# Some theoretical aspects of partial least squares regression.*

Inge S. Helland †

**Abstract**

We give a survey of partial least squares regression with one $y$-variable from a theoretical point of view. Some general comments are made on the motivation as seen by a statistician for this kind of studies, and the concept of soft modelling is criticized from the same angle. Various aspects of the PLS algorithm are considered, and the population PLS model is defined. Asymptotic properties of the prediction error are briefly discussed, and the relation to other regression methods are commented upon. Results indicating positive and negative properties of PLSR are mentioned, in particular the recent result of Butler, Denham and others which seem to show that PLSR can not be an optimal regression method in any reasonable way. The only possible path left towards some kind of optimality, seems then to be through first trying to find a good motivation for the population model and then possibly finding an optimal estimator under this model. Some results on this are sketched.

KEY WORDS: Biased regression methods, continuum regression, PCR, PLS, PLS algorithm, PLSR, population model, prediction, prediction error, regression, relevant components, ridge regression, shrinkage.

---

*Invited paper to appear in a special isue on PLS of Chemometrics and Intelligent Laboratory Systems.

†Department of Mathematics, University of Oslo, P.O.Box 1053 Blindern, N-0316 Oslo, Norway. E-mail: ingeh@math.uio.no

## 1 Introduction.

I am grateful for this opportunity to say something about partial least squares regression from a statistician's point of view. Statisticians on the one hand and on the other hand certain groups of data analysts, including those working with PLS-regression, now seem to be on their way to develop completely separate cultures. It may well be that this development can be fruitful for some limited time, but my own conviction is that in the long run it would be best for all parties if we tried to find some kind of synthesis between the various cultures.

My colleague Emil Spjøtvoll often used to say that 'things can be understood on many different levels'. This was usually said in some pedagogical setting, but it does make sense in, say, methodological research, too. In particular, when new ideas are being developed, the most fruitful approach is often to let rigor rest for a while, and let intuition reign - at least in the beginning. New methods may require new concepts and new approaches, in extreme cases even a new language, and it may then be impossible to describe such ideas precisely in the old language. Also, we all have a limited brain capacity, and if every effort is spent on rigor and precision, there may be no room left for innovation.

I think it is right to say that this point has not always been quite appreciated by those of my collegues that work within theoretical statistics. Mathematical statistics, which should be the umbrella of a large body of methodological research, has at least to some degree developed into a purely deductive science. To a certain extent this is appropriate, when research is made in areas where the conceptual foundation is well established. However, there do exist areas, also in statistics, where it is wise to have an open mind towards different solutions, even if these, at least in the beginning, and at least from a mathematical point of view, may have to be formulated in a slightly intuitive way.

As a consequence of this way of thinking, I have felt for many years that it is important for mathematical statisticians to be open to impulses from the outside. This was my main reason for catching interest when seeing the initial empirical success of the PLS-algorithm in regression, and this was also my reason for bringing with me several preprints by Harald Martens, Svante Wold and other chemometricians on a sabbathical to Edinburgh in 1986, where I spent some time trying to find a structure in it all, as seen from a statistician's point of view. Since then I have returned to the PLS-algorithm from several mathematical directions, some of my colleagues would say more often than the algorithm - which I may agree is after all not more that an algorithm - deserves.

It is important to say at this point that this for me has never been primarily 'an attempt to help chemometricians' - or worse: an endeavour to make chemometrics more academic and respectable. To me this has more been an attempt to extract all the structure from the method that I was able to find, in order to investigate the link to the world of mathematical statistics. And of course the general urge to look for good prediction methods was there in addition, but this has never been the only motivation.

To be very brief, the main structural idea for mathematical statistics which in my view have emerged from this process is simply this: In certain cases it seems to pay to replace the full statistical model with a particular reduced model. The full consequences of this as a general principle in a statistical setting remain to be explored fully. If some general theory of model reduction could be developed, several areas of applied statistics would be affected; there even seems to be some possibility of finding a link towards quantum theory. However, many problems remain to be investigated here.

For a more conventional discussion from the statistical point of view of the use of PLS and related methods, see Brown (1993). The most popular book on the use of such methods in chemometry is still Martens and Næs (1989). Some technical results on PLS have been presented recently in the chemometrical journals, in several statistical journals and also in journals on linear algebra. We will mention some of these, but will concentrate on the main developments.

