = \bar{y}_{\setminus i} + \mathcal{B}(X_{\setminus i}\,, Y_{\setminus i}\,; \alpha)'\,(x_i - \bar{x}_{\setminus i}).$$

Index $\setminus i$ means that observation i was excluded from the training data. Other functions $\alpha_{\text{best}}(X, Y)$ also occur. For example, in RR the rule can be based on the ridge trace.

Each coordinate of the vector x is called an x-variable or *explanatory variable*, each coordinate of y is called a y-variable or *response variable*. We can form new explanatory variables by taking linear combinations of the given ones. We denote them $\Sigma_j c_j x_j = x'c$ $(= c'x)$, where the p-vector c is said to be a coefficient vector. Similarly, new response variables can be formed as $\Sigma_k d_k y_k = y'd$.

1.3 Factor-based methods

A similarity between many methods is that the function B of equation (1) is constructed in an iterative manner, by deriving so-called "factors", or "components". Canonical Correlation Regression (CCR), PLSR and PCR are a few examples of this kind. In the multivariate case, the procedure is as follows. A first linear combination of x-variables and a first linear combination of y-variables are selected, denote them $x'c_1$ and $y'd_1$, such that the latter is well predicted by the former. The criterion by which the pair is optimal specifies the procedure. Further pairs ($x'c_2$, $y'd_2$), ($x'c_3$, $y'd_3$), ... are then found, all optimal under additional constraints, which typically demand that each new predictor $x'c_l$ be uncorrelated with all the previous ones, $x'c_1$, ... $x'c_{l-1}$. Correlation refers to the training data, of course, so the constraint is equivalent to the two n-vectors Xc_i and Xc_j being orthogonal whenever $i \neq j$. The procedure is repeated until some stopping criterion is met, yielding a set of (say) $a \leq p$ mutually uncorrelated regressors. Note that throughout the iterative procedure, there is no requirement of orthogonality between the vectors Yd_l, or even of linear independence, between them. Only in special cases are the variables Yd_l uncorrelated, for example in CCR.

When a pairs have been determined, ordinary multivariate least squares regression is used to construct a predictor for y. In terms of linear algebra, one constructs an $n \times a$ matrix

$$T = XC \tag{5}$$

from the $p \times a$ matrix C of columns $c_1, \ldots, c_a$, and analogously an $n \times a$ matrix

$$U = YD. \tag{6}$$

The predictor (1) is constructed by OLSR of U on T, that is, one computes $B_* =$

$(T'T)^{-1}T'U$, and transforms back to the original variables, obtaining

$$B = CB_*D^+, \tag{7}$$

where D^+ is the Moore-Penrose inverse of D.

The number a of factors to use is not specified within this construction. We regard a as a metaparameter, and note that the procedure described becomes a method only jointly with some stopping rule for a.

2 A procedure with two continuous metaparameters

2.1 Variance as criterion for selection

There are many ways to define what makes a pair ($x'c$, $y'd$) optimal. Firstly, and most obviously, the correlation between $y'd$ and $x'c$ is a relevant measure. Secondly, it also seems reasonable that the sample variance of $y'd$ should be considered (*i.e,* the norm of the vector Yd). If a certain linear combination $y'd$ is nearly constant over the training data, the best predictor of ${y_0}'d$ might be to use this constant regardless of x_0. Thirdly, one should also consider the variance of the linear combination $x'c$ (*i.e,* the norm of the vector Xc). The arguments for avoiding regressors with small sample variance are well-known and need not be repeated here.

Many succesful regression methods are based on the three criteria listed above. For example, PLSR as well as PCR are based on discarding components with small $|Xc|$ or $|Yd|$. The methods give different results because the three criteria R^2, $|Xc|$ and $|Yd|$ are given different importance in the process of factor selection.

Consider any regression method for which the first factor is determined by maximizing some function $F(R^2, |Xc|^2, |Yd|^2)$ for c and d on the unit spheres $\mathbf{S}^p \subset \mathbf{R}^p$ and $\mathbf{S}^q \subset \mathbf{R}^q$ respectively. Formally, for the first factor:

$$(c_1,\, d_1) = \arg\max\{F(R^2, |Xc|^2, |Yd|^2); |c| = |d| = 1\} \tag{8}$$

It is natural that the function F should be monotone in all its three arguments. Given this, the maximum of (8) will be found on a two-dimensional surface on $\mathbf{S}^p \times \mathbf{S}^q$, as the following proposition states.

Proposition 1 *Let $F:\ \mathbf{R}^3 \to \mathbf{R}$ be a function that is monotone in all three arguments. Then, the p-vector c and q-vector d defined as*

$$(c,\, d) = \arg\max\{F(R^2, |Xc|^2, |Yd|^2); |c| = |d| = 1\}$$

will be eigenvectors of the matrices M_x and M_y respectively, where

$$M_x(\alpha_x,\, \alpha_y) \;=\; ((1-\alpha_x)X'X + \alpha_x I_p)^{-1}X'Y((1-\alpha_y)Y'Y + \alpha_y I_q)^{-1}Y'X\,, \qquad (9)$$

$$M_y(\alpha_x,\, \alpha_y) \;=\; ((1-\alpha_y)Y'Y + \alpha_y I_q)^{-1}Y'X((1-\alpha_x)X'X + \alpha_x I_p)^{-1}X'Y \qquad (10)$$

for some numbers α_x and α_y (depending on the criterion function F)

Proof: Using Proposition 2.1 from Björkström & Sundberg (1999), we can argue as follows from formula (8): First, suppose d is fixed. Then the solution c will be proportional to a "ridge regressor" with Yd as response variable:

$$c \propto ((1-\alpha_x)X'X + \alpha_x I_p)^{-1}X'Yd, \qquad (11)$$

for some number α_x. Analogously, by symmetry, when c is fixed we get

$$d \propto ((1-\alpha_y)Y'Y + \alpha_y I_q)^{-1}Y'Xc, \qquad (12)$$

for some α_y. Eliminating d between equations (11) and (12), we see that

$$c \propto ((1-\alpha_x)X'X + \alpha_x I_p)^{-1}X'Y((1-\alpha_y)Y'Y + \alpha_y I_q)^{-1}Y'X\, c.$$

Similarly, eliminating c between equations (12) and (11), we see that

$$d \propto\ ((1-\alpha_y)Y'Y + \alpha_y I_q)^{-1}Y'X((1-\alpha_x)X'X + \alpha_x I_p)^{-1}X'Y\, d.$$

Thus, c and d are eigenvectors of the matrices (9) and (10), which proves the proposition. ■

The structure of the matrices $M_x(\alpha_x,\, \alpha_y)$ and $M_y(\alpha_x,\, \alpha_y)$ perhaps stands out clearer if we note that when c is an eigenvector of $M_x(\alpha_x,\, \alpha_y)$, then Xc will be an eigenvector of the matrix

$$H_x(\alpha_x)\, H_y(\alpha_y)$$

where $H_x(\alpha_x) = X((1-\alpha_x)X'X + \alpha_x I_p)^{-1}X'$ and $H_y(\alpha_y) = Y((1-\alpha_y)Y'Y + \alpha_y I_q)^{-1}Y'$. The two matrices $H_x(\alpha_x)$ and $H_y(\alpha_y)$ are ridge type versions of the usual "hat matrix", well known in linear regression. Analogously, Yd is an eigenvector of $H_y(\alpha_y)\, H_x(\alpha_x)$.

