Model reduction for prediction in regression models.

Inge Helland[*]

Abstract

We look at prediction in regression models under mean square loss for the random x case with many explanatory variables. Model reduction is done by conditioning upon only a small number of linear combination of the original variables. For each dimension a simple theoretical condition on the selection matrix is motivated from the mean square error. The corresponding reduced model will then essentially be the population model considered earlier for the chemometricians' partial least squares algorithm. Estimation of the selection matrix under this model is briefly discussed, and analoguous results for the case with multivariate response and for the classification case are formulated. Finally, it is shown that an assumption of multinormality may be weakened to assuming elliptically symmetric distribution, and that some of the results are valid without any distributional assumption at all.

KEYWORDS AND PHRASES: classification; expected squared prediction error; invariant space; model reduction; partial least squares regression; prediction; random x; regression analysis.

1 Introduction.

The regression model in its usual form

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}, \tag{1}$$

where $\mathbf{X}$ is $n \times p$ and $\mathbf{e}$ is $\mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, is one of the most successfull known statistical models from an applied point of view; yet its very form is defective in one respect, since any model that is conditioned upon a set of variables like the x-variables here necessarily contains no information about the distribution of these variables themselves. Even in the common situation where these are observed variables, not fixed

[*] Department of Mathematics, University of Oslo, Box 1053 Blindern, N-0316 Oslo, Norway. (ingeh@math.uio.no)

design-variables, it is still the undisputable procedure in nearly every application to take all information from the model conditioned upon all the x's as in (1). As an example of a conflict arising from this, it is very difficult to interprete the squared multiple correlation R^2 in any reasonable way without taking the distribution of the x-variables into account (see Helland, 1987 and references there). Some arguments for the ordinary conditioning can be given when the direction of prediction is from $\mathbf{x}$ to y, but other forms of conditioning are possible, and may also be useful, as will be seen below.

An even more well-known problem is implied by the situation when p is large - say of the same order as n or even larger. Then β cannot (or can hardly) be estimated by least squares because of collinearity. As a consequence, the standard regression method cannot be used directly to predict new y-variables from a new set of x-variables. This is in fact one of the great paradoxes of statistics: An increase in information in terms of an increase in the number of explanatory variables may typically in this sense make prediction more difficult, not easier.

There are lots of statistical methods whose object is to improve upon this situation: Subset selection, ridge regression, shrinkage methods, principal component regression, partial least squares regression and so on, and a lot has been written on the pros and cons of the various methods. In this paper we will not concentrate on methods, but on models. It is known that the ordinary regression methods usually function well when the number of explanatory variables is not too large compared to the number of observations. It seems also to be generally accepted that, roughly speaking, a large data set requires a more complicated model than a small data set. Taking the consequences of this way of thinking, a natural question is: With a given data size, how can a regression model be reduced in an optimal or near optimal way from the point of view of prediction? In general there are two ways to achieve a model reduction: through a change of conditioning and through parameter restriction, and in the simplest case these are equivalent, as will be shown below. The most important task, though, is to find the best reduced model, or at least some nearly best model, and this is not a trivial task in general.

Traditionally, statisticians are accustomed to keeping the same single model all the way from the initial model building to the final data analysis, but informally, model reduction has been used in all branches of applied statistics, both in estimation and in prediction problems. It is easy to find examples where it may pay to reduce the number of parameters in models when the data set is limited; a systematic likelihood-based theory for this has recently been given by Hjort (1998). Here we will present some main ideas for a general approach aimed at prediction in regression models, first for the case with multinormal observations. Inspiration for the theory comes from methods developed in chemometry, but we emphasize again that we will primarily discuss models, not methods, and that the arguments used here to reduce regression models, have little to do whith chemometric methodology.

When reading the paper, it may be useful to have the analogue to variable selection in mind. As is well known, this term denotes the methods where one starts with the class of all possible regression models with subsets of the original x-variables as explanatory variables (or some large subclass of this class), and then use data to choose between the models. In the present paper we look at the class of regression models with a set of *linear combinations* of the original x's as explanatory variables and limit the class to those who theoretically seem to give the best predictions. This choice will depend upon unknown parameters, and the estimation of these parameters correspond to the use of data to select model in the variable selection case. Also, there is a final choice of the size of the model to be made. We will give some hints below on possibilities for developing simple criteria for this, but from what is known up to now, cross validation seems to be the best available tool. This is also the method usually employed in chemometrical models.

Since the initial class of models in our approach is considerably larger than what we have in the variable selection case, one should expect to find better predictions with this approach. For the case of chemometric methods, this expectation is also confirmed by simulation studies (see, e.g., Frank and Friedman, 1993). Much work remains to be done in evaluating specific predictions, however.

In the next three sections we discuss model reduction in multiple regression models assuming multinormality, and then the reduced model is presented in Section 5. In Section 6 we look at parameter estimation in the reduced model. Section 7 considers the corresponding situation when there are several response variables, and in Section 8 we look at classification problems. In Section 9 we generalize the basic results to other distributions than the multinormal distribution, and discuss some consequences of the general results obtained here.

2 Reduction of regression models by choice of conditioning.

In this Section we will make the ideal assumption that $(\mathbf{x}', y)'$ has a multinormal distribution with zero expectation and joint covariance matrix

$$\begin{pmatrix} \Sigma_{xx} & \sigma_{xy} \\ \sigma'_{xy} & \sigma_{yy} \end{pmatrix}. \tag{2}$$

(In fact, the most central results below may be generalized to observations that are not multinormal; see Section 9.) We assume that our sample consists of n independent observations from this population, and we want to predict y_0 from $\mathbf{x}_0$, sampled from the same population, i.e., having the same joint distribution. This model will be called the *basic model.* All variables are assumed to have zero expectation in order to simplify notation; in practice this essentially means that we will do regression on centered variables. A model including expectations could have been used at the expense of a more cumbersome notation. If we condition the basic model upon all the x-variables in all the samples, we get a regression model of the form (1) with $\beta = \mathbf{\Sigma}_{xx}^{-1}\sigma_{xy}$ and $\sigma^2 = \sigma_{yy} - \sigma_{xy}'\mathbf{\Sigma}_{xx}^{-1}\sigma_{xy}$, but the basic model as it stands contains more information.

As in the introduction we assume the dimension p of $\mathbf{x}$ to be fairly large, so that the regression estimator from (1) will be nonexistent or unstable. A simple solution is then to pick out k variables, say the first k, and do regression upon them. Let $\beta = (\beta_1, \ldots, \beta_p)'$ and $\mathbf{x} = (x_1, \ldots, x_p)'$.

Lemma 1.

The regression model obtained from the basic multinormal model by conditioning upon all variables and then putting $\beta_{k+1} = \cdots = \beta_p = 0$ *has the same form as the model obtained by just conditioning upon* $\mathbf{Z} = (\mathbf{x}_1, \cdots, \mathbf{x}_k)$ *In the basic model. This form is* $\mathbf{y} = \mathbf{Z}\gamma + \tilde{\mathbf{e}}$ *with* $\tilde{\mathbf{e}} \sim \mathrm{N}(\mathbf{0}, \tilde{\sigma}^2\mathbf{I})$.

Proof.

Simple calculation shows that in each case we get a model for each unit of the form $y = \gamma'\mathbf{z} + \tilde{e}$, where $\mathbf{z} = (x_1, \cdots, x_k)'$ and $\tilde{e} \sim \mathrm{N}(0, \tilde{\sigma}^2)$. The relationship between the parameters here and the parameters in the original model is in general different for the two cases, but this doesn't matter if the new equation is to be used to develop predictors, say by least squares. An interesting possibility, which is related to what we do later, is to adjust the parameters of the restricted model so that they fit with the z-conditioned model. $\diamond$

Obviously the same result holds if some other set of regression variables than the first k is kept in the model. There exist many methods aiming at picking the optimal set of variables, i.e., the best subset regression model to use, but here we want to look at a considerably larger set of models for seeking one that is good for prediction purposes: Let $\mathbf{R}$ be a $p \times k$ matrix of full rank k, and consider the new variables $\mathbf{z} = \mathbf{R}'\mathbf{x}$, a general set of k linear combinations of the original regression variables. Note that subset selection is a special case of this, and that regression upon a $\mathbf{z}$ of this form also can be related to several well-known methods like principal component regression. As in Lemma 1, concentrating upon such a smaller-dimensional set of variables can either be interpreted in terms of a model reduction or in terms of a special choice of conditioning in the model. Let $\mathbf{U}$ be any $p \times (p - k)$ matrix such that $\mathbf{R}'\mathbf{U} = \mathbf{0}$ and such that $(\mathbf{R}\ \mathbf{U})$ has full rank p.

Lemma 2.

For the multinormal case, the regression model obtained by

(a) conditioning upon $\mathbf{X}$ *and formally assuming* $\mathbf{U}'\beta = \mathbf{0}$ *in the basic model, and the model obtained from the basic model by*

(b) conditioning upon only $\mathbf{z} = \mathbf{R}'\mathbf{x}$ *for each unit,*

have the same form $\mathbf{y} = \mathbf{Z}\gamma + \tilde{\mathbf{e}}$ *with* $\tilde{\mathbf{e}} \sim \mathrm{N}(\mathbf{0}, \tilde{\sigma}^2\mathbf{I})$ *and* $\mathbf{Z} = \mathbf{XR}$*.*

Proof.

Same as for Lemma 1.

$\diamond$

Assume now that $\mathbf{R}$ is fixed, and that regression is done under the restricted model formulated in Lemma 2, i.e.,

$$\hat{\beta} = \mathbf{R}(\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R})^{-1}\mathbf{R}'\mathbf{X}'\mathbf{y}. \tag{3}$$

Under the assumption that $\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R}$ has full rank (which it will almost surely if $k < n$) the expectation and covariance matrix of $\hat{\beta}$ are

$$\mathrm{E}(\hat{\beta}|\mathbf{X}) = \mathbf{R}(\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R})^{-1}\mathbf{R}'\mathbf{X}'\mathbf{X}\beta \tag{4}$$

$$\mathrm{V}(\hat{\beta}|\mathbf{X}) = \mathbf{R}(\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R})^{-1}\mathbf{R}'\sigma^2, \tag{5}$$

where $\sigma^2 = \sigma_{yy} - \sigma_{xy}'\mathbf{\Sigma}_{xx}^{-1}\sigma_{xy}$.

This evaluation is done under the basic model conditioned upon the full matrix $\mathbf{X}$, which is a common procedure in statistics. Both under the restricted model in Lemma 2(a) and under the conditioned model in Lemma 2(b) we will have $\beta = \mathbf{R}\gamma$ for some γ, and hence $\mathrm{E}(\hat{\beta}|\mathbf{X}) = \beta$, respectively $\mathrm{E}(\hat{\beta}|\mathbf{XR}) = \beta = \mathbf{R}\gamma$. Under the restricted model there is no change in $\mathrm{V}(\hat{\beta}|\mathbf{X})$; under the conditioned model of Lemma 2(b) we get the same formula, but with σ^2 replaced by $\tilde{\sigma}^2 = \sigma_{yy} - \sigma_{xy}'\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\sigma_{xy}$, which in general is larger than or equal to σ^2.

