Reduction of regression models under symmetry.*

Inge S. Helland †

Abstract

For collinear data nearly all regression methods that have been proposed, are equivariant under the rotation group in the x-space. It is argued that the regression parameter along orbits of the rotation group in principle always can be estimated in an optimal way as a Pitman type estimator. On the other hand it is argued that it may pay in general to reduce the parameter space of a statistical model when this space is high-dimensional. It follows that any reduction in the regression model then must take place via the orbit index of the rotation group. Further information can be found using the form of the loss function. This is used to discuss the choice of regression model and thereby the choice of regression method. The solution which seems to emerge from this, is closely related to the population version of the chemometricians' partial least squares regression. Estimation under the reduced model is briefly discussed, as is model reduction in the corresponding classification problem.

KEY WORDS: Collinearity; Model reduction; Multiple regression; Orbit; Partial Least Squares Regression; Pitman estimator; Rotation; Symmetry.

---

*Invited paper to be published in Viana, M. and D. Richards [Eds.] Algebraic Methods in Statistics. Contemporary Mathematics Series of the American Mathematical Society.

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

1 Introduction.

Collinear data in regression and in multivariate fields like classification is a major practical problem, and also a problem area where there exist a large number of seemingly unrelated methods, many of them derived by ad hoc arguments. The present paper is an attempt to introduce some general theory into this area.

Our point of departure is a rather heuristic statement, which will be made more concrete in the next sections: When the number of parameters in a statistical model is relatively large and the data set not that big, a sound advice is often to reduce the model. This can of course be done in very many ways, but in the present paper we consider a situation where we are given some help in the model reduction process:

Assume in general that there is a natural symmetry group attached to the model. Then the model parameter can be divided into an orbit index (maximal invariant under the group) and a parameter along the orbit. This last parameter will be invariantly measurable, and it can be estimated in a rather satisfying optimal way, at least in principle, by finding the minimum risk equivariant estimator along the orbit. Hence the potential for gaining anything via model reduction is limited to the orbit index. The reduced model should also be invariant under the chosen group. This can again be achieved by making the model reduction through reducing the orbit index parameters. Additional information is found from the specific form of the loss function.

A natural group to look at in many multivariate situations is the rotation group. In a regression context it is often natural to consider estimators that are equivariant under the group of rotations in the x-space. After a brief example to motivate the use of model reduction in regression as a general procedure, we discuss the characterization of the orbits of the rotation group in the parameter space and the degree of non-exactness of the same group. Thereafter this is used to discuss model reduction and implications to the construction of regression methods.

2 Model reduction in simple regression.

Look at the elementary linear regression model $y_i = \alpha + \beta x_i + \epsilon_i$, where the error terms ϵ_i are iid $N(0, \sigma^2)$, and where we for simplicity assume that the values $x_1, \ldots, x_n$ of the explanatory variable satisfy $\sum x_i = 0$. A straightforward calculation then shows that the mean squared prediction error at the fixed point x_0 under the nonrandom explanatory variable model equals

$$P = \frac{\sigma^2}{n} + x_0^2 \frac{\sigma^2}{\sum x_i^2} + \sigma^2. \tag{1}$$

.

We will start by illustrating on this simple example our first main point: In many situations it will pay to reduce the statistical model. In fact, this is very often done on an intuitive way by applied statisticians and by users of statistics. In my view, one of the most important open areas of theoretical statistics is to develop a sufficiently general theory of model reduction. Roughly, model reduction may be appropriate if the number of parameters in the original model is relatively large and the number n of data points is limited. This should be contrasted to standard statistical theory which concentrates on the ideal situation where n tends to infinity and the number of parameters is fixed. Like all tools, model reduction can be misused, but this should not prevent us from trying to investigate when such a reduction may be beneficial.

In the present simple situation, we have a small number of parameters, but even here model reduction can be considered. The strength of the example is that an explicit solution is easy to find. The most natural reduced model here will be one without slope: $\beta = 0$, leading to the prediction error implied by using the reduced model when data are generated by the full model:

$$P^R = \frac{\sigma^2}{n} + \beta^2 x_0^2 + \sigma^2. \tag{2}$$

Comparison with (1) shows that in terms of prediction error, model reduction pays (for all non-zero x_0) iff

$$\beta^2 < \frac{\sigma^2}{\sum x_i^2}. \tag{3}$$

This condition can also be written as $|t| < 1$, where t is the 'theoretical t-value' $\beta/\text{std}(\hat{\beta})$ with $\text{std}(\hat{\beta}) = \sigma/\sqrt{\sum x_i^2}$. In this form it can be shown that the condition also can be generalized to the question of deleting a single variable from a multiple regression model, a fact that also has been referred to in applied statistics books like Snedecor and Cochran (1989).