## 2 Statistical models, prediction and PLSR.

All real data contain noise, and the statistician's way to model this is through probability models. In my view here exist no other really successful alternative to this approach. In particular, the concept of 'soft modelling' seems to be very difficult to make precise. I will try to articulate this critique in a few brief remarks. First, take the soft model concept in its absolutely crudest form: Writing a regression model as

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

without saying anything more, it is completely empty of content: With data $(\mathbf{X}, \mathbf{y})$ we can let $\beta$ be anything, and then we can always define $\mathbf{e}$ by $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$. Then (1) will hold true, trivially. In practice, most soft modellers will probably want to say something about the error term: 'small', 'uncorrelated', 'unbiased' or other characteristics of noise, but it is difficult to make these characterizations precise without using probability models in some way or other.

A more common way to try to make (1) precise in a soft modelling context, is through saying that $\beta$ should be 'found' by least squares estimation. This gives a prescription for finding numbers on both sides of (1), but is still far from what most people would call a model. In particular, the distinction between the parameter in the model and the estimate of this parameter becomes very difficult to retain. For most statisticians it is crucially important to be able to distinguish between the data themselves and the hypothetical world behind, which the data should give information about.

Thus, from a statistical point of view, it is difficult to regard Partial Least Squares Regression - based as it is on a series of least squares fittings - as anything else than an algorithm. Thus it must at the outset be looked upon as a regression *method*, and be compared to other regression methods, not - at least not to begin with - as any radical new approach to regression or to model building. (We will later come back to how a particular form of model *reduction* can be motivated by PLS.)

On the other hand, in some sense one might say that the words 'soft modelling' point towards an aspect of statistical modelling that some statisticians seem to neglect: Most ordinary statistical models are idealizations, and thus say more than what we can read from data. We must even expect in practice that different researchers will use different models on the same data. Therefore, in a prediction context, we may hope that the statistical method used possesses certain robustness properties with respect to the detailed choice of model. This is a very difficult aspect of statistical methodology, which in practice is neglected in many investigations. But in my view the chemometrician's 'soft modelling' concept does not solve this problem in any way. On the contrary, an imprecise modelling concept may even give a context where it to some extent is difficult to state the problem precisely.

In this paper I will concentrate on PLSR as a prediction/ regression method, and say less directly about the use of PLS in finding latent 'loadings' and 'scorings', although these aspects will be touched upon several times. These methods inspired by factor analysis seem to function fairly well in practice, partly because the output given by necessity must be interpreted fairly crudely, but mainly because the methods can be coupled to the same population model as PLSR; see below. Historically, the PLS type latent variable soft models precede the regression method, see Wold (1985).

## 3 The algorithm.

Many chemometricians regard the PLSR1 algorithm as a very simple regression methods, and I agree that more difficulties emerge when we consider the PLSR2-algorithm with several $y$-variables or if algorithms with several blocks of variables are considered. My defense for concentrating on the simples case, is twofold: First, the PLSR1 algorithm is much used. Secondly, if you do not understand the simplest case, the hope that you will ever understand the more complicated case, is very meager.

The paper Helland (1988) started by proving formally the equivalence of two algorithms for PLSR1: The original algorithm with orthogonal scorings proposed in regression context by Wold et al. (1983) and the algorithm with orthogonal loadings proposed by Martens (1985). Both can be looked upon as special procedures where one wants to link the $n$-vector $\mathbf{y}$ of centered $y$-values to the $n \times p$ matrix $\mathbf{X}$ of centered $\mathbf{x}$-values through $k$ 'latent vectors':

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1' + \ldots + \mathbf{t}_k \mathbf{p}_k' + \mathbf{E}_k, \tag{2}$$

$$\mathbf{y} = \mathbf{t}_1 q_1 + \ldots + \mathbf{t}_k q_k + \mathbf{f}_k. \tag{3}$$

For example, the ordinary PLSR1-algorith can be *defined* by a successive use of (2) and (3) for $k = 0, 1, 2, \ldots$ together with a definition

$$\mathbf{t}_k = \mathbf{E}_{k-1} \mathbf{w}_k \quad \text{with weights} \quad \mathbf{w}_k = \mathbf{E}_{k-1}' \mathbf{f}_{k-1}, \tag{4}$$

and where the loadings are determined by least squares fit:

$$\mathbf{p}_k = \mathbf{E}_{k-1}' \mathbf{t}_k / \mathbf{t}_k' \mathbf{t}_k \ = \mathbf{X}' \mathbf{t}_k / \mathbf{t}_k' \mathbf{t}_k, \tag{5}$$

$$q_k = \mathbf{f}_{k-1}' \mathbf{t}_k / \mathbf{t}_k' \mathbf{t}_k \ = \mathbf{y}' \mathbf{t}_k / \mathbf{t}_k' \mathbf{t}_k. \tag{6}$$

The least 'logical' part of this definition of the algorithm is the determination of the weights in (4). Other possibilities can easily be imagined here.