A number of multivariate regression methods are in fact tantamount to maximizing a function like F in formula (8), at least as far as the first factor is concerned. Some examples are:

- Canonical correlation regression (CCR), where $F = R^2$ ($\alpha_x = \alpha_y = 0$)
- Partial least squares regression (PLSR), where

$$F = R^2\,|Xc|^2\,|Yd|^2 = (c'X'Yd)^2 = \text{ covariance squared},$$

- Principal components regression (PCR), where $F = |Xc|^2$.
- Reduced rank regression (RRR), in the terminology of Burnham et al (1996), and Brooks & Stone (1994; in a footnote). This procedure uses $F = R^2\,|Yd|^2$. Note that this is not standard usage of the term reduced rank regression.

In the next few sections we provide more examples. To that end, we need a result, the proof of which is given in Appendix A: When the function F in equation (8) is a product of powers of its arguments, *i.e,* when we can write

$$F(R^2,\,|Xc|^2,\,|Yd|^2) = |Xc|^{2a_x}\,|Yd|^{2a_y}\,R^{2b}, \tag{13}$$

for nonnegative numbers a_x, a_y, and b, then the resulting two parameters α_x and α_y satisfy the two equations

$$\alpha_x = \frac{a_x|Xc|^2}{a_x\,(|Xc|^2-1)\,+b}. \tag{14}$$

$$\alpha_y = \frac{a_y|Yd|^2}{a_y\,(|Yd|^2-1)\,+b}. \tag{15}$$

Note that (14) and (15) are not explicit formulas for α_x and α_y, but involve c and d, which are themselves functions of α_x and α_y. Nevertheless, the equations are useful, as we see in the next section.

2.2 Joint continuum regression (JCR)

In our notation, the first factor in Brooks & Stone's (1994) JCR is obtained by finding linear combinations of explanatory variables Xc and response variables Yd that maximize (cf. equation (1) in their paper):

$$F = |Xc|^{2\alpha/(1-\alpha)} |Yd|^2 R^2, \tag{16}$$

subject to $|c| = |d| = 1$. Inserting the exponents $a_y = b = 1$ from (16) into (14) and (15) we find $\alpha_y = 1$, and that α_x varies with the JCR parameter α according to:

$$\alpha_x = \frac{\alpha |Xc|^2}{\alpha (|Xc|^2 - 2) + 1}. \tag{17}$$

Thus, by setting $\alpha = 0$ we get the RRR method proposed by Burnham et al (1996). Setting $\alpha = 1/2$ gives $\alpha_x = 1$, and F simplifies to the squared covariance. This is equivalent with first factor PLSR. The limiting case $\alpha = 1$ gives $\alpha_x = |Xc|^2/(|Xc|^2 - 1)$. To see that this agrees with first-factor PCR, insert this α_x and $\alpha_y = 1$ in the definition (9) of $M_x(\alpha_x, \alpha_y)$ and get $M_x(\alpha_x, \alpha_y) = (X'X - |Xc|^2 I_p)^{-1} A$. Here, A is a temporary symbol for the remaining factors in (9). We see that if c is close to u_1, the largest eigenvector of $X'X$, then $|Xc|^2 \approx \lambda_1$, the largest eigenvalue of $X'X$. The matrix $(X'X - |Xc|^2 I_p)$ will be close to singular, and its inverse $(X'X - |Xc|^2 I_p)^{-1}$ will be dominated by one large column vector almost proportional to u_1. Thus, any vector c will yield approximately $M_x(\alpha_x, \alpha_y)c \propto (X'X - |Xc|^2 I_p)^{-1} Ac \propto u_1$, so that the largest eigenvector of $M_x(\alpha_x, \alpha_y)$ is u_1. This is first factor PCR.

Close to the end of their paper, Brooks & Stone mention an alternative to their JCR, namely, to maximize $R^2 |Xc|^{2\alpha/(1-\alpha)}$, *i.e.* fixing $\alpha_y = 0$ instead of $\alpha_y = 1$. This trajectory of methods thus includes canonical correlation regression (CCR), but not PLSR, except for the first factor.

2.3 Total least squares

Total least squares is a way to find an approximate solution to a system of linear equations $\mathbf{A}\,\mathbf{x} \approx \mathbf{b}$ by perturbing not only the right hand side $\mathbf{b}$ (as in the standard least squares approximation), but also the coefficient matrix $\mathbf{A}$. The multidimensional problem (the case where $\mathbf{b}$ is a matrix with more than one column)

is described in van Huffel and Vanderwalle (1991). In regression applications, Total Least Squares Regression (TLSR) means that X and Y are approximated by $\widehat{X}$ and $\widehat{Y}$ such that $\widehat{Y} = \widehat{X}\, B_{\mathrm{TLS}}$ for some $(p \times q)$-matrix B_{TLS}. The approximations are chosen so that the Frobenius norm $|\, [X\, Y] - [\widehat{X}\, \widehat{Y}]\, |_F$ is minimal. In Appendix B we show that $B_{\mathrm{TLS}} = -CD^{-1}$ where the columns of C and D are eigenvectors of (9) and (10). Thus, although TLSR is not based on identification of latent variables, it still is related to the procedures we define in equation (8) and Proposition 1.

2.4 A factor-based procedure with two continuous parameters (2PAR)

2.4.1 Subsequent factors

It was shown in Section 2.1 that if we restrict our consideration to the first factor ($a = 1$), then the set of methods obtained for varying α_x and α_y in equation (9) include all the optimal ones, as judged by any function of correlation and variances. However, regressors based on a single factor may be useful when Y is univariate (RR, LSRR) but they are not likely to be when the column span of Y is more than one-dimensional. We must therefore allow procedures which include successive pairs of eigenvectors of the matrices (9) and (10). Admittedly, for some important methods, including the NIPALS and SIMPLS versions of multivariate PLSR, this is **not** how later factors are defined. On the other hand, CCR and RRR do define later factors this way. There is also a version of multivariate PLSR called un-deflated PLS (UDPLS), suggested by Burnham et al (1996), where subsequent factors are identified in accordance with the procedure we now suggest. Therefore, there is some interest in exploring how well this class of methods compares with other types, notably the other multivariate PLS versions.

2.4.2 Definition of the method 2PAR

Consider the method obtained when the function $B(X, Y)$ of equation (1) is constructed as follows:

- Apply the procedure described in equations (5), (6) and (7), using as coefficient vectors the first eigenvectors of the matrices (9) and (10), respectively.

- Determine the number of factors, as well as the two parameters α_x and α_y by cross-validation.

Because of the two continuous metaparameters involved, we call this method "the two-parametric method" (2PAR). We next compare its predictive ability to that of some other methods, under various conditions.

3 Comparison of methods

3.1 Methods to be compared

To assess how the 2PAR method performs in competition with other regression methods, we undertake a study based on simulated data from a latent variable regression model. We compare the following six methods:

1a: Ridge regression, RR, separately for each response variable.

1b: "Least-squares ridge regression", LSRR (Björkström & Sundberg, 1999), *i.e.* RR is scaled so that the residual sum of squares is minimized. We do not include more than the first of several possible factors.

2: Univariate PLSR, that is PLSR applied to each response variable separately.

3a: Multivariate PLSR, using the NIPALS (PLS2) algorithm.

3b: Multivariate PLSR, using de Jong's (1993b) "Statistically Inspired Modification" (SIMPLS).

4: The two-parametric method 2PAR described in Section 2.4.

Among these methods, three involve continuous metaparameters (1a, 1b and 4). Four of them are based on iteratively determined factors (2, 3a, 3b and 4). Three of them have potential to exploit covariance between the y-variables (3a, 3b and 4), the others are mere concatenations of univariate methods. All six are of the form described in equation (3), based on different procedures $\mathcal{B}$, but all with selector function (4) based on "leave-one-out" cross-validation. A more detailed account of the metaparameters and the computational procedures is given in Appendix C.