In this paper we shall mainly concentrate on the random x regression, where the assumption on the error terms is that ϵ_i, given all the x-variables are iid $N(0, \sigma^2)$, and we typically may assume that $x_1, \ldots, x_n$ are iid $N(0, \sigma_x^2)$. Thus the assumption $\sum x_i = 0$ is replaced by $\text{E}(x_i) = 0$. Then the new prediction error is found by taking the expectation over the x-variables in (1) and (2), leading to the following criterion for model reduction:

$$\beta^2(n-2) < \frac{\sigma^2}{\sigma_x^2}. \tag{4}$$

This illustrates explicitly the statement made earlier that model reduction may be beneficial when the number of data points is small or moderate. Unfortunately, the criterion depends upon unknown parameters - a general problem in this area.

In this paper we will concentrate on the act of reducing a given model; the estimation process under the reduced model will not be fully discussed. In Bickel (1984) a general discussion can be found on the problem of estimating under a simple model when wanting to guard against deviations towards a larger model. A very general large sample discussion of model reduction using likelihood theory can be found in Hjort (1991). A number of specific examples of model reduction can also be found in these two papers and in references given there, while other specific cases are discussed elsewhere. For instance, a treatment of model reduction in Kalman filter models is given in Desai et al (1985). The idea of using rotation symmetry to find the best reduction of linear models in the way advocated in the present paper, seems to be new.

3 The multiple random x regression model under rotational symmetry.

Consider now in general a p-dimensional x-distribution and the corresponding $(p+1)$-dimensional joint (x, y)-distribution with expectation $(\mu_x^t, \mu_y)^t$ and covariance matrix

$$\begin{pmatrix} \Sigma_{xx} & \sigma_{xy} \\ \sigma_{xy}^t & \sigma_y^2 \end{pmatrix}. \tag{5}$$

From now on we will usually replace these covariance parameters by the equivalent parameter set $\theta = (\Sigma, \beta, \sigma_y^2)$, where $\Sigma = \Sigma_{xx}$ and $\beta = \Sigma_{xx}^{-1}\sigma_{xy}$. The parameter of interest is β, and we will be interested in prediction problems, assuming a linear conditional expectation $\mathrm{E}(y|x) = \alpha + \beta^t x$ (this includes the multinormal case, and, more generally, the case of elliptical distributions; see references in Helland, 2000a). We also assume a linear predictor $\hat{y}_x = \hat{\alpha} + \hat{\beta}^t x$. From this, considering the expectation of $(\hat{y}_x - y)^2$ over future x's and using $\hat{\alpha} = \bar{y} - \hat{\beta}^t \bar{x}$, one can see that a natural loss function (cp. also Theorem 1 in Helland and Almøy, 1994) is

$$L(\theta; \hat{\beta}) = (\hat{\beta} - \beta)^t \Sigma (\hat{\beta} - \beta). \tag{6}$$

The data set from which estimation shall be made, consists of n independent observations from this model, summarized in the usual way as (X, y). When n is large compared to p, nobody would try to challenge the ordinary least squares estimate $\hat{\beta}_{LS} = (\mathcal{X}^t\mathcal{X})^{-1}\mathcal{X}^t y$, where $\mathcal{X}$ is the centered X-matrix. But the difficult situations arise when p is relatively large, even of the same order as n, or more generally, if $\mathcal{X}^t\mathcal{X}$ may have some extremely small eigenvalues.

Most of the solutions that have been proposed for prediction in regression models: principal component regression, latent root regression, ridge regression, partial least squares regression, continuum regression and so on, are equivariant under rotation in the x-space of the above model. Therefore, a natural task might be to try to find the best or nearly best estimator among these equivariant ones. For completeness we repeat the necessary definitions; see for instance Lehmann and Casella (1998) for further discussion.