Now to the question of how the matrix $\mathbf{R}$ can be chosen in the best possible way when the purpose is to get good predictions. As can be expected, the optimal choice will depend upon the parameters of the model, but we will not be too concerned about this problem now. In the next Sections we will look at other conditions while having in mind the fact that parameters need to be estimated, and after that we will turn to the estimation problem itself. For now we will just formulate the following simple condition, in practice to be looked upon as an unachievable ideal goal:

Condition 1: $\beta \in \mathrm{span}(\mathbf{R})$.

Here $\mathrm{span}(\mathbf{R})$ means the $k-$dimensional space spanned by the columns of $\mathbf{R}$. Since the mean square prediction error is uniquely determined by the expectation and variance of $\hat{\beta}$, it follows from the formulae (4)-(5) that conditions of interest must depend only upon this space, not on the whole matrix. It may be instructive to notice what Condition 1 means when $\mathbf{R}$ is a simple variable selection matrix: It just means that all 'correct' variables have been selected: All variables x_i that have been left out, have $\beta_i = 0$.

The following results are not unexpected, but fundamental:

Theorem 1.
(a) If Condition 1 holds, then $\mathrm{E}(\hat{\beta}|\mathbf{X}) = \beta$.
(b) Assuming that $\mathbf{\Sigma}_{xx}$ *is invertible, we have that* $\tilde{\sigma}^2 = \sigma^2$ *if and only if Condition 1 holds.*
(c) Assuming that $\mathbf{\Sigma}_{xx}$ *is invertible,* $\tilde{\beta} = \mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\sigma_{xy}$ *is equal to* $\beta = \mathbf{\Sigma}_{xx}^{-1}\sigma_{xy}$ *if and only if Condition 1 holds.*

Proof.
(a) We have already noted that $\beta = \mathbf{R}\gamma$ implies unbiasedness.
(b) We use the well-known general identity

$$\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}' = \mathbf{\Sigma}_{xx}^{-1} - \mathbf{\Sigma}_{xx}^{-1}\mathbf{U}(\mathbf{U}'\mathbf{\Sigma}_{xx}^{-1}\mathbf{U})^{-1}\mathbf{U}'\mathbf{\Sigma}_{xx}^{-1}, \tag{6}$$

(which can be proved by multiplying both sides by $\mathbf{U}$ and by $\mathbf{\Sigma}_{xx}\mathbf{R}$ and noting that these two matrices combine to a matrix of full rank.) Multiplying this identity by σ_{xy}' from the left and by σ_{xy} from the right, we see that the formulae for σ^2 and for $\tilde{\sigma}^2$ give the same value if and only if $\sigma_{xy}'\mathbf{\Sigma}_{xx}^{-1}\mathbf{U} = \beta'\mathbf{U} = 0$, which is equivalent to $\beta \in \mathrm{span}(\mathbf{R})$, i.e., Condition 1.
(c) Similar, using (6).

$\diamond$

3 The mean square prediction error.

In the evaluation of predictors we will take as a point of departure the expected squared prediction error $PRE = \mathrm{E}(y_0 - \hat{\beta}'\mathbf{x}_0)^2$. An important question is under what conditioning this expectation should be evaluated. Our answer is related to the main object here, namely to develop a general theory of prediction which functions well for the whole target population: At least for $(\mathbf{x}_0, y_0)$ the expectation should be with respect to the unconditional basic model. Not quite so strong argument can be given for taking expectation over the distribution of the matrix $\mathbf{X}$ of explanatory variables in the calibration set, one argument is that this gives an explicit formula which is easy to discuss. Alternatively, we can assume that the sample size n is so large that $n^{-1}\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R}$ can be approximated by $\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R}$ in the last part of the proof below. (Note that the convergence here in any ordinary matrix norm will be much better than the convergence of $n^{-1}\mathbf{X}'\mathbf{X}$ against $\mathbf{\Sigma}_{xx}$ if the dimension k is much smaller than p.)

Theorem 2.

Let $\mathbf{R}$ be a fixed $p \times k$ matrix of full rank k and assume $k < n - 1$, and let the estimated regression vector be given by equation (3). Then

$$PRE = \tilde{\sigma}^2 \frac{n-1}{n-k-1}, \tag{7}$$

where $\tilde{\sigma}^2 = \sigma^2 + \beta'(\mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\mathbf{\Sigma}_{xx})\beta$.

Proof.

Condition upon $\mathbf{z} = \mathbf{R}'\mathbf{x}$. In this conditioned model the residual variance is $\tilde{\sigma}^2 = \sigma_{yy} - \sigma_{xy}'\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\sigma_{xy}$. Since $\sigma^2 = \sigma_{yy} - \sigma_{xy}'\mathbf{\Sigma}_{xx}^{-1}\sigma_{xy}$, and since $\sigma_{xy} = \mathbf{\Sigma}_{xx}\beta$, the formula in the Theorem for $\tilde{\sigma}^2$ follows.

In the same conditioned model the regression vector is $\gamma = (\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\sigma_{xy}$ with least squares estimator $\hat{\gamma} = (\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R})^{-1}\mathbf{R}'\mathbf{X}'\mathbf{y}$. Expanding the square in $PRE = \mathrm{E}(y_0 - \hat{\gamma}'\mathbf{z}_0)^2$ and conditioning upon $\hat{\gamma}$, we find

$$PRE|\hat{\gamma} = \sigma_{yy} - 2\hat{\gamma}'(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})\gamma + \hat{\gamma}'(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})\hat{\gamma}.$$

Taking the conditional expectation of this, given $\mathbf{XR}$, gives

$$\begin{aligned} &PRE|\mathbf{XR} \\ &= \sigma_{yy} - 2\sigma_{xy}'\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\sigma_{xy} + \mathrm{tr}[(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})(\gamma\gamma' + (\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R})^{-1}\tilde{\sigma}^2)] \\ &= \tilde{\sigma}^2(1 + \mathrm{tr}[(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})(\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R})^{-1}]). \end{aligned} \tag{8}$$

Taking the expectation over $\mathbf{XR}$ and using a wellknown result from multivariate analysis (Anderson 1984, lemma 7.7.1) then gives (7).

$\diamond$

Corollary 1.

The expected squared prediction error PRE is $\sigma^2\frac{n-1}{n-k-1}$ if and only if Condition 1 holds; in all other cases PRE is larger than this.

Thus to obtain good predictions one should first try to achieve a situation where we can be as confident as possible that Condition 1 is satisfied at least in some approximate sense, and at the same time we should try to keep the dimension k as small as possible if this can be done without a substantial increase in $\tilde{\sigma}^2$.

4 Stepwise model selection and a reduction condition.

The regression vector β is an unknown quantity, so in whatever way $\mathbf{R}$ is determined, from data or in other ways, it is impossible to guarantee that the ideal Condition 1 holds. This problem is extra accute since it is impossible to estimate β accurately when the number of variables is large. It is therefore important to look at the behaviour of the mean square prediction error also when $\beta \notin \mathrm{span}(\mathbf{R})$. In this Section we will find a new reasonable condition to impose upon $\mathbf{R}$, a condition which applies to situations where the model dimension is increased stepwise and which is not so sensitive to the value of β.

Thus we start with a simple model of dimension $k = 1$, and increase the dimension stepwise. At each step k we then have $\mathbf{R} = \mathbf{R}_k$ as a $p \times k$ matrix of rank k, and at the next step $\mathbf{R}_{k+1} = (\mathbf{R}_k \; \mathbf{d}_k)$ for some vector $\mathbf{d}_k$. Then in general $\tilde{\sigma}^2 = \tilde{\sigma}_k^2$ in (7) will decrease or stay constant, while the factor $\frac{n-1}{n-k-1} \approx 1 + \frac{k}{n}$ will increase. The typical net effect will be that $PRE = PRE_k$ will be a convex function of k with a certain minimum. The aim in the end is to get a low minimum of PRE_k, and to achieve this, it is important to have the initial decrease in $\tilde{\sigma}_k^2$ at each step as large as possible. For the discussion which follows, it is useful to have in mind a hypothetical plot of PRE_k as a function of k with the ideal curve corresponding to $\tilde{\sigma}_k^2 = \sigma^2$ at the bottom of the plot. (See Fig. 1). In the plot of PRE_k against k there are two fixed points: $PRE_0 = \sigma_{yy}$ and $PRE_p = \sigma^2 \cdot \frac{n-1}{n-p-1}$ if $p < n-1$. (If $p \geq n-1$ the righthand end of the curve will essentially tend to infinity.) To get a minimum which is as low as possible between these two points, it is essential that the decrease $\tilde{\sigma}_k^2 - \tilde{\sigma}_{k+1}^2$ is large for small k.

It turns out that the optimal condition of this kind again is sensitive to the value of β. However, by being satisfied with a decrease which is as large as possible or nearly so *under most circumstances*, we get a condition which, while still depending upon unknown parameters, involve parameters which are more easy to estimate accurately.

The following result gives the formula for the decrease in the variance and the mathematically optimal value for this.

Theorem 3.
(a) Let

$$\mathbf{Q}_k = \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xx}\mathbf{R}_k(\mathbf{R'}_k\mathbf{\Sigma}_{xx}\mathbf{R}_k)^{-1}\mathbf{R'}_k\mathbf{\Sigma}_{xx}.$$

The decrease in $\tilde{\sigma}^2$ when going from $\mathbf{R}_k$ to $\mathbf{R}_{k+1} = (\mathbf{R}_k \; \mathbf{d}_k)$ (assuming $\mathbf{d}_k \notin \mathrm{span}(\mathbf{R}_k)$) is always nonnegative and is given by

$$\tilde{\sigma}_k^2 - \tilde{\sigma}_{k+1}^2 = \frac{(\beta'\mathbf{Q}_k\mathbf{d}_k)^2}{\mathbf{d'}_k\mathbf{Q}_k\mathbf{d}_k}. \tag{9}$$

(b) The maximal decrease is

$$\beta'\mathbf{Q}_k\beta,$$

and this is achieved if and only if $\mathbf{d}_k = \text{const.}\cdot\beta$ *plus arbitrary values in* $\text{span}(\mathbf{R}_k)$. *It may be convenient to replace* β *here by* δ_k, *given by* $\beta = \mathbf{R}_k\gamma_k + \delta_k$ *with* $\mathbf{R}'_k\mathbf{\Sigma}_{xx}\delta_k = \mathbf{0}$.

(c) With this choice of $\mathbf{d}_k$, $\tilde{\sigma}^2_{k+1}$ *will always achieve its smallest possible value* $\sigma^2 = \sigma_{yy} - \beta'\mathbf{\Sigma}_{xx}\beta$, *so the expected mean square error at step* $k+1$ *is then as small as possible.*

Proof.