3.2 Types of data, 2^4-design

We want to compare the methods under different circumstances with regard to near-collinearity and observation errors, both in X and in Y. In order to vary these conditions in a systematic way, we employ the general latent variable multivariate regression model (LVMR) of Burnham et al. (1999). The i:th observation in a data set is generated as

$$x_i' = t_i' P + \sigma_x e_i' \tag{18}$$

$$y_i' = t_i' Q + \sigma_y f_i' \tag{19}$$

where $(t_i',\ e_i',\ f_i')'$ has an $(a+p+q)$-dimensional Gaussian $\mathrm{N}(\mathbf{0}, \mathbf{I})$ distribution, and the outcomes are independent for $i = 1, \ldots, n$. The two matrices P and Q and the two standard deviations σ_x and σ_y give rise to four factors that are varied in a systematic way:

- "Factor P" denotes the condition number of P, which is substantially larger in some data sets than in others.
- "Factor Q" concerns the orientation of the a-dimensional eigenvectors of Q relative to those of P. As mentined in Appendix D, this is likely to affect the quality of the predictions.
- "Factor E" is the size of σ_x in (18).
- "Factor F" is the size of σ_y in (19).

The factor levels can be combined arbitrarily, yielding $2^4 = 16$ types of data. More details about the levels are provided in Appendix D, but generally Low is more favourable for prediction than High.

3.2.1 Size of the data sets

Our data sets have $p = 5$ explanatory variables and $q = 3$ response variables, The model has $a = 3$ latent variables. Each data set has 308 observations, of which the first $n = 8$ are used as training data and the remaining $n_0 = 300$ for validation. As

measure of the prediction errors, we use the root mean squared error of prediction, RMSEP = $\sqrt{\text{PRESS}/n_0}$ where PRESS is $\Sigma_{i=1}^{n_0}|y_i - \widehat{y}_i|^2$.

3.2.2 Variance reduction

We construct 100 data sets, "replicates", for each of the sixteen data types. However, instead of producing 1600 unrelated data sets, we simulate 100 triplets of matrices (T, E, F), and create one data set of each type with each triplet, by inserting different combinations of P, Q, σ_x and σ_y in the equations

$$X = T\,P + \sigma_x\,E \tag{20}$$

$$Y = T\,Q + \sigma_y\,F. \tag{21}$$

It is reasonable to assume that data sets created with the same random numbers will be more similar than two arbitrary sets. In the following analysis of variance, we will speak of a "triplet effect" to explain the part of the variation that can be ascribed to differences between the 100 outcomes of (T, E, F).

4 Results

4.1 Some immediate observations

Each of the 1600 data sets is regarded as an experimental unit. The six regression methods are applied to each unit and we regard the response as a six-dimensional vector of more or less correlated RMSEP-values. In Figures 1 and 2 we indicate by box plots the general magnitude and spread of the responses for two of the 16 data types. Figure 1 shows the case where all four factors are at their "low" levels. This combination of levels was designed to yield relatively favourable conditions for prediction: X and Y are well-conditioned matrices, and the error terms E and F are small. Figure 1 indicates only little difference between the methods. For a quick assessment of the general quality of the predictions, note that the term $\sigma_y f'$ in (19) is unpredictable by any method, so an RMSEP-value of about $|\sigma_y\,f'| = \sqrt{q}\,\sigma_y$ may serve as a lower bound on what can best be achieved. We have $q = 3$ and none of the six medians exceeds twice $\sqrt{3}\,\sigma_y$.

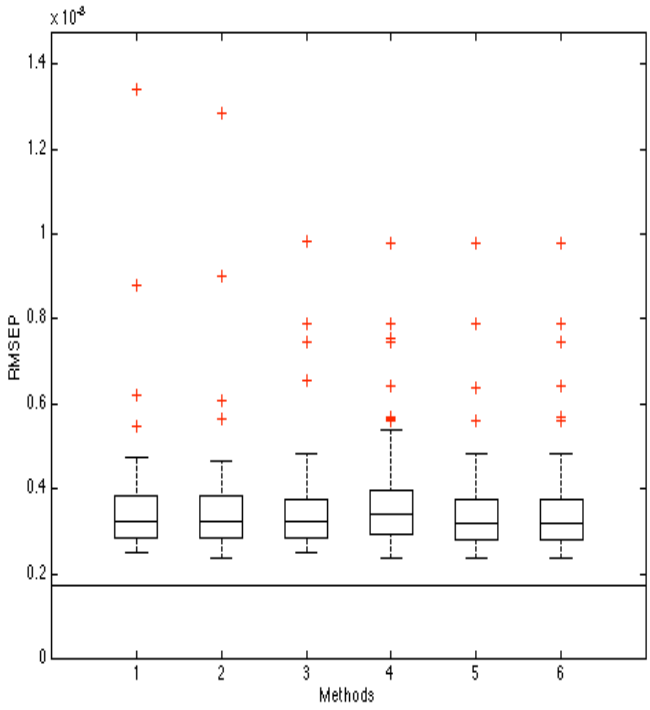

Figure 1: Box-plots for RMSEP when all factor levels are low ($i = 1$). Regression methods, from left to right: RR, LSRR, PLSR1, NIPALS, SIMPLS, 2PAR. The solid line shows $\sqrt{3}\sigma_y$. One unit on the vertical axis is 10^{-4}.

In Figure 2, the condition number of P is changed to its high level, while the other three factors remain low. All six methods do not respond equally to this change. In Figure 2, univariate PLSR, NIPALS, SIMPLS and 2PAR yield prediction errors that are roughly a factor of 10 larger than what can be ascribed to the term $\sigma_y f'$ alone. RR and LSRR are even worse, and clearly inferior to the other methods. Comparing graphs like Figures 1 and 2 for all the 16 data types, one can readily observe the following:

- **RR and LSRR behave quite similarly.** When the condition number for P is low, RR and LSRR are as good as all other methods, but when it is high, these methods are not as good as the others. In particular, they are worse when the level of factor E is low. Also, LSRR performs slightly better than RR.

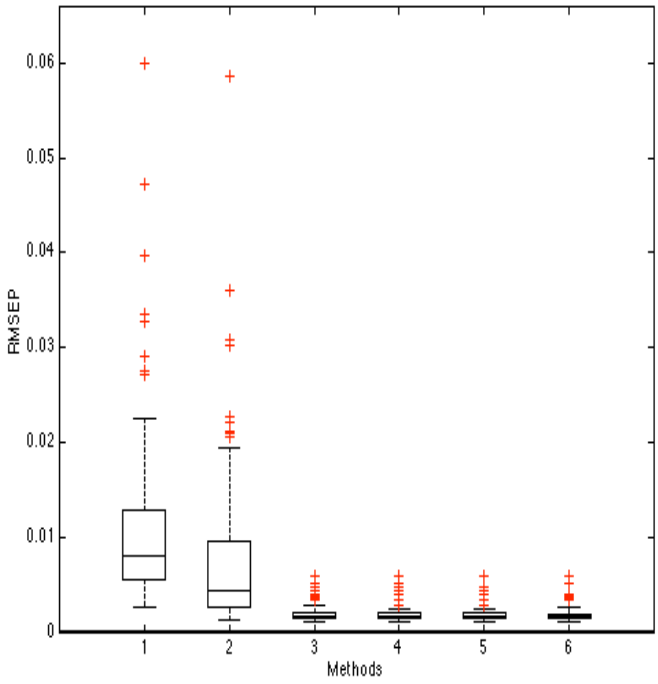

Figure 2: Box-plots for RMSEP when cond(P) is high and the other three factors are low ($i = 2$) . Regression methods, from left to right: RR, LSRR, univariate PLSR, NIPALS, SIMPLS, 2PAR. The solid line shows $\sqrt{3}\sigma_y$. Note the change of scale from Figure 1.