The rotation group in question has group elements g which can be identified by the orthogonal $p \times p$ matrices Q with determinant $+1$. In the sample space, the group G is given by the transformations $(\mathcal{X}, y) \to (\mathcal{X}Q, y)$, which induces the group $\bar{G}$ acting on the parameter space through the transformations

$$(\Sigma, \beta, \sigma_y^2) \to (Q^t \Sigma Q, Q^t \beta, \sigma_y^2). \tag{7}$$

The parametric function $\beta = \beta(\theta)$ is trivially seen to be invariantly estimable (also called permissible; see Helland, 2000b): $\beta_1 = \beta_2$ implies $Q^t\beta_1 = Q^t\beta_2$ for all Q. An estimator $\hat{\beta}$ is equivariant if we also have $\hat{\beta} \to Q^t\hat{\beta}$ when $(\mathcal{X}, y) \to (\mathcal{X}Q, y)$. For equivariant estimators the loss (6) will be invariant. Since the parameter space is closed under the transformations in $\bar{G}$, we have what Lehmann and Casella (1998) call an invariant estimation problem.

A difficulty, however, is that the group here is not transitive on the parameter space. For transitive groups the risk (expected loss) will be a constant function of the parameter, which is a strong indication that the problem of finding an equivariant estimator that minimimizes the risk uniformly, has a unique solution. Such best equivariant estimators are indeed found quite generally as the Pitman estimator and its generalizations to non-location groups.

In the present case, however, such uniqueness can only be expected when estimating parameters on the orbits of $\bar{G}$. When estimating orbit indices (maximally invariant parameters under $\bar{G}$), other methods must be used. It is only with respect to this last part of the estimation that we can expect any gain from trying to reduce the parameter space when p is large.

4 Non-exactness, orbits and maximal invariants in the parameter space.

We start by decomposing the covariance matrix Σ, that is, giving the general form that such a positive definite matrix might have, i.e.,

$$\Sigma = \sum_{k=1}^{q} \lambda_k P_k, \tag{8}$$

where the P_k are projection matrices upon orthogonal eigenvector spaces V_k (which we will call *strata* in analogy with the use of this term in Nelder, 1965). Here all the λ_k are positive and different (otherwise strata could have been combined). Without loss of generality we can take the λ_k's in decreasing order. Obviously, $q \leq p$.

We will assume that Σ has full rank p, which is equivalent to

$$\sum_{k=1}^{q} \dim(V_k) = \sum_{k=1}^{q} \mathrm{tr}(P_k) = p,$$

or again to the requirement that the direct sum of the spaces V_k equals the full p-dimensional Euclidean space R^p.

The first question we ask, is what transformations in the rotation group that conserve the matrix Σ.

Proposition 1.

The following are equivalent:

(a) $Q^t \Sigma Q = \Sigma$.

(b) $QP_j = P_j Q$ *for* $j = 1, \ldots, q$.

(c) Q *can be written as the commuting product of* q *rotations* Q_k*, where* Q_k *is a rotation only within one stratum* V_k.

The proof of this is given in Appendix 1.

The next question is simpler: Which transformations conserve the regression vector β? From $Q^t \beta = \beta$ follows that β is an eigenvector of Q with eigenvalue 1, or a more useful characterization: Q is a rotation around the axis determined by β. Combining these two results, we get:

Corollary 1.

The transformation Q *conserves* (Σ, β) *if and only if* Q *is the commuting product of* q *rotations* Q_k*, each acting on a single stratum* V_k*, such that for each stratum* V_k *upon which* β *has a non-zero component* $P_k\beta$ *we have that* Q_k *is a rotation around the axis given by* $P_k\beta$.

Since σ_y^2 is unaffected by the transformations, this corollary characterizes the transformations that conserve $\theta = (\Sigma, \beta, \sigma_y^2)$, i.e. it shows the degree of non-exactness of the rotation group in the parameter space.

The following theorem gives one way of characterizing the orbits of the same group in the parameter space. We omit the proof, since it uses the same techniques and is very similar to the proof of Proposition 1, now looking at the equation $Q^t \Sigma_1 Q = \Sigma_2$ instead of the previous $Q^t \Sigma Q = \Sigma$.

Theorem 1.

Two parameter values $\theta_1 = (\Sigma_1, \beta_1, \sigma_{y,1}^2)$ *and* $\theta_1 = (\Sigma_2, \beta_2, \sigma_{y,2}^2)$ *with* $\Sigma_r = \sum_{k=1}^{q_r} \lambda_k^{(r)} P_k^{(r)}$ *are on the same orbit of the rotation group if and only if:*

(1) $q_1 = q_2 = q$.

(2) $\lambda_k^{(1)} = \lambda_k^{(2)}$, $(k = 1, 2, \ldots, q)$.