We have $\tilde{\sigma}^2_k = \sigma_{yy} - \beta'\mathbf{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\mathbf{\Sigma}_{xx}\beta$ with $\mathbf{R} = \mathbf{R}_k$, and $\tilde{\sigma}^2_{k+1}$ is given by the same formula with $\mathbf{R} = (\mathbf{R}_k\ \mathbf{d}_k)$. By a straightforward calculation (see Appendix 1) we find (9). Since $\mathbf{Q}_k\mathbf{R}_k = \mathbf{0}$, we can replace β by δ_k in (9), and we can without loss of generality let $\mathbf{d}_k$ be in the space orthogonal to $\text{span}(\mathbf{R}_k)$. Then by Schwarz' inequality we see that the term subtracted in (9) is maximized for $\mathbf{d}_k = \text{const.}\cdot\delta_k$, and the value of the difference is then $\delta'_k\mathbf{Q}_k\delta_k$.

For the last assertion, insert $\beta = \mathbf{R}_k\gamma_k + \delta_k$ into the formula for $\tilde{\sigma}^2_k$ in the beginning of this proof, use (9) with $\beta \to \delta_k = \mathbf{d}_k$, and the same formula for β in $\sigma^2 = \sigma_{yy} - \beta'\mathbf{\Sigma}_{xx}\beta$ to prove $\tilde{\sigma}^2_{k+1} = \sigma^2$.

$\diamond$

This Theorem holds also for $k = 0$ with $\tilde{\sigma}^2_0 = \sigma_{yy}$, and then it shows that the smallest possible $\tilde{\sigma}^2_1$ is just the smallest globally, namely $\sigma^2 = \sigma_{yy} - \beta'\mathbf{\Sigma}_{xx}\beta$, that is, we get full reduction in one step, and this is obtained by taking $\text{span}(\mathbf{R}_1) = \text{span}(\beta)$, i.e., a version of Condition 1 again.

The interesting case, however, is when Condition 1 does not hold exactly, and we at a later step k have some more or less arbitrary selection matrix $\mathbf{R} = \mathbf{R}_k$. What conditions should one then impose on $\mathbf{R}_{k+1}$ in the next step in order that the reduction in $\tilde{\sigma}^2_k$ should be as large as possible? One may expect that it is not very crucial to have $\text{span}(\mathbf{R}_{k+1})$ determined in an exactly optimal way as long as (i) one goes in a direction which lead to a definite reduction in $\tilde{\sigma}^2$, (ii) the determination of the direction depends on parameters which are easy to estimate relatively accurately. On this background we will normalize $\mathbf{d}_k$ by $\mathbf{d}'_k\mathbf{d}_k = 1$ and concentrate on the numerator of (9). This numerator then gets its largest value when $\mathbf{d}_k$ is chosen along $\mathbf{Q}_k\beta$, a vector which has the form $\sigma_{xy} - \alpha$, where α belongs to $\text{span}(\mathbf{\Sigma}_{xx}\mathbf{R}_k)$. This leads to introducing a new condition:

Condition P: $\text{span}(\mathbf{R}_{k+1}) \subseteq \text{span}(\sigma_{xy}, \mathbf{\Sigma}_{xx}\mathbf{R}_k)$.

If we disregard complications that may rise from a possible increase of the denominator in (9), it turns out that a choice of $\text{span}(\mathbf{R}_{k+1})$ which does not satisfy

Condition P, can always be improved with respect to the decrease in the expected squared prediction error.

Theorem 4.

Assume that $\mathrm{span}(\mathbf{\Sigma}_{xx}\mathbf{R}_k) \not\subseteq \mathrm{span}(\mathbf{R}_k)$*. Suppose that we have found* $\mathbf{d}_k$ *and hence* $\mathbf{R}_{k+1} = (\mathbf{R}_k \ \mathbf{d}_k)$ *in such a way that Condition 2 does not hold. Assume further that*

$$\mathbf{d}'_k\mathbf{Q}_k\mathbf{d}_k \geq \mathbf{d}'_k\mathbf{P}\mathbf{Q}_k\mathbf{P}\mathbf{d}_k, \tag{10}$$

where $\mathbf{P}$ *is the projection operator upon the space spanned by* σ_{xy} *and* $\mathbf{\Sigma}_{xx}\mathbf{R}_k$*. Then one can always find another vector* $\mathbf{d}^*_k$*, and hence* $\mathbf{R}^*_{k+1} = (\mathbf{R}_k \ \mathbf{d}^*_k)$ *such that*

(i) Condition P holds for $\mathbf{R}^*_{k+1}$.

(ii) We have

$$\frac{(\beta'\mathbf{Q}_k\mathbf{d}^*_k)^2}{\mathbf{d}'^*_k\mathbf{Q}_k\mathbf{d}^*_k} \geq \frac{(\beta'\mathbf{Q}_k\mathbf{d}_k)^2}{\mathbf{d}'_k\mathbf{Q}_k\mathbf{d}_k} \tag{11}$$

for all β*. The reduction in* $\tilde{\sigma}^2_k$ *and therefore in* PRE_k *is therefore larger than or equal to what was obtained by the unstarred space extension.*

Proof. From the formula for $\mathbf{Q}_k$ we have

$$\mathbf{Q}_k\beta = \mathbf{\Sigma}_{xx}\beta - \mathbf{\Sigma}_{xx}\mathbf{R}_k(\mathbf{R}'_k\mathbf{\Sigma}_{xx}\mathbf{R}_k)^{-1}\mathbf{R}'_k\mathbf{\Sigma}_{xx}\beta = \sigma_{xy} - \alpha,$$

where $\alpha \in \mathrm{span}(\mathbf{\Sigma}_{xx}\mathbf{R}_k)$. Let L be the space spanned by σ_{xy} and $\mathbf{\Sigma}_{xx}\mathbf{R}_k$, so that $\mathbf{Q}_k\beta \in L$. Without loss of generality we can assume that $\mathbf{d}_k$ is perpendicular to $\mathrm{span}(\mathbf{R}_k)$, and then by assumption it follows that $\mathbf{d}^*_k$, the projection of $\mathbf{d}_k$ onto L will be nonzero. With this $\mathbf{d}^*_k$ the numerators will be the same on both sides of (11), while the denominator on the left hand side will be less than or equal to the denominator on the righthand side.

$\diamond$

As already stated, Theorem 4 is valid also for $k = 0$, when Condition P implies $\mathbf{R}_1 \propto \sigma_{xy}$. Since the dimension of $\mathbf{R}_k$ is always assumed to be $p \times k$, we therefore get a unique solution at each step from this:

Condition P': $\mathrm{span}(\mathbf{R}_k) = \mathrm{span}(\sigma_{xy}, \mathbf{\Sigma}_{xx}\sigma_{xy}, \cdots, \mathbf{\Sigma}^{k-1}_{xx}\sigma_{xy})$.

Remarks.

1. The technical condition (10) is really needed; it is not in general true that $\mathbf{Q} - \mathbf{PQP}$ is nonnegative definite when $\mathbf{Q}$ is positive definite and $\mathbf{P}$ is a projection. This is the reason why Condition P leads to the solution σ_{xy} in the first step instead of the theoretically optimal choice β.

2. It is important that Condition P involves σ_{xy}, which is easily estimable, in contrast to β, which is difficult to estimate for large p.

3. If covariances are replaced by estimated covariances in the above formulation of Condition P', we get the partial least squares regression algorithm, as shown by Helland (1988). An early discussion of PLSR can be found in Wold et al. (1984); it is now used routinely throughout the chemometrical literature.

4. The choice made here in Condition P is not unique; another alternative to the hard-to estimate $\mathbf{d}_k = \beta$, which maximizes (9), is to maximize $\frac{(\beta'\mathbf{Q}_k\mathbf{d})^2}{(\mathbf{d}'\mathbf{Q}_k\mathbf{d})^{1-\gamma}}$ for some fixed γ. The sample version of this leads to continuum regression (Stone and Brooks, 1990), which Sundberg (1993) has shown is closely related to ridge regression. Yet another alternative is to let the $\mathbf{d}_k$ be eigenvectors of $\mathbf{\Sigma}_{xx}$. We will show later that this leads to a population model which is equivalent to the one resulting from Condition P.

5. When the condition $\text{span}(\mathbf{\Sigma}_{xx}\mathbf{R}_k) \nsubseteq \text{span}(\mathbf{R}_k)$ in Theorem 4 does not hold, we see from Condition P' that $\text{span}(\mathbf{R}_{k+1}) \subseteq \text{span}(\mathbf{R}_k)$. This case will de discussed in the next section.

5 The reduced model with k relevant components.

Via Condition P we have now arrived at a sequence of models which are 'nearly optimal' in terms of expected squared prediction error at each step, and which is formulated in terms of linear combinations of the original variables with coefficients that are easily estimable from data. This model is conditional with respect to just this rather peculiar set of linear combinations, however, and it this sense bears no relation to the original regression model, which was conditional upon all the x-variables. To find a connection, we must replace the conditioning by a restriction of the parameters, which is done by using Lemma 2.

Theorem 5.

(a) In the sense formulated in Lemma 2, the $\mathbf{R}_k'\mathbf{x}$*-conditioned regression model satisfying Condition P has the same form as the ordinary* $\mathbf{x}$*-conditioned regression model under the following equivalent sets of parameter restrictions:*

(i) $\mathbf{\Sigma}_{xx}^k\beta \in \text{span}(\beta, \mathbf{\Sigma}_{xx}\beta, \ldots, \mathbf{\Sigma}_{xx}^{k-1}\beta)$.

(ii) $\mathbf{\Sigma}_{xx}^k\sigma_{xy} \in \text{span}(\sigma_{xy}, \mathbf{\Sigma}_{xx}\sigma_{xy}, \ldots, \mathbf{\Sigma}_{xx}^{k-1}\sigma_{xy})$.

(iii) There exists a set of k eigenvectors of $\mathbf{\Sigma}_{xx}$ *such that β belongs to the space spanned by these eigenvectors.*

(b) If the condition (a)(i), (a)(ii) or (a)(iii) hold for dimension k, but not for $k' < k$, there is a unique invariant space $\text{span}(\mathbf{R})$ *under* $\mathbf{\Sigma}_{xx}$ *of dimension k containing*

β. The regression defined in connection to Lemma 2 then have regression variables $\mathbf{z} = \mathbf{R}'\mathbf{x}$.

(c) If the conditions in (a) are satisfied, we can let $\mathrm{span}(\mathbf{R})$ *be the linear space mentioned in (a)(i), equivalently the linear space mentioned in (a)(ii) or equivalently the space spanned by the eigenvectors in (a)(iii).*

Proof.

Let $\mathbf{U}$ be a matrix of full column rank $p - k$ whose columns are orthogonal to $\mathrm{span}(\mathbf{R}_k) = \mathrm{span}(\sigma_{xy}, \mathbf{\Sigma}_{xx}\sigma_{xy}, \ldots, \mathbf{\Sigma}_{xx}^{k-1}\sigma_{xy})$. Then $\mathbf{U}'\beta = 0$ is equivalent to $\beta = \mathbf{\Sigma}_{xx}^{-1}\sigma_{xy} \in \mathrm{span}(\mathbf{R}_k)$. Multiplying the resulting linear relation by $\mathbf{\Sigma}_{xx}$, we see that this is equivalent to the statement that condition (a)(ii) holds either for this k or some lower values of k. The equivalence between (a)(i) and (a)(ii) is easy to see by a similar reasoning, and the equivalence to (a)(iii) was proved in Helland (1990). The rest of the Theorem follows from Helland (1990) and Næs and Helland (1993), where also other equivalent sets of conditions are given.