If $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0p})'$ is a set of $\mathbf{x}$-measurements on a new unit, one can use this to define new scorings $t_{j0}$ in a similar way as in the algorithm above, and then predict the corresponding $y$-observation in step $k$ by

$$\hat{y}_{k0} = \bar{y} + \sum_{j=1}^{k} t_{j0} (\mathbf{t}_j' \mathbf{t}_j)^{-1} \mathbf{t}_j' \mathbf{y}. \tag{7}$$

It turns out that this prediction can be written in several equivalent ways. One is terms of the matrix of weights

$$\mathbf{W}_k = (\mathbf{w}_1, \ldots, \mathbf{w}_k).$$

Using this, we can write

$$\hat{y}_{k0} = \bar{y} + (\mathbf{x}_0 - \bar{\mathbf{x}})'\mathbf{b}_k, \tag{8}$$

where

$$\mathbf{b}_k = \mathbf{W}_k(\mathbf{W}_k'\mathbf{X}'\mathbf{X}\mathbf{W}_k)^{-1}\mathbf{W}_k\mathbf{X}'\mathbf{y}. \tag{9}$$

A completely different-looking equivalent formula turns out to be

$$\mathbf{b}_k = \mathbf{W}_k(\mathbf{P}_k'\mathbf{W}_k)^{-1}\mathbf{q}_k \tag{10}$$

in terms of the loadings.

An important side result which comes up during the derivation of these results, is that the weights $\mathbf{w}_1, \ldots, \mathbf{w}_k$ span the same space as the Krylov sequence

$$\mathbf{s}, \mathbf{S}\mathbf{s}, \ldots, \mathbf{S}^{k-1}\mathbf{s},$$

where $\mathbf{S} = \mathbf{X}'\mathbf{X}$ and $\mathbf{s} = \mathbf{X}'\mathbf{y}$.

## 4 The population model.

Let us now turn to the statistician's way of viewing the world.

A crucial concept in all statistical modelling is that of a parameter or a set of parameters: The unknown reality behind our observations which we try to find out as much as possible about. One way to approach the parameters is to look at the idealized situation where we have infinite amount of information, i.e., we let the amount of data tend to infinity. The simplest example is an expectation $\mu$, which can be looked upon as the limit of the mean $\bar{x}$ as the number $n$ of observations tend to infinity.

In other cases it is more natural to look upon the statistical model as a deliberate idealization. For instance: No real data are exactly normally distributed, but even so, regression models with Gaussian error terms have turned out to be very useful idealizations.

In the same way as in the case of the mean, if we want to look for a statistical model behind the PLSR-algorithm, the parameters of that model should be found by letting $n \to \infty$. Then, in particular, in the notation of the last paragraph of the previous Section, $n^{-1}\mathbf{S}$ will tend to the covariance matrix $\mathbf{\Sigma}$ of the $x$-variables, while $n^{-1}\mathbf{s}$ tends to the covariance vector $\sigma$ between $y$ and $\mathbf{x}$. The ordinary regression vector is then $\beta = \mathbf{\Sigma}^{-1}\sigma$. Note, however, that taking the PLSR 'model' with $k < p$ components seriously, turns out to imply a special structure for this regression vector. This is in fact the essence of the PLS population model.

One way to procede, is by taking a principal component decomposition of $\mathbf{\Sigma}$, which we always can do:

$$\mathbf{\Sigma} = \sum_{j=1}^{p} \nu_j \eta_j \eta_j'. \tag{11}$$

Here $\nu_j$ are the eigenvalues and $\eta_j$ the eigenvectors of $\mathbf{\Sigma}$. From the results of the previous Section one can convince oneself that the population version of the space spanned by the PLSR weights is also spanned by

$$\sigma, \mathbf{\Sigma}\sigma, \mathbf{\Sigma}^2\sigma, \ldots, \tag{12}$$

where

$$\mathbf{\Sigma}^{k-1}\sigma = \sum_{j=1}^{p} (\nu_j)^{k-1} \eta_j \eta_j' \sigma. \tag{13}$$

Using known results on Vandermonde determinants and principal components, several results on the population version of PLSR were proved from this in Helland

(1990), among other things: *The population PLSR space has dimension $m$ if and only if there are $m$ different eigenvalues $\nu_j$ with corresponding eigenvectors $\eta_j$ such that $\eta_j'\sigma \neq 0$.*

This should be compared to the formula for the population version of the principal component regression vector:

$$\beta_{PCR} = \sum_{j=1}^{k} (\nu_j)^{-1} \eta_j \eta_j' \sigma. \tag{14}$$

In this equation we naturally want to have as few terms as possible. This can be achieved in two ways: By deleting terms with $\eta_j'\sigma = 0$ and by rotating in eigenspaces with equal eigenvalue $\nu_j$ such that we get only one eigenvector, hence only one term in (14) for each different eigenvalue. Doing this, we see from the previous result that the (minimal) number of terms in the sum (14) will always be equal to the dimension of the PLSR-space.