(3) There is a rotation matrix Q *such that* $P_k^{(2)} = Q^t P_k^{(1)} Q$, $(k = 1, \ldots, q)$. *In particular, the set of dimensions of* $\{V_k^{(2)}\}$ *must match the dimensions of* $\{V_k^{(1)}\}$.

(4) The same Q *gives* $\beta_2 = Q^t \beta_1$, *or equivalently,* $P_k^{(2)} \beta_2 = Q^t (P_k^{(1)} \beta_1)$, $(k = 1, \ldots, q)$.

(5) $\sigma_{y,2}^2 = \sigma_{y,1}^2$.

From this result we also get the maximal invariant of the parameter group, since this always equals the index of the orbits.

Corollary 2.

The orbit index of the parameter group is given by

(i) For the ordered set of strata: Their relative orientation, their dimensions and the corresponding eigenvalues λ_k.

(ii) The norms of the projected regression coefficients $\gamma_k = ||P_k \beta||$.

(iii) σ_y^2.

Proof.

From Theorem 1 it is only left to prove that it is enough to characterize the β-dependence of the orbits by the norms γ_i of each stratum component. This

can be seen from Proposition 1. By that result, the matrix Q of Theorem 1 (3) has a non-uniqueness corresponding to any set of rotations within some or all of the strata. Using such a rotation within stratum i, we see that the direction of the stratum component of β within each stratum is arbitrary, and only its norm is fixed.

5 Optimizing on orbits.

Assume now data (X, y) from n independent units. In this section we will also assume multinormality - at least as a useful approximation. The estimation of the expectations is then trivial, and by sufficiency the estimation of the covariance structure can be assumed to depend only on

$$S = S_{xx} = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t,$$

$$s = s_{xy} = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

$$s_y^2 = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Now concentrate on a fixed orbit of the rotation group in the parameter space. This group is compact, and thus has a unique Haar measure, which can be taken as a prior probability. Then using well known results generalizing the Pitman estimator (see, for instance Helland, 2000b; the essential result is also given in Berger, 1980) and using the loss function (6), the minimum risk estimator, given the orbit, will be

$$\hat{\beta}_O = \mathrm{E}_{post}(\Sigma_{xx})^{-1}\mathrm{E}_{post}(\sigma_{xy}) = \mathrm{E}_{post}(\Sigma)^{-1}\mathrm{E}_{post}(\Sigma\beta), \tag{9}$$

where E_{post} is the corresponding aposteriori expectation. The minimal risk (given orbit) will then be

$$R_{min} = \beta^t\Sigma\beta - \mathrm{E}_{post}(\Sigma\beta)^t\mathrm{E}_{post}(\Sigma)^{-1}\mathrm{E}_{post}(\Sigma\beta), \tag{10}$$

because $\mathrm{E}_{post}(\beta^t\Sigma\beta) = \beta^t\Sigma\beta$ since $\beta^t\Sigma\beta = \sum_k \lambda_k\gamma_k^2$ is a constant of the orbit.

As ususal, the posterior density is proportional to the product of the likelihood and the prior. Using a known form for the multinormal likelihood, found by first taking the likelihood of X, and the multiplying by the conditional likelihood of y, given X, we find:

Theorem 2.

Fix an orbit of the rotation group in the above situation. Then the minimum risk equivariant estimator for β, given the orbit, is given by

$$\begin{gathered}\hat{\beta}_O = D^{-1}N = \\ (\int Q^t\Sigma Q \exp[\tfrac{n}{2\sigma^2}(2\beta^t Qs - \beta^t QSQ^t\beta) - \tfrac{n}{2}\mathrm{tr}(Q^t\Sigma^{-1}QS)]d\gamma(Q))^{-1} \\ (\int Q^t\Sigma\beta \exp[\tfrac{n}{2\sigma^2}(2\beta^t Qs - \beta^t QSQ^t\beta) - \tfrac{n}{2}\mathrm{tr}(Q^t\Sigma^{-1}QS)]d\gamma(Q)),\end{gathered} \tag{11}$$

where (Σ, β) correspond to a fixed point on the orbit and $d\gamma(Q)$ is Haar measure for the rotation group with group elements identified with orthogonal matrices Q. Also, $\sigma^2 = \sigma_y^2 - \beta^t\Sigma\beta = \sigma_y^2 - \sum_k \lambda_k\gamma_k^2$ is the residual variance, a constant of the orbit.