$\diamond$

One way to state these conditions together is that $\mathrm{span}(\mathbf{R}_k)$ *should be a k-dimensional invariant space under $\mathbf{\Sigma}_{xx}$ containing β*:

Condition 1: $\beta \in \mathrm{span}(\mathbf{R}_k)$
Condition 2: $\mathrm{span}(\mathbf{\Sigma}_{xx}\mathbf{R}_k) = \mathrm{span}(\mathbf{R}_k)$.

It is important for understanding that β now will be a parameter vector in the reduced model, in general not equal to the true regression vector of the original model.

As pointed out by von Rosen (1994), there is a substancial mathematical theory of invariant spaces, and the concept has also been used to some extent in linear model theory (see, e.g., Kruskal, 1968). It is obvious that there always is an invariant space containing β of dimension p, namely the whole space R^p. Theorem 5(a) expresses equivalent conditions implying that there exist invariant spaces (containing β) of smaller dimensions. In all cases $k < p$ this imposes restrictions upon the parameter space. The restrictions imposed are naturally nested in k.

Theorem 5 was arrived at by taking Condition P as a point of departure, a condition that aims at making (9) large for small k. Another possible point of departure is to let the coloumns of $\mathbf{R}_k$ be eigenvectors of $\mathbf{\Sigma}_{xx}$: $\mathbf{R}_k = (\mathbf{v}_1, \ldots, \mathbf{v}_k)$. Inserting this successively into (9), we find $\tilde{\sigma}_k^2 - \tilde{\sigma}_{k+1}^2 = \lambda_k(\beta'\mathbf{v}_k)^2$, where λ_k is the eigenvalue corresponding to $\mathbf{v}_k$. Again this reduction should be taken as large as possible for small k, but since β is hard to estimate, it is difficult to say what succession of eigenvectors this leads to. In any case Theorem 5 shows that this leads to exactly the same reduced model as Condition P.

The essential results of this section will also hold if $\mathbf{\Sigma}_{xx}$ does not have full rank, and the results (except for the reference to Lemma 2) do not depend on multinormality. The question of nonnormality will be taken up in a broader setting in Section 9.

6 Estimation of the matrix R.

Theorem 5 gives formulae for the desired invariant space of dimension k and conditions for its existence, everything expressed in terms of the unknown model parameters. For the use of the present ideas in practice, this space must be determined in some approximate way, either by using data or by other means. Heuristically one should expect that it is not too important to determine this space very accurately: If Condition 1 is satisfied exactly (with β as the true regression parameter), we do not need further conditions; if it only holds approximately, then (exact or approximate) validity of Condition 2 will help minimizing the effect of this. If Condition 2 is only approximately satisfied, the result is only small additional terms in the expected squared prediction error PRE. A crucial point is the determination of the dimension k. If it is too large, Theorem 2 shows that PRE will increase; if it is too small, the requirement needed to get an approximate invariant space of dimension k may be too severe.

To begin with, we will fix k and look at the determination of span($\mathbf{R}$). An obvious solution is to use the space in Theorem 5 (a)(ii) with covariances replaced by estimated covariances, i.e., take $\hat{\mathbf{R}} = (\mathbf{s}_{xy}, \mathbf{S}_{xx}\mathbf{s}_{xy}, \ldots, \mathbf{S}_{xx}^{k-1}\mathbf{s}_{xy})$ where $\mathbf{S}_{xx} = n^{-1}\mathbf{X}'\mathbf{X}$ and $\mathbf{s}_{xy} = n^{-1}\mathbf{X}'\mathbf{y}$. As shown in Helland (1990), this is equivalent to the well-known, but still somewhat controversial partial least squares method PLSR proposed by chemometricians. Numerous publications, in particular in the two chemometrical journals have shown that the method functions reasonably well, but critique has been raised by statisticians, for instance in Frank and Friedman (1993). From the present point of view a critical remark against PLSR is the following: The assumption that there exists an invariant space containing β of dimension k implies a restriction of the model parameters, as expressed by Theorem 5 (a)(ii). The resulting parameter estimates resulting from the above PLSR-formula for $\hat{\mathbf{R}}$ does not satisfy the corresponding restriction (with probability 1) when $k < p$. This may imply a loss in efficiency.

Nevertheless, asymptotic developments in Helland and Almøy (1994) and simulations in Almøy (1996) seem to indicate that PLSR functions well under the restricted model when k is small. Less is known in precise terms about its prediction ability under the original basic model as formulated above.

Principal component regression is another much used and reasonably well functioning method. In our setting it is given by the estimates $\hat{\mathbf{R}}$ connected to the criterion in Theorem 5 (a)(iii). One wellknown problem with the method is that there are several ways to determine which eigenvectors of $\mathbf{S}_{xx}$ to include in $\hat{\mathbf{R}}$, the two most common solutions being those with the largest eigenvalues or those with the largest values of a t-statistics connected to y.

A full discussion of various estimation procedures is beyond the scope of the present paper. We will limit ourself to listing some mathematical results that may be of relevance in this connection. Most of the results are relatively easy to prove, though some proofs, while based on relatively simple ideas, require more details. We will assume a sequence of data sets $(\mathbf{X}, \mathbf{y})$ (size n) with $\hat{\mathbf{R}}_{(n)}$ being a sequence of estimators of $\mathbf{R}$ (all of the same dimension $p \times k$), and we take $\hat{\beta}_n = \hat{\mathbf{R}}_{(n)}(\hat{\mathbf{R}}'_{(n)}\mathbf{X}'\mathbf{X}\hat{\mathbf{R}}_{(n)})^{-1}\hat{\mathbf{R}}'_{(n)}\mathbf{X}'\mathbf{y}$. For a vector $\mathbf{v}$ and a vector space L, let $d(\mathbf{v}, L)$ be the distance from $\mathbf{v}$ to L, i.e., the length $\|\mathbf{v} - \mathrm{P}_L\mathbf{v}\|$, where $\mathrm{P}_L\mathbf{v}$ is the projection of $\mathbf{v}$ on L. Finally, the expected squared prediction error is $PRE_{(n)} = \mathrm{E}(y_0 - \hat{\beta}'_n\mathbf{x}_0)^2$.

1. If $\hat{\mathbf{R}}_{(n)}$ is determined from an independent training set or if $\hat{\mathbf{R}}_{(n)}$ converges in probability to a constant matrix $\mathbf{R}$, then the expected squared prediction error $PRE_{(n)}$ converges to σ^2 if and only if

$$d(\beta, \mathrm{span}(\hat{\mathbf{R}}_n)) \xrightarrow{\mathrm{P}} 0. \tag{12}$$

If this condition does not hold, every subsequence limit of $PRE_{(n)}$ will be larger or equal to σ^2, with strict inequality for at least one subsequence limit.

2. If Condition 1 holds for the limiting matrix $\mathbf{R}$ so that $\beta = \mathbf{R}\gamma$ for some γ, then, as in Helland and Almøy (1994), as $n \to \infty$

$$PRE_{(n)} = \sigma^2(1 + \frac{k}{n}) + \frac{1}{n}\gamma'\mathrm{E}(\mathbf{W}'\mathbf{Q}\mathbf{W})\gamma + o(\frac{1}{n}), \tag{13}$$

where $\mathbf{Q} = \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\mathbf{\Sigma}_{xx}$ and $\mathbf{W}$ is the limit in distribution of $\sqrt{n}(\hat{\mathbf{R}}_{(n)} - \mathbf{R})$.

3. As a possible alternative to crossvalidation for determining model dimension, one might hope to draw upon the estimated residual variance

$$\hat{\tilde{\sigma}}^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\beta}_n)'(\mathbf{y} - \mathbf{X}\hat{\beta}_n), \tag{14}$$

where $\hat{\beta}_n = \hat{\mathbf{R}}_n(\hat{\mathbf{R}}'_n\mathbf{X}'\mathbf{X}\hat{\mathbf{R}}_n)^{-1}\hat{\mathbf{R}}'_n\mathbf{X}'\mathbf{y}$. It is easy to prove that this estimator is asymptotically unbiased in the sense that $\mathrm{E}(\hat{\tilde{\sigma}}^2) \to \tilde{\sigma}^2$ as $n \to \infty$ with $\tilde{\sigma}^2 = \sigma^2 + \beta'\mathbf{Q}\beta$.

A serious difficulty though is that the rate of convergence of $\mathrm{E}(\hat{\tilde{\sigma}}^2)$ towards $\tilde{\sigma}^2$, although of order $O(1/n)$ will in general depend upon the dimension k. To get further, the next term in the expansion should be made explicit, so that this effect may be corrected for. This issue will not be pursued here. Note that for model choice situations where criteria like Mallow's C_p and the Akaike criterion are used, no estimation of a subspace is needed, so this problem does not occur.

4. One obvious candidate for an estimator $\hat{\mathbf{R}}_{(n)}$ is the maximum likelihood estimator under the restricted model, which is developed in Helland (1992). Unfortunately, this estimator requires heavy computation, and the empirical results in a prediction setting are not too convincing (Almøy, 1996). An alternative candidate is the conditional maximum likelihood estimator, where one conditions upon $\mathbf{XR}$ in the restricted model. Using essentially the same computations as in Helland (1992), we get that this estimator is found by minimizing

$$(s_y^2 - \mathbf{s}'_{xy}\mathbf{R}(\mathbf{R}'\mathbf{S}_{xx}\mathbf{R})^{-1}\mathbf{R}'\mathbf{s}_{xy})|\mathbf{R}'\mathbf{S}_{xx}^{-1}\mathbf{R}|, \tag{15}$$

Note that the product of two factors is to be minimized in (15); the first factor is minimized if and only if $\hat{\beta}_{LS} = \mathbf{S}_{xx}^{-1}\mathbf{s}_{xy} \in \text{span}(\hat{\mathbf{R}})$ and the second factor is minimized iff $\hat{\mathbf{R}}$ is spanned by the k eigenvectors of $\mathbf{S}_{xx}$ with the largest eigenvalues. In this way the resulting predictor will have a relation both to ordinary regression and the most common form of principal component regression.

As in Helland (1992) the minimization here can be done stepwise, first in one dimension and then by successively increasing the dimension, if we impose the constraint that the resulting estimated subspaces should be nested within each other. The minimization for each dimension can also be done for instance as in Helland (1992).