The next result may perhaps at the outset be even more surprising, but is proved from the same basis: *When this minimal number of terms is used, the population PLS regression vector and the population PCR regression vector are numerically equal.*

Two empirically well known results on the ordinary PLSR and PCR vectors may be understood heuristically from this: 1) The two methods will very often give similar results. 2) PLS regression may tend to require fewer components. The last result can be understood from the fact that the population version of PLSR automatically selects the minimal number of terms when eigenvalues are equal; there is no such automaticity in PCR.

This minimum number of terms in population PLSR/ PCR is called the number of relevant components. If **R** is the corresponding matrix of PLSR population weights, then the part of the $x$-space spanned by $\mathbf{R}'\mathbf{x}$ is called the space of relevant components for the regression. These spaces can be characterized in a number of ways. Here is one from Næs and Helland (1993): The space spanned by $\mathbf{R}$ is the minimal space such that 1) The vector $\beta$ belongs to this space, 2) The same space is also spanned by $\mathbf{\Sigma R}$. A simpler characteristic is perhaps 1) together with the requirement that the space should be spanned by *some* set of eigenvectors of $\mathbf{\Sigma}$.

Finally, a compact way of characterizing a model with $m$ relevant components, is the following; see Næs and Martens (1985) and Næs and Helland (1993): We can write

$$\mathbf{x} = \mathbf{Rz} + \mathbf{Uv}, \tag{15}$$

where $\mathbf{R}$ and $\mathbf{U}$ are fixed matrices of full column rank of dimension $p \times m$ and $p \times (p - m)$, respectively, and where $\mathbf{R}'\mathbf{U} = \mathbf{0}$, $\mathrm{cov}(\mathbf{v}, y) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{z}, \mathbf{v}) = \mathbf{0}$. Then the columns of $\mathbf{R}$ span the same space as the PLSR population weights, and $\mathbf{R}'\mathbf{x}$, or equivalently, $\mathbf{z}$, span the space of relevant components for the regression.

It is shown in Helland (1990) that the sample PLS algorithm as $n \to \infty$ converges to a corresponding population PLS algorithm, and that (2), (3) with $m$ terms under the model with $m$ relevant components can be translated into a factor type model with uncorrelated errors. This gives a way of making the latent structure interpretation of PLS precise, and it is not too difficult to see the connection to (15). The sample loadings and scores that are used in practice are estimates of the loadings and scores in the population model.

Like in all statistical modelling, this population structure gives only probability distributions for the result of the (sample) PLSR algorithm; in addition it says something approximately about the result for the case when $n$ is very large. The rôle of the population model is very important from a theoretical point of view, however: It gives a precise formulation of the ideal reality that lies behind a PLS regression with $m$ terms: A population model with $m$ relevant components. As indicated before, every statistical model is an idealization; in a very specific way, this is the particular idealization which is coupled both to PLS regression and to PLS latent analysis.

## 5 Prediction error.

The next thing a statistician would do when an ideal model is established and it is known that some given estimator converges towards a model parameter as $n \to \infty$, is to try to study closer the distance between estimator and parameter. In the present case the parameter of interest is the regression vector $\beta$, and the estimator is the corresponding vector $\mathbf{b}$ found by, say PLSR or PCR. A natural way to measure the distance between the parameter and estimator in this case, is by

$$d = \mathrm{E}(\mathbf{b} - \beta)'\mathbf{\Sigma}(\mathbf{b} - \beta), \tag{16}$$

which is closely related to prediction error.

For ordinary multiple regression, a straightforward, but but not quite trivial calculation shows that

$$d = \tau^2 \frac{p}{n - p - 2}, \tag{17}$$

where $\tau^2$ is the error variance of the regression equation (cp. Helland and Almøy, 1994). An obvious observation here is that this error can be quite big when the number $p$ of $x$-variables is large; it will even get infinite when $p$ approaches the number $n$ of observations. An obvious remedy might be to the reduce the number of variables, but this will usually lead to an increase in $\tau^2$. What PCR and PLSR do, is in effect to use a reduced number of *linear combinations* of $x$-variables as regressors. The fact that these linear combinations have to be estimated from data, again leads to an increase in the variance. The resulting asymptotic distances are calculated in Helland and Almøy (1994). Simulations related to the same distances are carried out in Almøy (1996). In all these cases an ideal model is assumed where the number of relevant components is fixed at some number $m$.