To calculate explicitly the integrals of (11) seems to be impossible except in some special cases. For instance, when $\Sigma = S = I$, it follows from Gradshteyn and Ryzhik (1994), formula 4.641.2 together with some simple derivations, cp. also Example 4 in Helland (2000b), that

$$\hat{\beta}_O = \frac{I_{\frac{p}{2}}(\frac{n}{\sigma^2}\|\beta\| \cdot \|s\|)\|\beta\|}{I_{\frac{p}{2}-1}(\frac{n}{\sigma^2}\|\beta\| \cdot \|s\|)\|s\|}s,$$

where $I_q(\cdot)$ is the modified Bessel function of the second kind.

This is unfortunate, for if it was possible to give (11) as an explicit formula, this formula could be of definite practical importance: For a fixed orbit, the solution above is the best equivariant estimator. Practically all known regression methods for collinear data are equivariant under the rotation group. This means that, given that some definite established method can be put into the form of first estimating the orbit index of the rotation group, and given that we estimate this orbit index in the same way as in the method in question, the estimator derived from (11) will dominate this method.

In fact, more than this can be said. The estimator (11) does not depend on the whole orbit index as given in Corollary 2.

Proposition 2.
The parameter dependence of (11) is given by
(i) σ^2.
(ii) The number m *of strata such that* $\gamma_k \neq 0$.
(iii) For given σ^2 *and* m*, a vector function* $\hat{\beta}_O((\lambda_1, \gamma_1), \ldots, (\lambda_m, \gamma_m); \text{data})$ *which for all data-values is symmetric under any permutation of the* m *arguments.*

Proof.
Fix σ^2, and look first at the 'denominator' D in $\hat{\beta}_O$. It depends upon the following quantities:

$$\begin{array}{c} Q^t \Sigma Q = \sum_i \lambda_i (Q^t e_i)(Q^t e_i)^t, \\ Q^t \beta = \sum_i \beta_i (Q^t e_i), \\ \text{tr}(Q^t \Sigma^{-1} Q S) = \sum_i \frac{1}{\lambda_i} (Q^t e_i)^t S (Q^t e_i), \end{array} \tag{12}$$

where $\{e_i\}$ is a set of eigenvectors for Σ and $\beta_i = \beta^t e_i$. When the set $\{Q^t e_i, i = 1, \ldots, p\}$ is given a uniform distribution under rotation, the expectation of any such function will depend upon $\{(\lambda_i, \beta_i), i = 1, \ldots, p\}$ in a symmetric way. One particular such symmetric function is the number q of strata. So assume, say, that $V_1 = \text{span}(e_1, \ldots, e_t)$, say, is a stratum with constant λ_1. Then, in particular, the function D should be symmetric under any rotation within this stratum. This means that it can be regarded as a function of $((\lambda_1, \gamma_1), (\lambda_1, 0), \ldots, (\lambda_1, 0))$ together with the rest of the paramaters, or equivalently a function of (λ_1, γ_1) and these parameters. Hence an alternative, and more convenient way to state the result above, is that, given σ^2 and q, we have that D is a symmetric function of $\{(\lambda_k, \gamma_k), k = 1, \ldots, q\}$.

A similar analysis can be made for the 'numerator' N, but looking at this part, we are also lead to a refinement. In the same way as the stratum number q, the function $m =$ (number of strata such that $\gamma_k \neq 0$) is also a symmetric function of $\{(\lambda_i, \beta_i), i = 1, \ldots, p\}$ (obviously, $m \leq q$). Now take a closer look

at (11). Without loss of generality, we can let Σ in this formula be diagonal with eigenvalues on the diagonal. Then we see that any contribution from the strata with $\gamma_k = 0$ will vanish. Thus we can take $\hat{\beta}_0$ as the symmetric function of the m bivariate orbit parameters as stated.

Corollary 3.

The minimal risk given the orbit, R_{min}, depends upon σ^2 and m, and given these parameters it is a function of $((\lambda_1, \gamma_1), \ldots, (\lambda_m, \gamma_m))$ which is symmetric under any permutation of the m arguments.

6 Model reduction.

Consider in general a statistical model with parameter $\theta \in \Theta$, an open set in some topological space, and let $\bar{G}$ be a non-transitive group on this space. Let $L(\hat{\theta}, \theta)$ be the loss function of an invariant estimation problem in this setting. Fix $\theta_0 \in \Theta$. We are interested in the effect of reducing the statistical model by restricting θ to some set $\Gamma \subset \Theta$ such that $\theta_0 \in \Gamma$. From what has been said earlier, an optimal equivariant estimator can be found quite generally on the orbit of $\bar{G}$ containing θ_0. Thus we can without loss of generality let Γ contain this orbit. This gives the minimal useful choice of Γ. In practice, Γ should be chosen as a set Φ in the orbit index space of $\bar{G}$ taken together with the corresponding orbits. Our task is to say something about the choice of Φ. Let us specialize again to the regression setting.