A simpler solution is achieved by assuming that $\mathbf{\Sigma}_{xx} = \mathbf{\Sigma}_0$ is known, when we maximize the likelihood by minimizing

$$(s_y^2 - \sum_{j=1}^{k} \frac{(\mathbf{s}'_{xy}\mathbf{e}_{i_j})^2}{\lambda_{i_j}}) \prod_{j=1}^{k} \lambda_{i_j}^{-1} \tag{16}$$

with an obvious notation for eigenvectors and eigenvalues. One should expect this to be a reasonable solution also when $\mathbf{\Sigma}_{xx}$ is unknown and eigenvectors from $\mathbf{S}_{xx}$ are used. A further simplification is to approximate the first paranthesis in (16) by a product, so that no minimization over subsets is needed. Unfortunately, simulations indicate that the corresponding predictor seems to behave roughly like the principal component predictor based upon selecting components by a t-test (T. Almøy, private communication), a predictor which in most cases is known to be inferior to the ordinary principal component predictor.

5. Other estimators of $\mathbf{R}$ can be found by taking into account invariance properties of the model. This is presently being investigated.

7 Model reduction in regression with several response variables.

A multivariate extension of (1) is

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{17}$$

where $\mathbf{Y}$ is $n \times q$, $\mathbf{B}$ is an $p \times q$ parameter vector, and where the rows of $\mathbf{E}$ are independent multivariate normal $(\mathbf{0}, \boldsymbol{\Omega})$. Again we also have interest in the marginal $\mathbf{x}$-distribution, which is assumed multinormal $(\mathbf{0}, \boldsymbol{\Sigma}_{xx})$. (For most purposes, multinormality may be dispensed with; see Section 9.) If then (17) is looked upon as representing n independent conditional distributions, the joint distribution will also be multinormal, and $\mathbf{B} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$.

Take $(\mathbf{x}_0, \mathbf{y}_0)$ from the same population, so that $\mathbf{y}_0 = \mathbf{B}'\mathbf{x}_0 + \mathbf{e}_0$ with $\mathbf{e}_0 \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$. For an estimator $\hat{\mathbf{B}}$ one gets a predictor $\hat{\mathbf{B}}'\mathbf{x}_0$, and when evaluating this, it is natural to weight the dependent variables by the inverse error covariance matrix. This gives

$$\begin{aligned} PRE &= \mathrm{E}((\mathbf{y}_0 - \hat{\mathbf{B}}'\mathbf{x}_0)'\boldsymbol{\Omega}^{-1}(\mathbf{y}_0 - \hat{\mathbf{B}}'\mathbf{x}_0)) \\ &= \mathrm{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{yy}) - 2\mathrm{E}(\mathrm{tr}(\boldsymbol{\Omega}^{-1}\hat{\mathbf{B}}'\boldsymbol{\Sigma}_{xy})) + \mathrm{E}(\mathrm{tr}(\boldsymbol{\Omega}^{-1}\hat{\mathbf{B}}'\boldsymbol{\Sigma}_{xx}\hat{\mathbf{B}})), \end{aligned} \tag{18}$$

where $\boldsymbol{\Sigma}_{yy} = \mathbf{B}'\boldsymbol{\Sigma}_{xx}\mathbf{B} + \boldsymbol{\Omega}$.

Consider now some fixed $p \times k$ matrix $\mathbf{R}$ of full rank $k < p$ and the estimator

$$\hat{\mathbf{B}} = \mathbf{R}(\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R})^{-1}\mathbf{R}'\mathbf{X}'\mathbf{Y}.$$

Using this for prediction gives a similar predictor for each y-variable as used in Section 2, but with the same $\mathbf{R}$ for each variable. Taking conditional expectation of $\mathbf{Y}$, given $\mathbf{X}$ we get

$$\mathrm{E}(\hat{\mathbf{B}}|\mathbf{X}) = \mathbf{R}(\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R})^{-1}\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{B}$$

In particular, $\mathrm{E}(\hat{\mathbf{B}}|\mathbf{X}) = \mathbf{B}$ if

Condition 1. $\mathrm{span}(\mathbf{B}) \subseteq \mathrm{span}(\mathbf{R})$.

As in the case with one response variable, this condition also minimizes PRE. To look further upon the expected squared prediction error when the condition does not hold, we will use the approximation $n^{-1}\mathbf{R}'\mathbf{X}'\mathbf{X}\mathbf{R} = \mathbf{R}'\mathbf{S}_{xx}\mathbf{R} \sim \mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{R}$, and we will condition upon $\mathbf{X}\mathbf{R}$, as in the proof of Theorem 2. Then $\mathrm{E}(\hat{\mathbf{B}}|\mathbf{X}\mathbf{R}) = \mathbf{R}(\mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\boldsymbol{\Sigma}_{xy}$. Using (17) in the formula for $\hat{\mathbf{B}}$

$$\begin{aligned} &\mathrm{E}(\hat{\mathbf{B}}'\boldsymbol{\Sigma}_{xx}\hat{\mathbf{B}}|\mathbf{X}\mathbf{R}) \\ &\sim \mathbf{B}'\boldsymbol{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{B} + n^{-2}\mathrm{E}[\mathbf{E}'\mathbf{X}\mathbf{R}(\mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\mathbf{X}'\mathbf{E}|\mathbf{X}\mathbf{R}]. \end{aligned} \tag{19}$$

To insert the expressions above into (18) we need that the rows $\mathbf{e}_i'$ of $\mathbf{E}$ satisfy

$$\mathrm{V}(\mathbf{e}_i|\mathbf{XR}) = \tilde{\mathbf{\Omega}} = \mathbf{\Omega} + \mathbf{B}'(\mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\mathbf{\Sigma}_{xx})\mathbf{B}.$$

Then $\mathbf{f}_i = \mathbf{\Omega}^{-\frac{1}{2}}\mathbf{e}_i$ are independent with covariance matrix $\mathbf{\Omega}^{-\frac{1}{2}}\tilde{\mathbf{\Omega}}\mathbf{\Omega}^{-\frac{1}{2}}$, and writing out in terms of the $\mathbf{f}_i$'s, we find that

$$\mathrm{E}(\mathbf{E}\mathbf{\Omega}^{-1}\mathbf{E}'|\mathbf{XR}) = \mathbf{I}\cdot\mathrm{tr}(\mathbf{\Omega}^{-1}\tilde{\mathbf{\Omega}}).$$

So taking first conditional and then unconditional expectation in equation (18) gives

$$PRE \sim (q + \mathrm{tr}(\mathbf{\Omega}^{-1}\mathbf{B}'(\mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\mathbf{\Sigma}_{xx})\mathbf{B}))(1 + \frac{k}{n}). \tag{20}$$

This is the multivariate generalization of formula (7). Again it is natural to determine the space $\mathrm{span}(\mathbf{R})$ successively: $\mathbf{R}_{k+1} = (\mathbf{R}_k\ \mathbf{d}_k)$. Very much of the previous development is exactly as before. Let $\mathbf{Q}_k = \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xx}\mathbf{R}_k(\mathbf{R}_k'\mathbf{\Sigma}_{xx}\mathbf{R}_k)^{-1}\mathbf{R}_k{}'\mathbf{\Sigma}_{xx}$ as before, and let α_k be the first factor on the righthand side of formula (20) when $\mathbf{R} = \mathbf{R}_k$ is inserted. The generalization of formula (9) is then

$$\alpha_k - \alpha_{k+1} = \frac{\mathbf{d}_k'\mathbf{Q}_k\mathbf{B}\mathbf{\Omega}^{-1}\mathbf{B}'\mathbf{Q}_k\mathbf{d}_k}{\mathbf{d}_k'\mathbf{Q}_k\mathbf{d}_k}. \tag{21}$$

(See Appendix 1.) For the formulation of the following Theorem we refer to the population PLS2 algorithm defined and discussed briefly in Appendix 2.

Theorem 6.
Assume $\mathbf{\Omega} = \mathbf{I}$.

(a) The population PLS2 algorithm minimizes at each step the numerator of (21) if $\mathbf{d}_k'\mathbf{d}_k = 1$. *This algorithm terminates at* $\mathbf{R} = \mathbf{R}_k$ *if and only if Condition 1 then holds.*

(b) If this algorithm terminates at step k, then also

$$\textbf{Condition 2}: \ \mathrm{span}(\mathbf{\Sigma}_{xx}\mathbf{R}_k) = \mathrm{span}(\mathbf{R}_k),$$

and we have that $\mathbf{R}_k$ *spans a minimal space satisfying both Condition 1 and Condition 2.*

(c) Alternatively, this space can be characterized as the smallest space spanned by k eigenvectors of $\mathbf{\Sigma}_{xx}$ *which contains* $\mathrm{span}(\mathbf{B})$.

(d) The parameter restrictions formulated in (b) or (c) above constitute the restrictions generalizing those of Theorem 5 to the multivariate case.

Proof.

(a) In the notation of Appendix 2, the numerator of (21) is maximized if and only if $\mathbf{d}_k = \mathbf{w}_{k+1}$ is an eigenvector of $\mathbf{Q}_k\mathbf{B}\mathbf{B}'\mathbf{Q}_k = \mathbf{\Sigma}_{xy}^{(k+1)}\mathbf{\Sigma}'^{(k+1)}_{xy}$ with maximal eigenvalue, and this is just the way the algorithm is defined in the appendix. The algorithm terminates at step k iff $\mathbf{\Sigma}_{xy}^{(k+1)} = \mathbf{0}$, which in the notation $\mathbf{D}_i = \mathbf{\Sigma}_{xx}^{-\frac{1}{2}}\mathbf{\Sigma}_{xy}^{(i)}$ and $\mathbf{S}_i = \mathbf{\Sigma}_{xx}^{\frac{1}{2}}\mathbf{R}_i$ means $\mathbf{D}_{k+1} = (\mathbf{I} - \mathbf{S}_k(\mathbf{S}_k'\mathbf{S}_k)^{-1}\mathbf{S}_k')\mathbf{D}_1 = \mathbf{0}$. This is equivalent to $\text{span}(\mathbf{D}_1) \subseteq \text{span}(\mathbf{S}_k)$, hence $\text{span}(\mathbf{B}) \subseteq \text{span}(\mathbf{R}_k)$.

(b) Since $\mathbf{Q}_k\mathbf{B} = \mathbf{\Sigma}_{xy} - \mathbf{\Sigma}_{xx}\mathbf{R}_k(\mathbf{R}_k'\mathbf{\Sigma}_{xx}\mathbf{R}_k)^{-1}\mathbf{R}_k'\mathbf{\Sigma}_{xx}$ determines the next vector $\mathbf{w}_{k+1} = \mathbf{d}_k$ in $\mathbf{R}_{k+1} = (\mathbf{R}_k\ \mathbf{d}_k)$, it is clear that

$$\textbf{Condition P}: \ \text{span}(\mathbf{R}_{k+1}) \subseteq \text{span}(\mathbf{\Sigma}_{xy}, \mathbf{\Sigma}_{xx}\mathbf{R}_k).$$

When the algorithm stops at step k, then $\mathbf{R}_{k+1}$ may be replaced by $\mathbf{R}_k$ here. Since we know that $\mathbf{\Sigma}_{xy} = \mathbf{\Sigma}_{xx}\mathbf{B}$ with $\text{span}(\mathbf{B}) \subseteq \text{span}(\mathbf{R}_k)$, it follows that $\text{span}(\mathbf{R}_k) \subseteq \mathbf{\Sigma}_{xx}\mathbf{R}_k$. Since the matrices on both sides have the same rank, equality follows. From the proof in (a) it follows that the space $\text{span}(\mathbf{R}_k)$ is minimal with the property of containing $\text{span}(\mathbf{B})$ among all nontrivial sequences of spaces satisfying Condition P at each step.