The asymptotic expressions turn out to be relatively complicated, and will not be reproduced here. Qualitatively, it turns out that the difference between PCR and PLSR in most cases is relatively small. No method dominates the other. PCR does best when the irrelevant eigenvalues are relatively small or relatively large; PLSR does best for intermediate irrelevant eigenvalues. Since the difference is very small for small irrelevant eigenvalues, and since large irrelevant eigenvalues seem to be very rare, this can be interpreted as an, admittedly relatively weak, argument for PLSR in this comparison. The conclusions above are confirmed in the systematically designed simulation study by Almøy (1996).

## 6 Links to other regression methods.

There exist a large number of regression methods that have been proposed for near collinear data: In addition to PCR and PLSR we have ridge regression, latent root regression, various methods suggested by calibration theory or by Bayesian theory, variable selection methods, James-Stein shrinkage etc.. Often these methods give fairly similar results, as already discussed for PCR and PLSR. This fact may make the situation somewhat easier for the user of regression methods, but of course 'often' here does not mean 'always'. One may perhaps still have the hope that with so many similar regression methods available, it might in some future be possible to find one selected method which turns out to be optimal in a fairly natural canonical way.

Stone and Brooks (1990) have proposed a regression method which contains ordinary least squares (OLS), PCR and PLSR as special cases. It is a stepwise procedure, where a generalized criterion is maximized in each step. This criterion depends on a parameter $\alpha$, where $0 \leq \alpha \leq 1$ and $\alpha = 0$ gives OLS, $\alpha = 1/2$ gives PLSR and $\alpha = 1$ gives PCR. The method was demonstrated to perform well on some selected examples, but a weak point in practice is that crossvalidation has to be used to determine both $\alpha$ and the number of steps. Since crossvalidation already is an Achilles' heel in all stepwise procedures like PLSR, this becomes a double problem in continuum regression.

The criterion used by Stone and Brooks (1990) is in reality well known when specified to the three methods: OLS is based on maximizing the empirical correlation between $y$ and $\hat{\beta}'\mathbf{x}$, PCR is based upon maximizing the variance of the normalized linear combination of $x$-variables at each step, while PLSR can be derived from maximizing at each step the covariance between $y$ and such a linear combination. (See also Höskuldsson, 1988.) A heuristic comparison of the three methods together with ridge regression on the basis of such characterizations can be found in Frank and Friedman (1993). Their qualitative discussion using this point of departure can be looked upon as an explanation why these methods are so similar in their performance.

In Sundberg (1993) and Björkström and Sundberg (1999) it was shown that the first step in continuum regression can be written in the form of a generalized ridge regression. Unfortunately, the PLSR case corresponds to a ridge parameter tending to infinity.

## 7 Known positive and negative properties of PLSR.

Most of the comparisons between regression methods have been done via simulation studies. The simulations by Almøy (1996) using the model with $m$ relevant components have already been mentioned. An extensive discussion of PLSR from a statistical point of view, including systematic Monte Carlo simulations, is given in Frank and Friedman (1993). In these simulations ridge regression comes out best in an overall assessment, followed closely by PLSR and PCR, while variable selection does not perform as well as the other methods. The small difference between PLSR, PCR and ridge regression is commented upon by the authors by saying that one would not sacrifice much average accuracy over a lifetime by using one of them to the exclusion of the other two. In the discussion, Svante Wold gives arguments to the effect that ridge regression would probably have performed differently under a different simulation design.

Aldrin (1997) demonstrates by using simulation that regression methods, in particular ridge regression, can be improved on by adjusting the length of the regression vector so that a measure of the prediction error is minimized. In particular, his length modified ridge regression dominates PLSR in the simulations. His discussion also takes into account relations to the continuum regression by Stone and Brooks

(1990).

In fact, it is a well known characteristic of biased regression methods that the length of the corresponding regression vector in general tends to be not too large, and that this aspect has a positive effect on the prediction error. For PLSR a definite mathematical result in this direction can be shown: In de Jong (1995) and in Goutis (1996) it is shown that PLSR shrinks in the sense that one always have $|\mathbf{b}_{PLSR}| \leq |\mathbf{b}_{OLS}|$.

Another nice mathematical property of PLSR has been proved by de Jong (1993): With the same number of components, PLSR will always give a higher coefficient of determination $R^2$ than PCR.