Proposition 3.

Assume the multiple regression model, and let $\bar{G}$ be the group of rotations in the parameter space. Let β be estimated as in Theorem 2 for each fixed orbit. Then the only model reductions that can affect the estimate are given by combinations of some or all of the following possibilities:

- *fix the residual variance σ,*
- *fix the number m,*
- *fix some specific symmetric function of $((\lambda_1, \gamma_1), \ldots, (\lambda_m, \gamma_m))$.*

This follows immediately from Proposition 2.

Looking at these three possibilities, the first one seems not very attractive from an applied point of view; it will hardly lead to any decrease in the prediction error. Also, the last type of model reduction may be an option when one has particular model information, but it seems very difficult to specify any such function which can serve as an input to a general regression method. Furthermore, this reduction method only makes sense in combination with the second method. This leaves one in general with the option of fixing the number m at some value $m < p$ as definitely the most natural approach to model reduction from the present point of view.

It is interesting, then, that a practical regression method closely related to this particular choice has been used by chemometricians by decades now in a large spectrum of applications. The method is claimed to have great success, empirically, and it has obtained quite an amount of attention. A special issue of the journal Chemometrics and Intelligent Laboratory Systems devoted to this particular method is planned to appear shortly.

7 Partial least squares regression in population form.

Partial least squares regression (PLS) was proposed as a method to solve collinearity problems in multiple regression/ calibration by Wold et al. (1983), and has gained an increasing popularity, especially among chemometricians, see Martens and Næs (1989), and also the forthcoming special journal issue mentioned above. The method was proposed as an algorithm, in fact, several algorithms were available; two of them are shown to be equivalent in Helland (1988). In the present context the corresponding population algorithm, treated in Helland (1990), is of special interest. It can be formulated as follows, taking the $p + 1$ dimensional variable (x, y) as the point of departure. Replacing variables by datavectors and (co)variances by their sample counterparts here, leads to the ordinary PLS algorithm.

1. Initialization: $e_0 = x - \mu_x, \;\; f_0 = y - \mu_y.$

Next, do for $a = 1, 2, \ldots$:

2. Weights and scores: $w_a = \mathrm{Cov}(e_{a-1}, f_{a-1}), \;\; t_a = e_{a-1}^t w_a.$

3. Loadings by 'least squares':

$$p_a = \mathrm{Cov}(e_{a-1}, t_a)/\mathrm{Var}(t_a), \;\; q_a = \mathrm{Cov}(f_{a-1}, t_a)/\mathrm{Var}(t_a).$$

4. New residuals: $e_a = e_{a-1} - p_a t_a, \;\; f_a = f_{a-1} - q_a t_a.$

Stopping at step $a = k$, this leads to the bilinear representation

$$x = \mu_x + p_1 t_1 + \ldots + p_k t_k + e_k, \;\; y = \mu_y + q_1 t_1 + \ldots + q_k t_k + f_k, \qquad (13)$$

where loadings and scores are determined by the algorithm. The corresponding sample equation is much used in the chemometric literature in the same way as statisticians use factor analysis, and it is often claimed - and in fact also demonstrated - that the corresponding plots are useful for interpreting complex data sets.

Rather than interpretation, our main theme in the present paper is prediction at some set of x-values x^0. Then the natural procedure is to use the partial least squares algorithm to construct new scores t_a^0 from x^0 and in analogy with the last equation in (13) take

$$\hat{y}_{k,PLS} = \mu_y + q_1 t_1^0 + \ldots + q_k t_k^0. \qquad (14)$$

Proposition 4.

The above predictor can be written

$$\hat{y}_{k,PLS} = \mu_y + \beta_{k,PLS}^t (x^0 - \mu_x), \qquad (15)$$

with

$$\beta_{k,PLS} = W_k (W_k^t \Sigma W_k)^{-1} W_k^t \sigma, \;\; \textit{where } \sigma \textit{ here is } \sigma_{xy} \textit{ and } W_k = [w_1, \ldots, w_k]. \qquad (16)$$

Proof.
The corresponding sample case is proved in Helland (1988), Theorem 3.1.