- **NIPALS and SIMPLS are particularly similar.** Comparing the two multivariate forms of PLSR, we note (as already pointed out by de Jong, 1993b), that they mostly give very similar results. Only for five of the 16 data types is the correlation in RMSEP between NIPALS and SIMPLS less than 1.0000 (to four decimal places). The five data types are

i	Level of factor P-Q-E-F	Corr. NIPALS SIMPLS
1	L-L-L-L	0.83
2	H-L-L-L	0.88
3	L-H-L-L	0.81
9	L-L-L-H	0.99
11	L-H-L-H	0.99

- **Skew distribution of RMSEP.** We note that most of the RMSEP distribu-

tions are skew, with all outliers found on the upper side.

In Figures 1 and 2, data are shown for each method separately, but the methods have sources of variation in common: Data type and dataset (triplet). For example, the outliers in any of Figures 1 and 2 do largely represent the same datasets for all methods. Because of the skewness, and since the effects of data type and data set are more likely to be multiplicative than additive, we prefer to continue the analysis with the natural logarithms of the RMSEP:s. Note that on the log (or ln) scale, PRESS, MSEP and RMSEP are equivalent.

4.2 Analysis of variance of ln(RMSEP)

For each combination $\{ij\}$ of the factors data type i, $i = 1, \ldots, 16$ and triplet j, $j = 1, \ldots, 100$, let Z_{ij} be the six-dimensional column vector consisting of the ln(RMSEP)-values for the six methods. We express these data additively in terms of factor effects:

$$Z_{ij} = \widetilde{\mu} + \widetilde{\alpha}_i + \widetilde{\beta}_j + \widetilde{\epsilon}_{ij} \tag{22}$$

where the terms can be interpreted in a standard ANOVA manner: For each component, $\widetilde{\mu}$ is the population average, $\widetilde{\alpha}_i$ is the effect of data type i, $\widetilde{\beta}_j$ is an effect of triplet j, and all these terms are estimated so that the residuals $\widetilde{\epsilon}_{ij}$ are as small as possible (their sum of squares is minimized).

The population average $\widetilde{\mu}$ and the data-type effects $\widetilde{\alpha}_i$ are regarded as unknown parameters to be estimated. On the other hand, since the triplets are generated at random, $\widetilde{\beta}_j$ and the residuals $\widetilde{\epsilon}_{ij}$ are considered as random vectors.

Since our main purpose is to compare methods, we write each term on the right hand side of (22) as its mean value over the six methods, plus a deviation. Using superscript k to denote the k:th component of a vector, and overbar to denote average over the six methods, we get ($k = 1, \ldots, 6$):

$$\left.\begin{array}{l} \widetilde{\mu}^{(k)} = \overline{\mu} + \mu^{(k)} \\ \widetilde{\alpha}_i^{(i)} = \overline{\alpha}_i + \alpha_i{}^{(k)} \\ \widetilde{\beta}_j^{(k)} = \overline{\beta}_j + \beta_j{}^{(k)} \\ \widetilde{\epsilon}_{ij}^{(k)} = \overline{\epsilon}_{ij} + \epsilon_{ij}{}^{(k)} \end{array}\right\} \tag{23}$$

(Alternatively, the first line of (23) could have been written $\widetilde{\mu} = \overline{\mu}\,\mathbf{1}_6 + \mu$, where $\mathbf{1}_6 = (1,1,1,1,1,1)'$, and analogously for the other lines). The variables $\overline{\epsilon}_{ij}$ are iid, and the vectors ϵ_{ij} have degenerate covariance matrices. Since for example the covariance between NIPALS and SIMPLS is different for different data types, it seems appropriate to let the variances $\mathrm{Var}(\widetilde{\epsilon}_{ij})$ depend on i. Consequently, the variances $\mathrm{Var}(\overline{\epsilon}_{ij}) = \sigma_i{}^2$ and the matrices $\mathrm{Var}(\epsilon_{ij}) = \Sigma_i$ will also vary with i.

It is straightforward to estimate all the entries in (22) and (23) in terms of the data $Z_{ij}{}^{(k)}$. Appendix E gives the expressions. The following discussion builds on the results.

4.2.1 Global mean and effects of data type

The average over all data types, methods, and replicates yields a $\overline{\mu}$ estimate of -5.02. The estimated deviations vector μ is given by

Method k =	RR	LSRR	PLS1	NIPALS	SIMPLS	2PAR
$\mu^{(k)}$estimate	0.30	0.14	-0.08	-0.11	-0.11	-0.14

The two extremes, RR and 2PAR, differ by 0.44 units. Since the response variable is the logarithm of RMSEP this means that the prediction errors with RR exceed those with 2PAR by a factor of $e^{0.44}$, or roughly 60 %. However, this is on average over all types of data. Adding the estimated effects of data type, we obtain $\widetilde{\mu} + \widetilde{\alpha}_i$, $i = 1, \ldots, 16$, see Table 1. We observe that for most types of data sets, all six methods perform almost equally well. The range $\alpha^{\max} - \alpha^{\min}$ is never wider than 0.20, except for the types of data where factor P is high and factor E is low (rows 2, 4, 10 and 12). This means that the choice of method is important only when matrix P is ill-conditioned and the errors in the X-variables are small.

Trivially, different types of data are differently predictable. It is of some interest to see in what ways different methods are sensitive to different aspects of the data. To that end, the 2^4-structure of the sixteen data types can equivalently be represented

i	Level of factor P-Q-E-F	Method average $(\bar{\mu}+\bar{\alpha}_i)$	RR	LSRR	PLS1	NIPALS	SIMPLS	2PAR	range
1	L-L-L-L	−7.98	−0.00	−0.01	0.00	0.04	−0.02	−0.01	0.06
2	H-L-L-L	−5.85	1.16	0.60	−0.39	−0.44	−0.46	−0.48	1.64
3	L-H-L-L	−7.48	−0.01	−0.01	0.00	0.02	0.00	0.00	0.03
4	H-H-L-L	−3.83	1.36	0.74	−0.52	−0.52	−0.51	−0.55	1.91
5	L-L-H-L	−6.09	−0.03	−0.03	0.06	0.02	0.02	−0.02	0.09
6	H-L-H-L	−3.97	0.07	0.00	0.03	−0.02	−0.02	−0.06	0.13
7	L-H-H-L	−5.31	−0.04	−0.04	0.04	0.02	0.02	−0.01	0.08
8	H-H-H-L	−2.00	0.12	0.02	−0.03	−0.03	−0.03	−0.06	0.18
9	L-L-L-H	−5.93	−0.01	−0.02	0.04	0.01	0.00	−0.02	0.06
10	H-L-L-H	−5.43	0.80	0.35	−0.23	−0.30	−0.30	−0.32	1.12
11	L-H-L-H	−5.91	−0.02	−0.02	0.04	0.01	0.01	−0.02	0.06
12	H-H-L-H	−3.81	1.33	0.73	−0.50	−0.52	−0.52	−0.54	1.87
13	L-L-H-H	−5.65	−0.04	−0.04	0.09	0.01	0.01	−0.03	0.13
14	H-L-H-H	−3.96	0.07	0.00	0.04	−0.02	−0.02	−0.06	0.13
15	L-H-H-H	−5.16	−0.04	−0.04	0.07	0.01	0.01	−0.01	0.11
16	H-H-H-H	−2.00	0.12	0.02	−0.03	−0.03	−0.03	−0.06	0.18

Table 1: Averages, and deviations from average, by data type. The entries in the leftmost column are the estimated method averages, $\overline{\mu}+\overline{\alpha}_i$. The other six columns give $\mu^{(k)}+{\alpha_i}^{(k)}$ (eq. 23) for method k, $k=1,\ldots,6$. The last column is the difference between the largest and smallest deviation.

by main effects and interaction effects, see Table 2. (We define the effect of a factor to be the average increase observed in ln(RMSEP) when the factor is changed from low to high, and all other factors are unchanged.) From Table 2, we observe the following:

- On the whole, the two "ridge-type" methods (RR and LSRR) behave similarly, and the other four (the "factor-based" methods) are almost mutually identical, nowhere differing more than 0.04. This is seen from the pattern of positive and negative signs and the actual values in the four rightmost columns.