However, taking a closer look on the shrinkage properties of PLSR turns out to expose that the regression method has some serious defects. To understand this, we first do a decomposition of the mean square error of a general regression method with regression vector $\mathbf{b}$, where we (in contrast to what was done in Section 5) follow the common statistical tradition and regard the $x$-variables as non-stochastic:

$$\begin{aligned} MSE = \mathrm{E}(\mathbf{b}-\beta)'\mathbf{S}(\mathbf{b}-\beta) &= (\mathrm{E}\mathbf{b}-\beta)'\mathbf{S}(\mathrm{E}\mathbf{b}-\beta) + \mathrm{tr}(\mathbf{S}\mathrm{V}(\mathbf{b})) \\ &= \sum_{i=1}^{p} \lambda_i (\mathrm{E}a_i - \alpha_i)^2 + \sum_{i=1}^{p} \lambda_i \mathrm{Var}(a_i), \end{aligned} \tag{18}$$

where we have used the eigendecomposition

$$\mathbf{S} = \mathbf{X}'\mathbf{X} = \sum_{i=1}^{p} \lambda_i \mathbf{u}_i \mathbf{u}_i',$$

and have defined

$$a_i = \mathbf{b}'\mathbf{u}_i, \quad \alpha_i = \beta'\mathbf{u}_i.$$

It is easy to see that for ordinary least squares, the random variables $a_i$, say $a_i^0$

for this method, satisfy $\mathrm{E}(a_i^0) = \alpha_i$, so the first sum on the righthand side of (18), the bias part, vanishes.

For ridge regression and for PCR, it is also easy to see that we get $a_i = f(\lambda_i)a_i^0$, where $f(\lambda_i) = \lambda_i/(\lambda_i+\kappa)$, respectively $f(\lambda_i) = 1$ or 0. Here $\kappa$ is the ridge parameter, and for the PCR case $f(\lambda_i) = 1$ for the terms included in the regression.

Thus in both these cases we have $0 \leq f(\lambda_i) \leq 1$, which is a very good thing, for if we had $f(\lambda_i) > 1$ for some $i$, then both the corresponding bias term and the variance term in (18) would increase relative to ordinary least squares, and the method could have been easily improved on by simply replacing $f(\lambda_i)$ by 1 for this particular $i$. Since this is not the case, both ridge regression and PCR are true shrinkage methods in this very satisfying strong sense.

Now the question is: Is PLSR a shrinkage method in this sense? Heuristic discussions and calculations made by Frank and Friedman (1993) seemed to indicate that this was *not* the case. This suspicion has recently been confirmed by rigorous detailed mathematical calculations independently by Butler and Denham (2000) and by Lingjærde and Christophersen (2000): The functions $f(\lambda_i)$ (which must be made stochastic for PLSR) are in fact quite often larger than 1. This strongly indicates that PLSR nearly always can be improved in principle, so the regression method as such is not optimal in any reasonable way.

This strong and important negative result for PLSR can also be elucidated by using the population model of Section 4: The population model with $m < p$ steps is equivalent to a definite restriction of the original model parameters, say, by stating that the population weight at step $m+1$ is zero. In a way it is true that the sample PLS loadings, weights etc. give reasonable estimates of the corresponding population quantities, but these estimates have a very important defect: The sample estimates do not satisfy the restrictions implied by the population model. For instance, the probability that the sample weight at step $m + 1$ should vanish, is zero.

Thus any question about finding out in which sense the ordinary PLS algorith should be optimal, is in fact meaningless. The most we can do, is to state the following two questions:

1) In what settings are the model reduction assumed by the population PLS model the most meaningful one?

2) Given the population PLS model with $m$ steps, what are the best possible estimates of the parameters of this model?

In Helland (2000c) a fairly satisfying answer to 1) is given, and a discussion of 2) is also started.

## 8 Model reduction and PLSR.

For centered data $(\mathbf{x}, y)$ the covariance structure in itself gives a model containing $(p + 1)(p + 2)/2$ parameters. For multinormal data the whole model is completely specified by these parameters; we may assume this as a simplification, or simply ignore the rest of the structure, which is relatively unimportant in the search for a good predictor (Helland, 2000a).

The population PLS model with $m$ relevant components corresponds to a definite restriction of this model having a net number of parameters equal to $p(p+1)/2+m+1$. For $m = 0$ there is no correlation between $\mathbf{x}$ and $y$ in this model, for $m = p$ we get back the full model. Going from $m = 0$ to $m = p$ in steps gives a simple hierarchy of models.

In practice, nearly every statistical model is a simplification, and it is far from uncommon in statistics to reduce a model by reducing the number of parameters in order to try to explore some specific structure. One major difficulty is that the joint covariance model as such can be reduced in a large number of ways. An important problem is then to find out under what conditions the particular reduction implied by population PLS is natural. From a latent variable point of view it seems to be possible to say something about this. We will here be more interested in attacking the question in a regression/prediction setting.