The formula (16) indicates that the space $\mathcal{S}_k$ spanned by the columns of W_k is of some interest. Further progress can be made if we note that this space can also be spanned by a socalled Krylov sequence.

Proposition 5.
As long as w_k is nonzero, an alternative basis for $\mathcal{S}_k$ is given by the vectors

$$\sigma, \Sigma\sigma, \ldots, \Sigma^{k-1}\sigma.$$

Proof.
Compare the proof of Proposition 3.1 in Helland (1988).

Now we can start to make some algebraic conclusions from these results (for more details, see Helland, 1990): First, if there should be some value of k such that the corresponding weight vanishes, i.e., $w_k = 0$, then the PLS algorithm will be terminated at this value (and only at such values). Secondly, by Proposition 5, the first value $k = m$ such that $w_k = 0$ for $k > m$ is characterized by the property that the k vectors of the Krylov sequence will span the same m-dimensional space $\mathcal{S}_m$ also when $k > m$. If this happens, we will say that the population PLS algorithm stops at step m. Of course, the algorithm will always stop at $m = p$; the interesting case is when it stops before that. Note that by Proposition 4, the regression vector $\beta_{k,PLS}$ will always belong to the space $\mathcal{S}_m$.

From these observations there is in fact a short step to the main result of this section. This result gives a fairly precise interpretation of the result of the population PLS algorithm in the light of model reduction and rotational symmetry.

Theorem 3.

The population PLS algorithm stops at step m if and only if the number of strata in the population model with non-zero regression coefficient is equal to m. The resulting PLS regression wector $\beta_{m,PLS}$ will then equal $\beta = \Sigma^{-1}\sigma$.

Main idea of proof.

From (8) a linear dependency in the Krylov sequence is given by

$$\sum_{j=1}^{k} c_j \Sigma^{j-1}\sigma = \sum_{h=1}^{q}\{\sum_{j=1}^{k} c_j \lambda_h^{j-1} P_h\sigma\} = 0.$$

This can happen only if $k \geq m$, where now m is the number of strata with $P_h\sigma$ nonzero. Note that for $k = m$ there is a Vandermonde determinant among the different λ_h involved in this argument.

Thus with this natural specification of the population PLS algorithm, there is a close connection to the optimality results discussed in Section 5 and Section 6 above. Briefly, the termination of this algorithm at $m(< p)$ steps is equivalent to a reduced model of the type which results in a natural way from the orbit behavior of the optimal equivariant regressor under rotation invariance. And the regression vector that results from the corresponding population PLS algorithm is what it should be.

Now, of course, what is used in practice by chemometricians is the sample PLS algorithm, and nothing can be said about this from the above discussion, except a vague indication that it will be sensible if n is not too small and a good stopping rule is used. The stopping rule used in practice for PLS regression is based on cross validation using prediction performance, and has only an indirect link to the population stopping criterion mentioned above: If $w_k = 0$, then one should expect that the prediction at step $k-1$ is so good that this will affect the cross validation.

The opinions on the sample PLS algorithm are still rather varied. Many practicioners are satisfied with the fact that it gives some automatic procedure which empirically seems to give good results in many cases, and where they

also have a possibility to combine prediction will some latent variable analysis. Chemometricians have generalized the algorithm to cases with several y-variables and even to cases with many blocks of variables. On the other hand, sceptical remarks to the procedure have been raised by statisticians, for instance in Frank and Friedman (1993). Butler and Denham (2000) have just shown that the sample PLS procedure, while shrinking globally, has so poor shrinking properties at each factor that it seems impossible that it can satisfy any reasonable optimality property.

Here is where the situation stands today. The PLS algorithm will continue to be used by those who like it, even though it has no optimality properties. What matematical statisticians can hope for, is first to make better evaluations of the method. But a more interesting challenge will be if we can improve the method by using our common way of thinking, namely in terms of optimality. The purpose of the present paper has been to try to go some way in that direction.

By what has been discussed up to now, we have a population procedure which is equivalent to a model reduction with reasonably well defined good properties related to rotation invariance. We have integral formulae for the optimal estimated regression vector given the orbit indices. In addition to finding practical ways of calculating that integral - where MCMC and related techniques probably may help - the remaining challenge is to find good estimates of the remaining orbit index parameters.