- For all methods, the most important factor is the condition number of P. When it increases from 3 to 3^4, the RMSEP goes up by a factor of $e^{2.33} \approx 10$. This varies from $e^{2.33-0.25} \approx 8$ for 2PAR to $e^{2.33+0.66} \approx 20$ for RR.

Factor effect	Methods average	Individual method effect deviations					
		RR	LSRR	PLS1	NIPALS	SIMPLS	2PAR
P	2.33	0.66	0.33	-0.24	-0.25	-0.24	-0.25
Q	1.17	0.10	0.07	-0.07	-0.04	-0.03	-0.03
E	1.51	-0.55	-0.31	0.23	0.21	0.22	0.20
F	0.58	-0.05	-0.04	0.04	0.01	0.02	0.02
PQ	0.72	0.11	0.07	-0.06	-0.04	-0.04	-0.04
PE	0.24	-0.52	-0.29	0.18	0.21	0.20	0.21
PF	-0.47	-0.05	-0.03	0.00	0.03	0.02	0.03
QE	0.13	-0.08	-0.06	0.03	0.04	0.03	0.04
QF	-0.15	0.04	0.03	-0.02	-0.02	-0.02	-0.02
EF	-0.43	0.05	0.04	-0.03	-0.02	-0.03	-0.02
PQE	-0.06	-0.08	-0.06	0.04	0.03	0.04	0.03
PQF	0.05	0.04	0.03	-0.02	-0.02	-0.02	-0.02
PEF	0.32	0.05	0.03	-0.02	-0.02	-0.02	-0.02
QEF	0.07	-0.04	-0.03	0.02	0.02	0.02	0.02
PQEF	0.03	-0.04	-0.03	0.02	0.02	0.02	0.02

Table 2: Main and interaction effects of the four factors (cf. App. D). The Methods average column represents the factorial effects on average over all methods. The last six columns give the deviations from this average for each of the six methods.

- Also for all methods, factors Q and E are the most important ones after P. For factor-based methods, errors in x-variables (factor E) is more harmful than an unfavourable ordering of the latent variables (factor Q), while the converse holds for the ridge-type methods.

- For all methods, the fourth most important effect is the interaction between factors P and Q. The joint effect of High level in these factors is worse than the sum of main effects. The fifth is factor F (size of errors in the y-variables).

- Factor F interacts negatively with the other three factors. More precisely, once one or more of the factors P, E or Q are at high level (and thus deteriorate the conditions for prediction), the additional damage caused by large errors in the y-variables (Factor F = High) is minor.

- The individual methods deviations are seen to be substantial only for effects P, E and PE. Furthermore, the near equality of effect rows P, -E and -PE shows

that the methods essentially differ precisely for the combination P = High, E = Low. Under these conditions, corresponding to rows (i-values) 2, 4, 10, 12 in Table 1, RR and LSRR perform worse than the others, as noted already in Section 4.1, in connection with Figure 2.

4.2.2 Matrix triplet effects

We regard the terms $\widetilde{\beta}_j$ in (22) as independent outcomes of a six-dimensional random vector. The methods averages $\overline{\beta}_j$, $j = 1, \ldots, 100$, as defined in (23) form a somewhat right-skewed set of numbers, although not more so than to make a Gaussian model reasonable. Their mean is zero by construction, and the standard deviation is 0.26, which means that $\overline{\beta}_j$ contributes to the right hand side in (22) by roughly the same amount as, for example, an interaction effect between two of the systematic factors (cf. section 4.2.1). The estimated standard deviations of ${\beta_j}^{(k)}$ are:

Component	$\overline{\beta}_j$	RR	LSRR	PLS1	NIPALS	SIMPLS	2PAR
std estimate	0.26	0.13	0.16	0.08	0.07	0.07	0.07

The off-diagonal terms in $\widehat{\mathrm{Var}}(\beta_j)$ correspond to the following correlation coefficients:

	RR	LSRR	PLS1	NIPALS	SIMPLS	2PAR
RR	1.00	0.75	−0.87	−0.89	−0.88	−0.78
LSRR		1.00	−0.91	−0.93	−0.93	−0.77
PLS1			1.00	0.94	0.91	0.70
NIPALS				1.00	1.00	0.70
SIMPLS					1.00	0.72
2PAR						1.00

Again we see a tendency for the two ridge-type methods to be similar, and different from the four factor-based methods, those in the latter group also being mutually similar. All correlations between methods in different groups are negative. A typical ${\beta_j}^{(k)}$ is approximately half of a typical $\overline{\beta}_j$. Since we are interested in comparing different regression methods on the same data, we note that for a fixed j and two different k-values, the standard deviaton of the difference is $\mathrm{SD}\,(\,{\beta_j}^{(k_1)} - {\beta_j}^{(k_2)}) \approx 0.1$.

4.2.3 Residuals

The residuals $\widetilde{\epsilon}_{ij}$ in (22) represent lack of additivity and therefore also tell to what extent the model can predict which method will be best. As mentioned in connection with the decomposition (23), we allow the variance matrices $\mathrm{Var}(\widetilde{\epsilon}_{ij})$ to be different for different combinations of the factors P, Q, E and F. Consequently, the variances of the method averages, $\mathrm{Var}(\overline{\epsilon}_{ij}) = {\sigma_i}^2$ will vary with data type, i, as will the variance matrices for the deviations $\mathrm{Var}(\epsilon_{ij}) = \Sigma_i$. They are estimated together with the variances for the triplet effects, as sketched in Appendix F. Examples of results are shown in Table 3. We see that all $\widehat{\sigma}_i$ are between 0.11 and 0.22, so the term $\overline{\epsilon}_{ij}$ is comparable to $\bar{\beta}_i$ (since $\hat{\sigma}_\beta = 0.26$, cf. Section 4.2.2). Of more interest, however, are the variances for and correlations between the method deviations ϵ_{ij}, because they tell to what extent we can say that one method is better than another. In Table 3 we show the estimated standard deviations of $\overline{\epsilon}_{ij}$ and of the six components of ϵ_{ij}, together with the NIPALS/SIMPLS correlation. The table confirms the observation made in section 4.1, that the data types can be grouped in two categories: One for which NIPALS and SIMPLS yield substantially different predictions, one where those two methods agree almost perfectly. The first category turns out to consist of the data types where factor E = Low and at most one of the other factors is at level High (index $i = 1, 2, 3$ and 9). The other category consists of the remaining 12 data types. However, those two categories are not systematically different with regard to any other element in the covariance or correlation matrices.

In principle, Table 3 could be expanded to the right, with all possible pairwise correlations. However, the entries do not vary with the levels of P, Q, E or F in any regular way.