(c) A space of dimension k satisfies Condition 2 if and only if it is spanned by k eigenvectors of $\mathbf{\Sigma}_{xx}$.

(d) Lemma 2 can immediately be generalized to the multivariate case. The previous restriction $\mathbf{U}'\beta = \mathbf{0}$ now reads $\mathbf{U}'\mathbf{B} = \mathbf{0}$, where $\mathbf{U}$ is spanned by $p - k$ eigenvectors of $\mathbf{\Sigma}_{xx}$. This gives clearly the same restriction as in (c).

$\diamond$

So far for the population algorithm corresponding to a restricted population model. The natural sample estimate of $\mathbf{R}$ arising from this - again see Appendix 2 - then gives the PLS2-predictor from chemometry. There are some variants of these estimators/ predictors (see Holcomb et al. (1997) and references there), and at least some of them seem to perform poorly compared to other predictors proposed by the statistical community (Breiman and Friedman, 1996). There is probably scope both for improvements and for better comparisons, and one possible point of departure may be the present reduced model.

Note, however, that the above formulation already points at one weakness of the PLS2 algorithm. To get the relation sketched between the algorithm in its population form and the reduction in mean square error, we had to assume that the residual covariance matrix $\mathbf{\Omega}$ was the identity. The case where the residuals between different dependent variables are correlated, is one case where we expect to get some gain from a joint prediction. It seems likely that a modified algorithm, where first a preliminary estimate $\hat{\mathbf{\Omega}}$ is found, and then a maximization of $\mathbf{w}'\mathbf{X}'\mathbf{Y}\hat{\mathbf{\Omega}}^{-1}\mathbf{Y}'\mathbf{X}\mathbf{w}$ is done in each step instead of that of $\mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}$, will have the possibility of leading to better predictions.

Both in the chemometric literature and in the statistical literature (see the references above) there has been some discussion on when it pays to do single variable prediction and when it pays to include several y-variables simultaneously in the prediction. The present model formulation may throw some light upon this question. The multivariate prediction means parsimony in the sense that the same $\mathbf{R}$ is used for all variables. On the other hand, if the dimension k of the invariant space has to be increased much to ensure that all columns of $\mathbf{B}$ belong to this space, the net gain may be negative.

8 Classification.

The simplest classification method is linear discriminant analysis, where we assume p classification variables that are observed in each of 2 classes, $\mathbf{x}_1 \sim \mathrm{N}(\mu_1, \mathbf{\Sigma})$ in the first class and $\mathbf{x}_2 \sim \mathrm{N}(\mu_2, \mathbf{\Sigma})$ in the second class. Again the model parameters are estimated by a training set, say n observations from each of the two classes, and again the estimation is difficult if the number p of variables is large compared to n. In an interesting recent article Friedman (1997) argue that this problem is less in classification than in regression, but the problem may nevertheless be serious in many applications.

So assume that we reduce the number of variables to k by letting $\mathbf{R}$ be some fixed $p \times k$ matrix and taking $\mathbf{z}_1 = \mathbf{R}'\mathbf{x}_1$ and $\mathbf{z}_2 = \mathbf{R}'\mathbf{x}_2$. Then of course $\mathbf{z}_i \sim \mathrm{N}(\mathbf{R}'\mu_i, \mathbf{R}'\mathbf{\Sigma}\mathbf{R})$ $(i = 1, 2)$, and standard linear discriminant analysis can be done with the new variables $\mathbf{z}_i$.

Concentrate on the simple symmetric case with equal prior probability for the two classes and equal cost. Then (see for instance Ripley, 1996) the asymptotic probability of misclassification for the classification based upon the $\mathbf{z}_i$'s will be

$$\pi = 2\Phi(-\frac{1}{2}\delta), \tag{22}$$

where Φ is the cumulative standard normal distribution function, and where δ is positive with $\delta^2 = (\mu_1 - \mu_2)'\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}\mathbf{R})^{-1}\mathbf{R}'(\mu_1 - \mu_2)$. Again we are relatively confident with the formulae resulting from asymptotic calculation when the dimension of the variables involved is only k.

The probability of misclassification is small iff δ is not too small, and this is the objective for our conditions for the (theoretically) best possible choice of $\mathbf{R}$.

Condition 1. $\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) \in \mathrm{span}(\mathbf{R})$.

Theorem 7.

We have $\delta^2 \leq (\mu_1 - \mu_2)'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)$ *with equality if and only if Condition 1 holds. Hence this condition minimizes the probability of misclassification for given* k*, and this minimum is (asymptotically to the lowest order) the same for all values of* k*.*

Proof.

Multiplying the identity (6) from the left by $(\mu_1 - \mu_2)'$ and from the right by $\mu_1 - \mu_2$, we see that the minimum is reached for $\mathbf{U}'\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) = \mathbf{0}$, which is equivalent to Condition 1.

$\diamond$

A similar discussion as in the regression case involving stepwise model selection can be made. Instead we take a simpler approach showing a different property of the reduced model. We then introduce at once

Condition 2. $\mathrm{span}(\mathbf{\Sigma R}) = \mathrm{span}(\mathbf{R})$.

Theorem 8.

Assume a general $\mathbf{R}$ *is such that Condition 1 may or may not hold, and put* $\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) = \mathbf{R}\gamma + \epsilon$*, where* $\mathbf{R}'\epsilon = \mathbf{0}$*. Then either Condition 2 holds and* $\inf_\epsilon \delta^2 = \delta^2|_{\epsilon=\mathbf{0}} = \gamma'\mathbf{R}'\mathbf{\Sigma}\mathbf{R}\gamma$*, hence* $\sup_\epsilon \pi = \pi|_{\epsilon=\mathbf{0}}$ *for the probability of misclassification* π*, or Condition 2 does not hold and* $\sup_\epsilon \pi > \pi|_{\epsilon=\mathbf{0}}$.

Proof.

When $\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) = \mathbf{R}\gamma + \epsilon$ with $\mathbf{R}'\epsilon = \mathbf{0}$, we find

$$\delta^2 = \gamma'\mathbf{R}'\mathbf{\Sigma}\mathbf{R}\gamma + \epsilon'\mathbf{\Sigma}\mathbf{R}(\mathbf{R}'\mathbf{\Sigma}\mathbf{R})^{-1}\mathbf{R}'\mathbf{\Sigma}\epsilon + 2\gamma'\mathbf{R}'\mathbf{\Sigma}\epsilon.$$

The last two terms on the righthand side here depend upon ϵ. If Condition 2 holds, they vanish when $\mathbf{R}'\epsilon = \mathbf{0}$. Assume that Condition 2 does not hold. Then ϵ can be chosen such that the quadratic term in ϵ - call it Q - is positive. For such an ϵ replace ϵ by $t\epsilon$, where t is some scalar. Minimization over t leads to that the sum of these terms is negative when $t = -\gamma'\mathbf{R}'\mathbf{\Sigma}\epsilon/Q$.

$\diamond$

We hope to discuss estimators of $\mathbf{R}$ and corresponding classification procedures elsewhere.

Remark.

A similar discussion of classification error using a stepwise increase in dimension can be done in exactly the same way as in the regression case. The formula (22) is asymptotic, but this does not matter as long as it is used for selection matrices $\mathbf{R}$ with a small number k of columns. When k is large, sample variation will cause the classification error to be larger than what is given by (22), again parallell to the prediction error in the regression case. It is therefore again important to keep the number of linear combinations of variables on which classifications should be based, low. Again a restricted model, stating that only a small number of eigenvectors of $\mathbf{\Sigma}$ contribute in the cassification, seem to be useful.

On the other hand, the property used in Theorem 8 above, that Condition 2 in general implies an equivalence between the orthogonalities $\mathbf{R}'\delta = \mathbf{0}$ and $\mathbf{R}'\mathbf{\Sigma}\delta = \mathbf{0}$, has useful consequences in the regression case, also. (It can be used to derive alternative expressions for the expected squared prediction error.) Furthermore, a minimax property of prediction error - analoguous to that for classification error given in Theorem 8 - can be formulated for the regression case.

9 General theory and discussion.

For definiteness we will return to the situation of predicting a scalar variable y_0 from a vector $\mathbf{x}_0$ of variables, but generalizations to cover the situation in Section 8 should be fairly straightforward to make. We will also assume quadratic loss (and will assume that y has finite variance), so the general task is to minimize

$$PRE = \mathrm{E}(y_0 - \hat{y}_0)^2, \tag{23}$$

where $\hat{y}_0$ is a function of $\mathbf{x}_0$ and of a training set $\{(\mathbf{x}_i, y_i); i = 1, \ldots, n\}$ of independent observations, such that all $(\mathbf{x}_i, y_i); i = 0, 1, \ldots, n$ have the same, more or less unknown distribution.

The ordinary linear or nonlinear regression procedure is to go via the conditional expectation $\mathrm{E}(y_0|\mathbf{x}_0)$ and use as $\hat{y}_0$ an estimate $\hat{\mathrm{E}}(y_0|\mathbf{x}_0)$ of this from the training set. By adding and subtracting $\mathrm{E}(y_0|\mathbf{x}_0)$ to $y_0 - \hat{y}_0$ in the resulting equation (23), we find that it is equal to

$$PRE = \mathrm{E}[\mathrm{Var}(y_0|\mathbf{x}_0)] + \mathrm{E}[\mathrm{E}(y_0|\mathbf{x}_0) - \hat{\mathrm{E}}(y_0|\mathbf{x}_0)]^2. \tag{24}$$

The first term here is unrelated to the training set, so it is the second term that one should try to reduce in order to give good predictions. To get a feeling for this last term; in the linear regression case it is $\mathrm{E}((\hat{\beta} - \beta)'\mathbf{\Sigma}_{xx}(\hat{\beta} - \beta))$. The reduction of this term will be a problem if the number of parameters is large, which it generally will be if the number p of variables in $\mathbf{x}$ is large. So we may choose to reduce the model by conditioning on a smaller vector variable $\mathbf{z} = h(\mathbf{x})$ (say of dimension $k < p$) instead of on the whole vector $\mathbf{x}$. Note that in this general setting the analogue of

Lemma 1 and Lemma 2 do not necessarily hold, so choosing some smaller vector to condition on is in general something essentially different from equating some specific parameters in the model to zero.