In Helland (2000a) the prediction error itself was taken as a point of departure. Then it may be more natural to consider different conditionings of the joint model instead of different model reductions; for prediction purposes this can be shown to be equivalent. Hence assume that we want to predict $y$ by regressing on a linear combination $\mathbf{R}'\mathbf{x}$ of the $x$-data, where $\mathbf{R}$ is a fixed $p \times k$ matrix of rank $k$. Then, in in the same way as we found (17) the prediction error can be shown to be

$$PRE = \tilde{\tau}^2 \frac{n-1}{n-k-1}, \tag{19}$$

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

For fixed $n$ and $k$ this prediction error is as small as possible if and only if $\tilde{\tau}$ attains its minimal value $\tau$, which happens if and only if $\beta$ is in the space spanned by $\mathbf{R}$. This is the first condition needed to give the population PLS model, the second one ($\text{span}(\mathbf{\Sigma R}) = \text{span}(\mathbf{R})$) can be argued for qualitatively by looking at the reduction in $\tilde{\tau}$ from step to step as $k$ increases. A detailed discussion is given in Helland (2000a), where it is shown that under an extra technical condition PLSR gives the best model reduction from step to step. The technical condition seems to be unavoidable here.

The approach of Helland (2000c) is completely different. The point of departure is: Nearly all known regression methods are equivariant (i.e., the estimated regression vector transforms in the same way as the parameter vector) under rotations in the $x$-space. Thus this group of rotations seems to be of some significance when studying regression methods.

As a first step, the orbits of this group in the parameter space was determined: The set of parameter values $g\theta_0$, where $\theta_0$ is some fixed parameter value (in this case a vector of parameters), and where $g$ runs through the elements of the rotation group. It is known from theory that on each orbit one can in principle find a unique optimal estimator; this was given as a rather complicated multiple integral for the regression case in Helland (2000c). The parameters that are left to estimate, are then the indices of the orbits; all sensible model reductions must also be done on these parameters.

For the rotation group in the regression problem, the orbit indices turn out to be coupled in a natural way to the eigenspaces of $\mathbf{\Sigma}$ with different eigenvalues $\nu_j$. Going closer into this approach, turns out to give a very satisfying solution from a PLS modelling point of view: The best regression estimator, given the orbit, as explained above, turns out not to depend upon the whole orbit index. It may depend upon the residual variance $\tau^2$, on the number $m$ of relevant components which we defined from a population model point of view earlier, and on some symmetric function of the parameters$(\lambda_j, \gamma_j); j = 1, \ldots, m$, where $\gamma_j$ is the norm of the projection of the regression vector upon the eigenspace corresponding to $\lambda_j$.

This means that the simplest and most obvious solution of the model reduction problem is just to say that the number of eigenspaces with $\gamma_j \neq 0$ should be specified to some fixed number $m$. This specification leads directly to the population PLS model with $m$ relevant components, now apparently motivated in a purely theoretical setting. A similar solution can be devised for the PLS discriminant analysis model.

Thus it seems natural to accept the PLS population model. The remaining theoretical problem is to find the best possible estimator of the parameters under the model with $m$ relevant components. The maximum likelihood estimator of the model was discussed in Helland (1992). This turns out to be rather cumbersome to calculate, and more seriously, performs poorly in simulations (Almøy, 1996). The reason for this is probably straightforward: Maximum likelihood is a very good general procedure when the number of observations in a model is large compared to the number of parameters. For the population PLS model, even though the number of parameters has been slightly reduced for better prediction performance, the *total* number of parameters is still large, so a very large number of observations is required for maximum likelihood to perform well.

A more promising approach seems to be to use invariance as above, which may be shown to lead to a proven optimal estimator of a large part of the parameter space, namely for each fixed orbit. Several problems need to be solved before this approach can be made practical, however. The most serious problem is probably to find an efficient way to compute the multiple integral mentioned above; another problem is to find good estimates for the orbit parameters $((\lambda_i, \gamma_i); i = 1, \ldots, m)$.

Most people will of course stick to the ordinary sample PLS algorithm, and regard the search for better estimators as rather academical. The sample PLS method is definitely known to be suboptimal, however, and perhaps more seriously: Very little seems to be known in the context of PLSR about the ability of cross validation as a procedure to find the best number of components. A possible side result if one was able to find a workable estimation procedure based on statistical theory, might be that one could replace cross validation with some simple test procedure with known properties. Again, however, it is not clear that any workable solution in this direction can be found.

## 9 Discussion.

In general, it is clear to me that the ordinary tool of statisticians: probability models indexed by unknown parameters, has turned out to be very useful in a large number of applications. At present I see no way in which concepts like soft modelling can replace this tool fully or partly, even though developments using such loose concepts might indeed lead to useful methods. One problem is that without a precise modelling concept, it is very difficult to make assessments of the different methods. Thus in my view the only consistent way to proceed in the long run will be to try to connect such methods to statistical model of the ordinary kind.