Because of the random terms β_j and ϵ_{ij} in (22), it is by no means certain that the method with the most favorable systematic terms will yield the smallest prediction error *for a given data set*. Therefore we supplement the previous tables with Table 4 which shows for how many of the 100 data sets a method "wins" (i.e. yields smaller RMSEP than the others. It is interesting to note that a ridge-type method is "best" in roughly half of the simulations, except for the data types where P = High and

i	Data type P-Q-E-F	SD($\overline{\epsilon}_{ij}$)	RR	LSRR	PLS1	NIPALS	SIMPLS	2PAR	Corr. NIP / SIM
1	L-L-L-L	0.16	0.16	0.19	0.10	0.16	0.10	0.10	0.31
2	H-L-L-L	0.15	0.25	0.40	0.20	0.16	0.18	0.15	0.79
3	L-H-L-L	0.11	0.16	0.15	0.11	0.11	0.11	0.12	0.16
4	H-H-L-L	0.22	0.32	0.51	0.20	0.20	0.20	0.19	1.00
5	L-L-H-L	0.21	0.16	0.18	0.16	0.09	0.09	0.09	1.00
6	H-L-H-L	0.12	0.08	0.12	0.14	0.08	0.08	0.07	1.00
7	L-H-H-L	0.16	0.16	0.18	0.09	0.09	0.10	0.11	1.00
8	H-H-H-L	0.19	0.18	0.19	0.09	0.10	0.10	0.10	1.00
9	L-L-L-H	0.18	0.16	0.18	0.14	0.10	0.10	0.09	0.87
10	H-L-L-H	0.15	0.23	0.28	0.16	0.13	0.13	0.14	1.00
11	L-H-L-H	0.17	0.16	0.18	0.12	0.10	0.09	0.09	0.94
12	H-H-L-H	0.21	0.31	0.50	0.19	0.20	0.19	0.18	1.00
13	L-L-H-H	0.17	0.18	0.20	0.19	0.12	0.12	0.09	1.00
14	H-L-H-H	0.11	0.08	0.12	0.14	0.08	0.08	0.07	1.00
15	L-H-H-H	0.13	0.15	0.17	0.13	0.09	0.09	0.10	1.00
16	H-H-H-H	0.19	0.18	0.19	0.09	0.10	0.10	0.09	1.00

Table 3: Estimated standard deviations for the residual terms for all sixteen types of data sets. Column 3 shows estimated standard deviations, of $\overline{\epsilon}_{ij}$, columns 4 - 9 for each component of ϵ_{ij}. Column 10 shows the correlation coefficient between ${\epsilon_{ij}}^{(NIPALS)}$ and ${\epsilon_{ij}}^{(SIMPLS)}$.

E = Low.

5 Conclusions

In this paper we have demonstrated connections between various methods for construction of predictors for use in multivariate linear regression when explanatory and/or response variables are near-collinear. We have extended several conventional methods, in the form of a method with two continuous meta-parameters, and with the number of factors as an additional parameter. A method comparison has been carried out based on data with $n = 8$, $\dim(x) = 5$ and $\dim(y) = 3$, simulated in accordance with a latent variable multivariate multiple regression model. Within this framework, defined in (18) and (19), different types of data can be simulated. Six methods were compared, evaluating a PRESS-based performance measure by a test set. We conclude that for most data types the choice of method does not

	Data type					
i	P-Q-E-F	RR	LSRR	PLS1	PLSR	2PAR
1	L-L-L-L	25	28	14	18	15
2	H-L-L-L	0	3	29	31	37
3	L-H-L-L	27	20	17	17	19
4	H-H-L-L	0	4	29	30	37
5	L-L-H-L	27	27	10	23	13
6	H-L-H-L	8	26	19	18	29
7	L-H-H-L	22	24	8	20	26
8	H-H-H-L	13	29	14	23	21
9	L-L-L-H	23	29	14	20	14
10	H-L-L-H	0	5	31	27	37
11	L-H-L-H	25	21	19	9	26
12	H-H-L-H	0	4	41	21	34
13	L-L-H-H	25	27	14	18	16
14	H-L-H-H	9	24	21	17	29
15	L-H-H-H	23	28	13	15	21
16	H-H-H-H	13	29	20	21	17

Table 4: Number of simulations where the method "wins". Each row sums to 100. Columns for NIPALS and SIMPLS have been merged in column PLSR, because of the similarity between those two methods

affect the quality of the predictions. In the cases where method is important, factor-based methods predict better than one-factor ridge-type methods. Concatenation of univariate PLSR leads to somewhat larger errors than "genuinely" multivariate methods (SIMPLS, NIPALS 2PAR). On average over 100 replicates, 2PAR turns out best (if only marginally) in 8 of the 16 data types. In no case does 2PAR rank worse than third. When 2PAR is surpassed, the better methods are almost always RR and/or LSRR. Only for one of the data types is 2PAR second to another factor-based method. However, because of the amount of computation involved and in the light of Table 4, the improvement does not appear high enough to motivate a general recommendation of the method.

REFERENCES

Björkström, A. and Sundberg, R. (1999): A Generalized View on Continuum Regres-

sion. *Scand. J. Statist.* **26**: 17–30.

Breiman, L. and Friedman, H. (1997): Predicting multivariate responses in multiple linear regression. (With discussion) *J. R. Statist. Soc. B*, **59**: 3–54.

Brooks, R. and Stone, M. (1994) Joint continuum regression for multiple predictands. *J. Am. Statist. Ass.*, **89**, 1374–1377.

Brown, P. J. (1993) *Measurement, Regression, and Calibration*. Oxford Univ. Press, Oxford.

Burnham, A.J., MacGregor, J.F. and Viveros, R. (1999): Latent variable multivariate regression modelling. *Chemom. Intell. Lab. Syst.*, **48**, 167–180.

Burnham, A.J., Viveros, R. and MacGregor, J.F. (1996): Frameworks for latent variable multivariate regression. *J. Chemometrics* **10**, 31–45

de Jong, S. (1995) PLS shrinks. *J. Chemometrics*, **9**, 323–326.

de Jong, S. and Farebrother, R.W. (1994) Extending the relationship between ridge regression and continuum regression. *Chemom. Intell. Lab. Syst.*, **25**, 179–181.

de Jong, S. (1993a) PLS fits closer than PCR *J. Chemometrics*, **7**, 551–557.

de Jong, S. (1993b) SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **18**, 251–263.

de Jong, S. and Kiers, H. A. L. (1992) Principal covariates regression. Part I. Theory. *Chemom. Intell. Lab. Syst.* **14**, 155–164.

de Jong, S., Wise, B.M. and Ricker, L.N. (2001) Canonical partial least squares and continuum power regression. *J. Chemometrics*, **15**, 85–100.

Frank, I.E. and Friedman, J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

Goutis, C. (1996) Partial Least Squares Algorithm Yields Shrinkage Estimators. *Annals of Statistics* **24:2**, 816-824.

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

James, W. & Stein, C. (1961) Estimation with quadratic loss. In *Proceedings of the 4th Berkeley symposium*, **1**, 361 - 379. University of California Press, Berkeley.

Kalivas, J. H. (1999) Interrelationships of multivariate regression methods using eigenvector basis sets *J. Chemometrics*, **13**, 111–132.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. (With discussion) *J. R. Statist. Soc. B*, 36, 111–147.

Stone, M. and Brooks, R. J. (1990) Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. (With discussion) *J. R. Statist. Soc.* B, **52**, 237-269; Corrigendum (1992) **54**, 906–907.

Sundberg, R. (1999) Multivariate calibration – direct and indirect regression methodology. (With disc.) *Scand. J. Statist.* **26**, 161 – 207.

Sundberg, R. (1993) Continuum regression and ridge regression. *J. R. Statist. Soc.* B, **55**, 653–659.

van Huffel, S. and Vandewalle, J. (1991) *The Total Least Squares Problem. Computational Aspects and Analysis*. SIAM

Wise, B.M. and Ricker, L.N. (1993) Identification of finite impulse response models with continuum regression. *J. Chemometrics*, **7**, 1–14.