An important question is how to find a sensible new vector $\mathbf{z}$ to condition upon. Theoretically, the first requirement to be satisfied is that the change from $\mathbf{x}$ to $\mathbf{z}$ should not increase substantially the first term in (24). It is easy to find a simple theoretical condition for no such increase to take place: Look at the version of the well-known identity $\mathrm{Var}(y) = \mathrm{E}[\mathrm{Var}(y|\mathbf{x})] + \mathrm{Var}[\mathrm{E}(y|\mathbf{x})]$, conditioned upon $\mathbf{z}$, and take the expectation of this to get

$$\mathrm{E}[\mathrm{Var}(y|\mathbf{z})] = \mathrm{E}[\mathrm{Var}(y|\mathbf{x})] + \mathrm{E}[\mathrm{Var}(\mathrm{E}(y|\mathbf{x})|\mathbf{z})], \tag{25}$$

so by conditioning upon $\mathbf{z}$ instead of upon $\mathbf{x}$ in (24) we find

$$\begin{array}{ll} PRE = \mathrm{E}[\mathrm{Var}(y|\mathbf{z})] & +\mathrm{E}[\mathrm{E}(y|\mathbf{z}) - \hat{\mathrm{E}}(y|\mathbf{z})]^2 \\ = \mathrm{E}[\mathrm{Var}(y|\mathbf{x})] + \mathrm{E}[\mathrm{Var}(\mathrm{E}(y|\mathbf{x})|\mathbf{z})] & +\mathrm{E}[\mathrm{E}(y|\mathbf{z}) - \hat{\mathrm{E}}(y|\mathbf{z})]^2 \end{array}. \tag{26}$$

Hence the term independent of estimation in (24) does not increase when going from $\mathbf{x}$ to $\mathbf{z}$ if and only if $\mathrm{Var}(\mathrm{E}(y|\mathbf{x})|\mathbf{z}) = 0$, i.e., if $\mathrm{E}(y|\mathbf{x})$ is a function of $\mathbf{z}$ (almost surely), which then necessarily must be $\mathrm{E}(y|\mathbf{z})$. This leads to

Condition 1. $\mathrm{E}(y|\mathbf{x}) = \mathrm{E}(y|\mathbf{z})$ (a.s).

Theorem 9.
(a) The first term in (24) is constant in going from $\mathbf{x}$ *to* $\mathbf{z}$ *if Condition 1 holds. In all other cases this term will increase.*

(b) For the case of a linear function $\mathbf{z} = \mathbf{R}'\mathbf{x}$ *and of multinormal observations (or more generally, assuming that all conditional expectations are linear; see below), the Condition 1 here is equivalent to the previous Condition 1.*

Proof. (a) is already proved. For (b), use Theorem 1, either (b) or (c) there; the generalization to other cases with linear conditional expectation is straightforward, using Theorem 10 below.

◇

To find further conditions we will limit ourselves to linear functions $\mathbf{z} = \mathbf{R}'\mathbf{x}$. Furthermore, we will assume that the conditional expectation of y, given $\mathbf{x}$ is linear in $\mathbf{x}$; i.e., $\mathrm{E}(y|\mathbf{x}) = \beta'\mathbf{x}$. Then, from (25), the increase in conditional variance of y when going from conditioning on $\mathbf{x}$ to conditioning on $\mathbf{z}$ and hence the increase in the non-estimative part of PRE, will be

$$\mathrm{E}[\mathrm{Var}(y|\mathbf{z})] - \mathrm{E}[\mathrm{Var}(y|\mathbf{x})] = \mathrm{E}[\mathrm{Var}(\beta'\mathbf{x}|\mathbf{R}'\mathbf{x})]. \tag{27}$$

In the multinormal case this expression is

$$\beta'(\boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\boldsymbol{\Sigma}_{xx})\beta$$

This was the basic result needed for the mean square error calculations of Section 2, and it is important, though perhaps surprising to some that no distributional assumptions at all (except for finite variance) is needed for a closely related result to be valid.

Theorem 10.

Assume that $\mathbf{x}$ *has a finite covariance matrix* $\boldsymbol{\Sigma}_{xx}$*. Then we have the following:*

(a) If the conditional expectation of $\mathbf{x}$*, given* $\mathbf{R}'\mathbf{x}$ *is linear:* $\mathrm{E}(\mathbf{x}|\mathbf{R}'\mathbf{x}) = \mathbf{A}\mathbf{R}'\mathbf{x}$*, then*

$$\mathrm{E}[\mathrm{Var}(\beta'\mathbf{x}|\mathbf{R}'\mathbf{x})] = \beta'\mathbf{Q}\beta,$$

where $\mathbf{Q} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\boldsymbol{\Sigma}_{xx}$*.*

(b) In general

$$\mathrm{E}[\mathrm{Var}(\beta'\mathbf{x}|\mathbf{R}'\mathbf{x})] = \beta'\tilde{\mathbf{Q}}\beta,$$

where $\tilde{\mathbf{Q}} = \mathbf{Q} - \mathbf{QMQ}$ *for a nonnegative definite matrix* $\mathbf{M}$ *which is such that* $\tilde{\mathbf{Q}}$ *is nonnegative definite. So the difference in variance, hence the approximate difference in mean square error when we disregard estimation error, is given by this expression.*

(c) If the conditional expectation of $\mathbf{x}$*, given* $\mathbf{R}'\mathbf{x}$ *is linear, then with* $\mathbf{z} = \mathbf{R}'\mathbf{x}$ *we have* $\mathrm{E}(y|\mathbf{z}) = \gamma'\mathbf{z}$*, where* $\gamma = (\mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\sigma_{xy}$*. Furthermore, if* $\mathrm{Var}(y|\mathbf{x}) = \sigma^2$*, we have* $\mathrm{Var}(y|\mathbf{z}) = \tilde{\sigma}^2 = \sigma^2 + \beta'\tilde{\mathbf{Q}}\beta$*. The model restriction* $\mathbf{U}'\beta = \mathbf{0}$*, where* $\mathbf{R}'\mathbf{U} = \mathbf{0}$ *and* $(\mathbf{R}\ \mathbf{U})$ *has full rank, leads to the same conditional expectation* $\mathrm{E}(y|\mathbf{x}) = \gamma'\mathbf{z}$*.*

Proof.

In case (a) we find by multiplying $\mathrm{E}(\mathbf{x}|\mathbf{R}'\mathbf{x}) = \mathbf{A}\mathbf{R}'\mathbf{x}$ from the right by $\mathbf{x}'\mathbf{R}$ and then taking expectation that $\mathbf{A} = \boldsymbol{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{R})^{-1}$.

In general, given $\mathbf{R}$, choose $\mathbf{U}$ such that $\mathbf{R}'\mathbf{U} = \mathbf{0}$ and such that $(\mathbf{R}\ \mathbf{U})$ has full rank p. Then from equation (6) we have $\mathbf{Q} = \mathbf{U}(\mathbf{U}'\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{U})^{-1}\mathbf{U}'$, and by the same equation

$$\beta'\mathbf{x} = \beta'\boldsymbol{\Sigma}_{xx}\mathbf{R}(\mathbf{R}'\boldsymbol{\Sigma}_{xx}\mathbf{R})^{-1}\mathbf{R}'\mathbf{x} + \beta'\mathbf{Q}\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{x}, \tag{28}$$

so

$$\begin{aligned} &\mathrm{E}(\mathrm{Var}(\beta'\mathbf{x}|\mathbf{R}'\mathbf{x})) = \mathrm{E}(\mathrm{Var}(\beta'\mathbf{Q}\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{x}|\mathbf{R}'\mathbf{x})) \\ &= \mathrm{E}(\beta'\mathbf{Q}\boldsymbol{\Sigma}_{xx}^{-1}[\mathrm{E}(\mathbf{x}\mathbf{x}'|\mathbf{R}'\mathbf{x}) - \mathrm{E}(\mathbf{x}|\mathbf{R}'\mathbf{x})\mathrm{E}(\mathbf{x}|\mathbf{R}'\mathbf{x})']\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{Q}\beta) \\ &= \beta'\mathbf{Q}\boldsymbol{\Sigma}_{xx}^{-1}\mathrm{E}(\mathbf{x}\mathbf{x}')\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{Q}\beta - \beta'\mathbf{QMQ}\beta \\ &= \beta'\mathbf{U}(\mathbf{U}'\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{U})^{-1}\mathbf{U}'\beta - \beta'\mathbf{QMQ}\beta = \beta'\mathbf{Q}\beta - \beta'\mathbf{QMQ}\beta, \end{aligned} \tag{29}$$

where $\mathbf{M} = \boldsymbol{\Sigma}_{xx}^{-1}\mathrm{E}[\mathrm{E}(\mathbf{x}|\mathbf{R}'\mathbf{x})\mathrm{E}(\mathbf{x}|\mathbf{R}'\mathbf{x})']\boldsymbol{\Sigma}_{xx}^{-1}$ and equation (6) has been used again.

In case (a) we find by the calculation above that $\mathbf{QM}$ vanishes; in general $\mathbf{M}$ is nonnegative definite. Nonnegative definiteness of $\tilde{\mathbf{Q}}$ follows since the expected variance calculated above must be nonnegative.

The proof of (c) is found by first noting that $\mathrm{E}(y|\mathbf{x}) = \beta'\mathbf{x}$, and then taking the conditional expectation of this given $\mathbf{z} = \mathbf{R}'\mathbf{x}$. The formula for $\mathrm{Var}(y|\mathbf{z})$ follows from the same result used to prove (25). The formula for $\mathrm{E}(y|\mathbf{x})$ under the restriction $\mathbf{U}'\beta = \mathbf{0}$ follows from equation (6).

$\diamond$

The consequence of this is that all essential results of the Sections 2-5 are valid if we assume that the conditional expectation of $\mathbf{x}$, given $\mathbf{R}'\mathbf{x}$ is linear. The class of distributions for which this is valid, includes the elliptical (or elliptically symmetric) distributions; see Devlin et al. (1976) and references there, in particular Kelker (1970). In addition, Theorems 3 and 4 hold with some - admittedly nontrivial - modifications essentially without any distributional assumptions at all on the variables, assuming only finite moments and $\mathrm{E}(y|\mathbf{x}) = \beta'\mathbf{x}$. The detailed proof of this will be omitted, but the main idea is that these proofs are directly based on the formula for the expected squared prediction error, which by Theorem 10(b) is asymptotically valid in general if we only replace the matrix $\mathbf{Q}$ by $\tilde{\mathbf{Q}}$, and on the fact that this matrix also has a vanishing product with $\mathbf{R}$ and that it is small when $\mathbf{Q}$ is small. As a consequence, the two basic conditions, Condition 1 and the invariance condition on span($\mathbf{R}$) on the reduced model are of some relevance for any linear model with random x's, whatever the distribution of these x's, and whatever the conditional distribution of y, given these x's. The discussion on estimation in Section 6 is also quite generally valid, but the maximum likelihood estimate of Section 7 is of course distribution dependent.

One way to put this, is that the chemometricians in some sense seem to have been on the right track when they have used the term 'soft models' in connection to their PLS regression. The general feeling among statisticians still is that chemometricians are imprecise in some of their terminology, but on the other hand it seems to be easier to develop new and fruitful ideas on an intuitive level than if full rigor is demanded at each step. Recent issues of the chemometrical journals contain ideas that go far beyond what has been discussed in this paper. (This has also recently reached statistical journals; see for instance Durand and Sabatier (1997) and references there.)