However, I do feel that the statistician's tool may have to be supplemented in various directions. One example of an activity which falls outside the ordinary theoretical statistical paradigm, is model reduction, which has been done informally for a long time by applied statisticians, but for which no general theory exists. The population model for PLSR may provide one particular case which may give us a clue to such a general theory. Other possible clues exist, for instance symmetry, which can be explored systematically by using group theory; a survey of this area from the point of view of theoretical statistics has recently been given in Helland (2000b).

As I see it, it is essentially important for the development of theoretical statistics that one tries to keep in touch with various applications, and also with other groups working with inference and stochastic modelling, like financial analysts, geophysicists, control engineers, quantum physicists and chemometricians. It is also of course a hope that these groups may, at least to some extent, benefit from contact with statisticians. Some of these inference groups have other traditions and may be partly speak other languages than ours. So translation is an important task. To put it simply: The purpose of science, also methodological science, is to seek the truth. And when you look for something, it is often wise to look several places.

One of my hopes for the future is that more scientists with different background, statisticians and non-statisticians, will engage in interdisciplinary work on methodological questions. This is a work that takes time and efforts, but there is very much which remains to be done. Setting it all in a wider context, and perhaps exagerating a little, this is a kind of activity which may feel extra meaningful in the world as we know it, where cultural antagonisms are the sources of some of our most serious and devestating problems.

The ultimate goal of an activity of this kind should perhaps not primarily be compromises between different schools, but where possible something much more ambitious, namely syntheses between various ways of thinking. Such goals must by necessity take long time to reach, and it requires researchers who during the process are able to understand the set of explicit and implicit values of all scientific communities involved. In the meantime, the concept of complementarity, inherited from quantum theory, may perhaps have something to offer in this setting: It may make sense sometimes to look at one and the same problem from several different angles, even if these different viewpoints are not at the outset mutually consistent. Typically, however, the inconsistency may be due to the following: Each point of view is necessarily connected to some simplification, but we can often imagine that different simplifications of the same basic paradigm are used for the different viewpoints. Thus model reduction - or something related - may seem to be the issue again.

## References


Aldrin, M. (1997). Length modified ridge regression. *Computational Statistics & Data Analysis* **25**, 377-398.

Almøy, T. (1996). A simulation study on comparison of prediction methods when only a few components are relevant. *Computational Staistics & Data Analysis* **21**, 87-107.

Björkström, A. and R. Sundberg (1999). A generalized view on continuum regression. *Scand. J. Statist.* **26**, 17-30.

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

Butler, N.A. and M.C. Denham (2000). The peculiar shrinkage properties of partial least squares regression. *J. R. Statist. Soc.* B **62**, 585-593.

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

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

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

Goutis, C. (1996). Partial least squares algorithm yields shrinkage estimators. *Ann. Statist.* **24**, 816-824.

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 regression 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. (2000a). Model reduction for prediction in regression models. *Scand. J. Statist.* **27**, 1-20.

Helland, I.S. (2000b). Statistical inference under a fixed symmetry group. Submitted to Ann. Statistics.

Helland, I.S. (2000c). Reduction in regression models under symmetry. Invited contribution to: Viana, M. and D. Richards [Ed.] *Algebraic Methods in Statistics.* Contempory Mathematics Series of the American Mathematical Society.

Helland, I.S. and T. Almøy (1994). Comparison of prediction methods when only a few components are relevant. *J. Amer. Statist. Ass.* **89**, 583-591.

Höskuldsson, A. (1988). PLS regression methods. *J. Chemometrics* **2**, 211-228.

Lingjærde, O.C. and N. Christophersen (2000). Shrinkage structure of partial least squares. *Scand. J. Statist.* **27**, 459-473.

Martens, H. (1985). *Multivariate Calibration.* Dr. techn. Thesis. Technical

University of Norway, Trondheim.

Martens, H. and T. Næs (1989). *Multivariate Calibration.* Wiley, New York.

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

Næs, T. and H. Martens (1985). Comparison of prediction methods for collinear data. *Commun. Statist. Simulat. Comput.* **14**, 545-576.

Stone, M. and R.J. Brooks (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regeression (with discussion). *J. R. Statist. Soc.* B **52**, 237-269; corrigendum, **54** (1992), 906-907.

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

Wold, H. (1985). Partial Least Squares. In: Kotz, S. and N.L. Johnson. *Encyclopedia of Statistical Sciences.* Wiley, New York.

Wold, S., H. Martens and H. Wold (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Proc. Conf. Matrix Pencils* (A.Ruhe, B. Kågstrøm, eds.) Lecture Notes in Mathematics, Springer Verlag, Heidelberg, 286-293.