Appendices

A Interpretation of the two parameters

In Section 2.1 we stated without proof how the two "ridge parameters" α_x and α_y depend on the coefficients of the function F of equation (8) when F is a product of powers of its arguments, These results, formulas (14) and (15), are derived here. Thus, assume

$$F(R^2, |Xc|^2, |Yd|^2) = |Xc|^{2a_x} |Yd|^{2a_y} R^{2b}, \tag{24}$$

with a_x, a_y and b all nonnegative. We maximize F subject to the two constraints $|c|^2 = 1$ and $|d|^2 = 1$. Lagrange's method implies that at the required point, $\nabla \log F$ is in the span of $\nabla|c|^2$ and $\nabla|d|^2$. We have

$$\log F = a_x \log |Xc|^2 + a_y \log |Yd|^2 + b \log R^2.$$

With $K = c'X'Yd$ = covariance, we can write $R^2 = K^2/(|Xc||Yd|)^2$ and

$$\log F = (a_x - b) \log |Xc|^2 + (a_y - b) \log |Yd|^2 + b \log K^2.$$

Thus,

$$\nabla \log F = \frac{a_x - b}{|Xc|^2} \nabla |Xc|^2 + \frac{a_y - b}{|Yd|^2} \nabla |Yd|^2 + \frac{b}{K^2} \nabla K^2.$$

Evaluating the gradients and introducing Lagrangian multipliers we get

$$\frac{a_x - b}{|Xc|^2} \begin{pmatrix} 2X'Xc \\ 0 \end{pmatrix} + \frac{a_y - b}{|Yd|^2} \begin{pmatrix} 0 \\ 2Y'Yd \end{pmatrix} + \frac{2b}{K} \begin{pmatrix} X'Yd \\ Y'Xc \end{pmatrix} = 2\kappa_1 \begin{pmatrix} c \\ 0 \end{pmatrix} + 2\kappa_2 \begin{pmatrix} 0 \\ d \end{pmatrix} \tag{25}$$

The upper part reads

$$2\frac{a_x - b}{|Xc|^2} X'Xc + \frac{2b}{K} X'Yd = 2\kappa_1 c.$$

Multiplying this equation by c' and using $|c| = 1$ gives $\kappa_1 = a_x$. We thus have

$$\frac{a_x - b}{|Xc|^2} \left(X'X - \frac{a_x |Xc|^2}{a_x - b} I_p \right) c \propto X'Yd.$$

Comparison with equation (11) now demonstrates that

$$\alpha_x/(1 - \alpha_x) = -a_x |Xc|^2/(a_x - b),$$

that is we get the following expression for the first "ridge parameter" as function of the exponents a_x, a_y and b:

$$\alpha_x = \frac{a_x |Xc|^2}{a_x \left(|Xc|^2 - 1\right) + b}. \tag{26}$$

By analogy, the other parameter satisfies

$$\alpha_y = \frac{a_y |Yd|^2}{a_y \left(|Yd|^2 - 1\right) + b}. \tag{27}$$

In CCR, we use the exponents $a_x = a_y = 0$, $b = 1$. In PLSR we use $a_x = a_y = b = 1$, and in RRR we use $a_x = 0$, $a_y = b = 1$.

B Connection between 2PAR and TLSR

We show here that $B_{\mathrm{TLS}} = -C\,D^{-1}$. Since the column vectors of C and D are eigenvectors of (9) and (10), this expression illustrates that TLSR is closely related with 2PAR.

The equation $\widehat{Y} = \widehat{X}\,B_{\mathrm{TLS}}$ can be written

$$[\widehat{X} \quad \widehat{Y}] \begin{pmatrix} B_{\mathrm{TLS}} \\ -I_q \end{pmatrix} = 0 \tag{28}$$

which shows that the concatenated matrix $[\widehat{X} \quad \widehat{Y}]$ must have a null space of dimension at least q. If the null space dimension of $[X \quad Y]$ is less than q, one needs to replace some of the smallest singular values of $[X \quad Y]$ by zeros. Any right singular vector of $[\widehat{X} \quad \widehat{Y}]$ can be written $(c'\,d')'$ where c is a p-vector and d is a q-vector. The collection of all the q singular vectors corresponding to singular value 0 can be written as a $(p+q) \times q$ matrix $(C'\,D')'$, and we have

$$[\widehat{X} \quad \widehat{Y}] \begin{pmatrix} C \\ D \end{pmatrix} = 0.$$

Multiplication by D^{-1} and comparison with (28) shows that $B_{\mathrm{TLS}} = -C\,D^{-1}$. Further, any singular vector of $[X \quad Y]$ is an eigenvector of $\begin{pmatrix} X' \\ Y' \end{pmatrix} [X \quad Y]$. Thus there exists a number λ such that

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} [X \quad Y] \begin{pmatrix} c \\ d \end{pmatrix} = \lambda \begin{pmatrix} c \\ d \end{pmatrix}$$

From this, it is straightforward to conclude that c and d are eigenvectors of matrices of the types (9) and (10). The parameters α_x and α_y will both equal $\lambda/(\lambda - 1)$, a number that can be negative.

C Metaparameters and evaluation of the selector function

For methods 1a and 1b, the metaparameter is a q-vector of ridge constants (one for each response variable). For univariate and multivariate PLSR, the metaparameter is the number of factors, *i.e.*, an integer between 1 and p. In the univariate

case (Method 2) we determine this number for each response variable separately, so the parameter is a vector of q integers. For method 4, the metaparameter is the triplet (α_x, α_y, a), where a is the number of factors. For all methods, the function $\alpha_{\text{best}}(X, Y)$ is based on leave-one-out cross-validation. We evaluate the function α_{best} (equation 4) by trial and error: A large number of candidate α-values are tried, and the one that yields the smallest PRESS-value is chosen.

- For the ridge parameter α in RR and LSRR, we test 101 α-values, corresponding to $\alpha = e^{-i/10}$, $i = 0, \ldots, 100$. The candidates thus range from $\alpha = e^{-10} \approx 4.5 \text{ x } 10^{-5}$ to $\alpha = e^0 = 1$. The two limiting values $\alpha = 0$ and $\alpha = 1$ correspond to OLSR and one-factor PLSR, respectively.

- In the three forms of PLSR, (2, 3a and 3b), the number a of factors constitutes the metaparameter. The possible values are the integers 1, 2, ... p, and the choice $a = p$ yields the OLSR regressor. In case 2, we permitted a to be different for the three response variables.

- For 2PAR (Method 4 above), the metaparameter consists of two continuous parameters α_x and α_y, and an integer a denoting the number of factors. The latter is one of the integers $1, \ldots, \min(p, q)$. As for α_x and α_y, we try 300 pairs of values, 100 along each edge of the triangle shown in Figure 3. The set of combinations thus has $\min(p, q) \times 300$ elements.