The way of thinking of statistical models that is promoted in this paper, points further than specific chemometric methods, however. We make explicit the fact that in the case with many unknown parameters it is useful to have at least two different statistical models under considerations at the same time: The 'correct' model that adequately describes reality in all details and the 'simplified', reduced model which is used for estimation of parameters and for statistical analysis in general. In this paper we introduce for the linear model case specific theoretical conditions to assure that the reduced model functions as well as possible for prediction purposes. In this way we get a nested sequence of reduced models, and the order can be found by cross-validation or in other ways.

The possible danger of overfitting of models that may result from this way of thinking, needs to be further analyzed. If the order of the model is found by cross-validation, one will probably be reasonably safe, but in the multinormal case there also seem to be some possibility of using estimated prediction error found from the ordinary regression mean square, a possibility that should be investigated further.

There may be some intuitive arguments to the effect that because irrelevant information is thrown away, the estimates from the reduced model may have certain robustness properties. Exact results in this direction may be difficult, but will be welcome.

Practical algorithms for computing estimates are not touched upon at all in this paper. It is well known from partial least squares regression that the best formulas or algorithms for theoretical understanding are usually not the best for numerical computations.

As a final point, once the ideas behind this paper are accepted, there seems to be potential for extending in several directions. Logistic regression and other loglinear models (with or without link functions) with many explanatory variables is an immediate possibility, likewise multivariable models with many parameters. An interesting challange is the situation where the parameters are not in linear form, but where one neverless perhaps may give a similar theory to the present theory if the parameters are related by some group symmetry.

Appendix 1. Proof of (9) and related formulae.

Let $\mathbf{S}_k = \sqrt{\mathbf{\Sigma}_{xx}}\mathbf{R}_k$, $\mathbf{c}_k = \sqrt{\mathbf{\Sigma}_{xx}}\mathbf{d}_k$ and $\mathbf{P}_k = \mathbf{S}_k(\mathbf{S}_k'\mathbf{S}_k)^{-1}\mathbf{S}_k'$. Then $(\mathbf{I}-\mathbf{P}_k)\mathbf{c}_k$ - if nonzero - is orthogonal to span$(\mathbf{S}_k)$, and therefore

$$\mathbf{P}_{k+1} - \mathbf{P}_k = \frac{(\mathbf{I}-\mathbf{P}_k)\mathbf{c}_k\mathbf{c}_k'(\mathbf{I}-\mathbf{P}_k)}{\mathbf{c}_k'(\mathbf{I}-\mathbf{P}_k)\mathbf{c}_k}.$$

Multiplying this equation from the left and from the right by $\sqrt{\mathbf{\Sigma}_{xx}}$ then gives

$$\mathbf{\Sigma}_{xx}\mathbf{R}_{k+1}(\mathbf{R}_{k+1}'\mathbf{\Sigma}_{xx}\mathbf{R}_{k+1})^{-1}\mathbf{R}_{k+1}'\mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xx}\mathbf{R}_k(\mathbf{R}_k'\mathbf{\Sigma}_{xx}\mathbf{R}_k)^{-1}\mathbf{R}_k'\mathbf{\Sigma}_{xx} = \frac{\mathbf{Q}_k\mathbf{d}_k\mathbf{d}_k'\mathbf{Q}_k}{\mathbf{d}_k'\mathbf{Q}_k\mathbf{d}_k},$$

from which the equations (9) and (21) readily follow. By the the special form of $\mathbf{Q}_k$ $(= \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xx}\mathbf{R}_k(\mathbf{R}_k'\mathbf{\Sigma}_{xx}\mathbf{R}_k)^{-1}\mathbf{R}_k')$, this result can also be written

$$\mathbf{I} - \mathbf{\Sigma}_{xx}\mathbf{R}_{k+1}(\mathbf{R}_{k+1}'\mathbf{\Sigma}_{xx}\mathbf{R}_{k+1})^{-1}\mathbf{R}_{k+1}' = (\mathbf{I} - \frac{\mathbf{Q}_k\mathbf{d}_k\mathbf{d}_k'}{\mathbf{d}_k'\mathbf{Q}_k\mathbf{d}_k})(\mathbf{I} - \mathbf{\Sigma}_{xx}\mathbf{R}_k(\mathbf{R}_k'\mathbf{\Sigma}_{xx}\mathbf{R}_k)^{-1}\mathbf{R}_k').$$

By iterating this, we get

$$\mathbf{Q}_k = (\prod_{i=0}^{k-1}(\mathbf{I} - \frac{\mathbf{Q}_i\mathbf{d}_i\mathbf{d}_i'}{\mathbf{d}_i'\mathbf{Q}_i\mathbf{d}_i}))\mathbf{\Sigma}_{xx}. \qquad (30)$$

Appendix 2. Population version of the PLS2 algorithm.

The ordinary PLS2 algorithm (see, e.g., Holcomb at al., 1997) is determined from centered data $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{Y}_1 = \mathbf{Y}$ by for $i = 1, \ldots$ letting $\mathbf{w}_i$ be the normed eigenvector of $\mathbf{X}_i\mathbf{Y}_i'\mathbf{Y}_i'\mathbf{X}_i$ with maximal eigenvalue, and then defining successively

$$\mathbf{X}_{i+1} = \mathbf{X}_i(\mathbf{I} - \frac{\mathbf{w}_i\mathbf{w}_i'\mathbf{X}_i'\mathbf{X}_i}{\mathbf{w}_i'\mathbf{X}_i'\mathbf{X}_i\mathbf{w}_i})$$

$$\mathbf{Y}_{i+1} = \mathbf{Y}_i - \mathbf{X}_i\frac{\mathbf{w}_i\mathbf{w}_i'\mathbf{X}_i'\mathbf{Y}_i}{\mathbf{w}_i'\mathbf{X}_i'\mathbf{X}_i\mathbf{w}_i}.$$

By replacing sample (co)variances by population (co)variances $\mathbf{\Sigma}_{xx}^{(i)} = \mathrm{E}(\mathbf{x}_i\mathbf{x}_i')$ and $\mathbf{\Sigma}_{xy}^{(i)} = \mathrm{E}(\mathbf{x}_i y_i)$, this leads to $\mathbf{\Sigma}_{xx}^{(1)} = \mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{xy}^{(1)} = \mathbf{\Sigma}_{xy}$ and

$$\mathbf{\Sigma}_{xx}^{(i+1)} = (\mathbf{I} - \frac{\mathbf{\Sigma}_{xx}^{(i)}\mathbf{w}_i\mathbf{w}_i'}{\mathbf{w}_i'\mathbf{\Sigma}_{xx}^{(i)}\mathbf{w}_i})\mathbf{\Sigma}_{xx}^{(i)}, \quad \mathbf{\Sigma}_{xy}^{(i+1)} = (\mathbf{I} - \frac{\mathbf{\Sigma}_{xx}^{(i)}\mathbf{w}_i\mathbf{w}_i'}{\mathbf{w}_i'\mathbf{\Sigma}_{xx}^{(i)}\mathbf{w}_i})\mathbf{\Sigma}_{xy}^{(i)}.$$

Then taking $\mathbf{R}_k = (\mathbf{w}_1, \ldots, \mathbf{w}_k)$, this leads by (30) to the following relations between the quantities defined here and those defined in Appendix 1 and in the main body (Sections 4 and 7) of the paper:

$$\mathbf{w}_i = \mathbf{d}_{i-1}, \ \mathbf{\Sigma}_{xx}^{(i)} = \mathbf{Q}_{i-1}, \ \mathbf{\Sigma}_{xy}^{(i)} = (\mathbf{I} - \mathbf{\Sigma}_{xx}\mathbf{R}_{i-1}(\mathbf{R}_{i-1}'\mathbf{\Sigma}_{xx}\mathbf{R}_{i-1})^{-1}\mathbf{R}_{i-1})\mathbf{\Sigma}_{xy}.$$

In particular, $\mathbf{\Sigma}_{xy}^{(i)} = \mathbf{Q}_{i-1}\mathbf{B}$; which shows that the eigenvector defined in the population version of the PLS2 algorithm as formulated above, is just the vector needed to maximize the numerator of the righthand side of (21) when $\mathbf{\Omega} = \mathbf{I}$.

References

Almøy, T. (1996) A simulation study on comparison of prediction methods when only a few components are relevant. *Comp. Statist. and Data Anal.* **21**, 87-107.

Anderson, T.W. (1984) *An introduction to Multivariate Statistical Analysis*, John Wiley, New York.

Breiman, L. and J.H. Friedman (1996) Predicting multivariate responses in multiple linear regression (with discussion). *J. R. Statist. Soc.* B **57**, 3-37.

Devlin, S.J., R. Gnanadesikan, and J.R. Kettenring (1976) Some multivariate applications of elliptical distributions. In: S. Ikeda et al. [Ed.] *Essays in Propbability and Statistics.* Shinko Tsusho, Tokyo.

Durand, J.-F. and R. Sabatier (1997) Additive splines for partial least squares regression. *J. Amer. Statist. Ass.***92**, 1546-1554.

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

Friedman, J.H. (1997) On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* **1**, 55-78.

Helland, I.S. (1987) On the interpretation and use of R^2 in regression analysis. *Biometrics* **43**, 61-69.

Helland, I.S. (1988) On the structure of partial least squares regression. *Commun. Statist. -Simula.* **17**, 581-607.

Helland, I.S. (1990) Partial least squares and statistical models. *Scand. J. Statist.* **17**, 97-114.

Helland, I.S. (1992) Maximum likelihood regression on relevant components. *J. R. Statist. Soc.* B **54**, 637-647.

Helland, I.S. and T. Almøy (1994) Comparison of prediction methods when only a few components are relevant. *JASA* **426**, 583-591.

Hjort, N.L. (1997) Estimation in moderately misspecified models. *JASA* to appear.

Holcomb, T.R., H. Hjalmarson, M. Morari and M. L. Tyler (1997) Significant regression: A statistical approach to partial least squares. *J. Chemometrics* **11**, 283-309.

Kelker, D. (1970) Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyã* **32**, 419-430.

Kruskal, W. (1968) When are Gauss-Markov and least squares estimators identical? A coordinate-free approach. *Ann. Math. Statist.* **39**, 70-75.

Næs, T. and I.S. Helland (1993) Relevant components in regression. *Scand. J. Statist.* **20**, 239-250.

Ripley, B.D. (1996) *Pattern Recognation and Neural Networks.* Cambridge University Press, Cambridge, UK.

von Rosen, D. (1994) PLS, linear models and invariant spaces. *Scand. J. Statist.* **21**, 179-186.

Wold, S., Wold, H., Dunn, W.J. and Ruhe, A. (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *Siam. J. Sci. Stat. Comput.* **5**, 735-743.

Fig. 1. Three hypothetical curves showing the expected mean square prediction error as a function of the number of components inlcuded in the model. The bottom curve is the ideal one, assuming that Condition 1 holds exactly.