D Levels of the systematic factors

Since we want to compare the six methods under several different circumstances with regard to near-collinearity, error sizes, *etc*, we vary the parameters so they represent different types of data. We define sixteen types by letting four factors vary on two levels each (a "low" and a "high" level):

- **Factor P:** The matrix P is set to be

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2^{-w} & 0 & 0 & 0 \\ 0 & 0 & 3^{-w} & 0 & 0 \end{pmatrix} \tag{29}$$

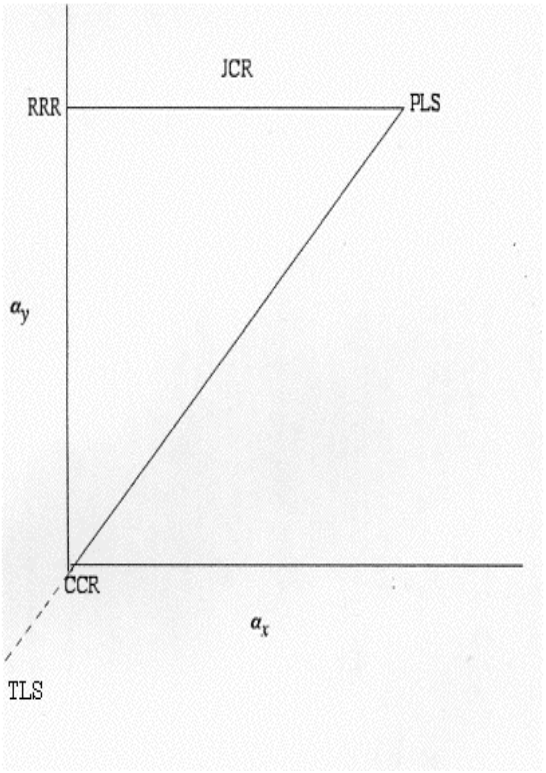

Figure 3: Possible combinations of α_x and α_y to be explored for the two-parametric method. The corners of the triangle correspond to CCR ($(\alpha_x, \alpha_y) = (0, 0)$) , PLSR ($(\alpha_x, \alpha_y) = (1, 1)$), and RRR ($(\alpha_x, \alpha_y) = (0, 1)$). The edge from PLS to RRR corresponds to Brooks & Stone JCR. TLS is not among the methods we compare, but it is based on regresssors corresponding to points on the dashed line.

where $w = 1$ for half of the data sets ("low" level), and $w = 4$ for the other half. Hence, the condition number for P differs by a factor of 3^3 between the low and high cases.

- **Factor Q:** The matrix Q is given by

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.1 \end{pmatrix} \tag{30}$$

(called its low level) in half of the simulations, and

$$Q = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{31}$$

(called high level) in the other half. The motivation for this is as follows. Because of the structure (29) of the matrix P, the latent variable t_1 exerts greatest influence on the x-variables, t_2 second largest and t_3 least. It seems then that the possibility to predict y from x would be best if the response variables were also affected primarily by t_1 and to lesser extents by the other two latent variables. By contrast, the worst situation would be if the y-variables were controlled mainly by t_3. Our two choices of Q thus probably represent two extreme situations.

- **Factor E:** In the term $\sigma_x\, e'$ in (18), the standard deviation σ_x is taken to be 10^{-4} in the low case and 10^{-3} in the high case.

- **Factor F:** In the term $\sigma_y\, f'$ in (19), the standard deviation σ_y is also taken to be 10^{-4} in the low case and 10^{-3} in the high case.

E Effect estimators

Define 6-vectors $\overline{Z}_{i\cdot} = (1/100)\Sigma_j Z_{ij}$, $\overline{Z}_{\cdot j} = (1/16)\Sigma_i Z_{ij}$, and $\overline{Z}_{\cdot\cdot} = (1/1600)\Sigma_{ij} Z_{ij}$. The following table then shows how to estimate the effects (eq. 23) in terms of the data:

	Term	Estimator	Comment
1	$\widetilde{\mu}$	$\overline{Z}_{\cdot\cdot}$	
2	$\widetilde{\alpha}_i$	$\overline{Z}_{i\cdot} - \overline{Z}_{\cdot\cdot}$	
3	$\widetilde{\beta}_j$	$\overline{Z}_{\cdot j} - \overline{Z}_{\cdot\cdot}$	
4	$\widetilde{\epsilon}_{ij}$	$Z_{ij} - \overline{Z}_{i\cdot} - \overline{Z}_{\cdot j} + \overline{Z}_{\cdot\cdot}$	
5	$\overline{\mu}$	$(1/6)\Sigma_{k=1}^{6} \overline{Z}_{\cdot\cdot}^{(k)}$	
6	$\overline{\alpha}_i$	$(1/6)\Sigma_{k=1}^{6} \widetilde{\alpha}_i^{(k)}$	$\widetilde{\alpha}_i$ from row 2
7	$\overline{\beta}_j$	$(1/6)\Sigma_{k=1}^{6} \widetilde{\beta}_j^{(k)}$	$\widetilde{\beta}_j$ from row 3
8	$\overline{\epsilon}_{ij}$	$(1/6)\Sigma_{k=1}^{6} \widetilde{\epsilon}_{ij}^{(k)}$	$\widetilde{\epsilon}_{ij}$ from row 4

The vectors without tilde are then estimated componentwise. For each $k = 1, \ldots, 6$:

	Term	Estimator	Comment
9	$\mu^{(k)}$	$\tilde{\mu}^{(k)} - \overline{\mu}$	rows 1 and 5 used
10	${\alpha_i}^{(k)}$	$\widetilde{\alpha}_i^{(k)} - \overline{\alpha}_i$	rows 2 and 6 used
11	${\beta_j}^{(k)}$	$\widetilde{\beta}_j^{(k)} - \overline{\beta}_j$	rows 3 and 7 used
12	${\epsilon_{ij}}^{(k)}$	$\widetilde{\epsilon}_{ij}^{(k)} - \overline{\epsilon}_{ij}$	rows 4 and 8 used

F Variance matrix estimators

With 16 data types and 100 triplets we obtain:

$$\mathrm{E}[\Sigma_{ij}(Z_{.j} - Z_{..})\,(Z_{.j} - Z_{..})'\,] \;=\; (100-1)\Big(16\,\mathrm{Var}(\widetilde{\beta}_j) \;+ \bar{V}\Big) \tag{32}$$

where $\bar{V}$ is the average of the $\mathrm{Var}(\widetilde{\epsilon}_{ij})$, $i = 1, \ldots, t$. In order to estimate each $\mathrm{Var}(\widetilde{\epsilon}_{ij})$ separately, we use the estimated residuals $(Z_{ij} - Z_{i.} - Z_{.j} + Z_{..})$ together with the expression, valid for each fixed i,

$$\begin{aligned} &\mathrm{E}[\Sigma_j(Z_{ij} - Z_{i.} - Z_{.j} + Z_{..})\,(Z_{ij} - Z_{i.} - Z_{.j} + Z_{..})'\,] \;= \\ &= \;(100-1)\Big((1 - \frac{2}{16})\mathrm{Var}(\widetilde{\epsilon}_{ij}) \;+\; \frac{1}{16}\bar{V}\Big) \end{aligned} \tag{33}$$

By letting the observed sums of squares and products estimate their expected values, we derive a system of linear equations from (33). Solving this system gives an estimate of $\mathrm{Var}(\widetilde{\epsilon}_{ij})$, for each i, and with the aid of (32) we calculate an estimate of $\mathrm{Var}(\widetilde{\beta}_j)$ also. Since $\beta_j \;=\; Q\,\widetilde{\beta}_j$, where Q is a matrix corresponding to column-centering, we obtain an unbiased estimate of $\mathrm{Var}(\beta_j)$ by $\widehat{\mathrm{Var}}(\beta_j) \;=\; Q\,\widehat{\mathrm{Var}}(\widetilde{\beta}_j)\,Q'$. Further, the variance matrices for $\widetilde{\epsilon}_{ij}$ can be estimated from (32) and (33), as can the variances $\mathrm{Var}(\bar{\epsilon}_{ij})$ and the matrices $\mathrm{Var}(\epsilon_{ij}) \;=\; \mathrm{Var}(Q\widetilde{\epsilon}_{ij})$, $i = 1, \ldots, 16$.