8 Estimation of orbit parameters.

In the formula (11), the point $(\Sigma, \beta, \sigma_y^2)$ is a fixed point on some orbit of the rotation group in the parameter space. But by Proposition 3 we may without loss of generality let Σ be diagonal with first diagonal elements $\lambda_1, \ldots, \lambda_m$ here, and then take $\beta = (\gamma_1, \ldots, \gamma_m, 0, 0, \ldots, 0)^t$. In addition to its data dependence, the solution will then depend upon m, these parameters and σ^2. This means that we are left with $2m + 1$ parameters to estimate. This is in general a very small number compared to the $(p+1)(p+2)/2$ parameters in the original

covariance structure.

Note, however, that any value of the parameter $(\Sigma, \beta, \sigma_y^2)$ which belongs to the same orbit (or more generally, to orbits which can be linked to that one through the parameter dependence corresponding to the one described in Proposition 2) will give the same estimate (11).

Taking maximal likelihood under multinormality as a natural approach to the estimation of these parameters, the last remark is a great aid together with the following simple remark: When $\hat{\theta}$ is a unique maximum likelihood estimator for θ, then, for any f which is constant on a subset Γ of the parameter space Θ and one-to-one on $\Theta \backslash \Gamma$, we have that $f(\hat{\theta})$ is the unique maximum likelihood estimator for $f(\theta)$.

Thus to find maximum likelihood estimators of the $2m + 1$ parameters described above, it is enough to give maximum likelihood estimators of the covariance parameters under the hypothesis that the number of strata in Σ with non-zero regression component ('number of relevant components', cp. Næs and Helland, 1993) is equal to m. This latter problem was discussed in Helland (1992).

The maximum likelihood solution found there, can be formulated as follows:

(a) With $A = S_{xx} - s_y^{-2} s_{xy} s_{xy}^t$ and $B = S_{xx}^{-1}$, find a $p \times m$ matrix $R = \hat{R}$ which minimizes

$$G(R) = (|R^t A R| / |R^t R|) \cdot (|R^t B R| / |R^t R|).$$

(b) With $\hat{R}$ as in (a), and taking $\hat{P}_R = \hat{R}(\hat{R}^t \hat{R})^{-1} \hat{R}^t$, the maximum likelihood estimates under the hypothesis of m relevant components are given by

$$\hat{\Sigma} = \hat{P}_R S_{xx} \hat{P}_R + (I - \hat{P}_R) S_{xx} (I - \hat{P}_R),$$

$$\hat{\beta} = \hat{R}(\hat{R}^t S_{xx} \hat{R})^{-1} \hat{R}^t s_{xy},$$

$$\hat{\sigma}_y^2 = s_y^2.$$

Note that $G(R)$ will be the same here if R is replaced by another matrix spanning the same space, similarly for the estimators. Hence what in effect is estimated, is an $m-$dimensional space, the space of relevant components.

In Helland (1992), an approximate algorithm for doing the minimization in (a) was proposed, but the prediction method based upon this algorithm performed rather poorly in simulations done by Almøy (1996). A reasonable conjecture is that performance will be better when combined with the integral (11), since the effective number of parameters estimated by the maximum likelihood part then will be reduced considerably, as discussed above. The computation needed in order to find these predictions will be rather heavy, however, a disadvantage which is particularly important if one wants to use crossvalidation to determine the number of relevant components, the most common procedure. An alternative emerging from the likelihood approach here, is using a likelihood ratio test.

In conclusion here, there are many questions concerning estimation and prediction under the reduced regression model which need further investigations. In addition to questions related to performance under the reduced model itself, the robustness questions in the spirit of the paper by Bickel (1984) will need to be adressed.

In addition to addressing optimality problems of the kind discussed above, it is also of interest to do further investigations on the performance of the ordinary sample partial least squares algorithm, since this is the algorithm used by chemometricians today. Asymptotic expansions, as in Helland and Almøy (1994) is one of several ways of approaching such questions.

9 Classification.

The sample partial least squares algorithm is also used by chemometricians outside the regression context, for instance when doing classification. A large number of applications of this can be found in the chemometrical literature, and the technique has also spread outside the field chemometrics, for instance to psychiatry (Gottfries et al, 1995). It is obvious that techniques of this

kind meet a latent need among researchers with near collinear data, both in regression problems and in classification problems.

In this Section we will discuss model reduction under rotational symmetry in the classification problem using ordinary discriminant analysis as the classification method, and point at similarities with and differences from the regression situation. We will concentrate on the simplest possible case: Two-class discriminant analysis with equal prior for the two classes, equal covariance matrix for the observations, equal cost of misclassification and equal number n of observations in the two classes. We have in mind a situation where the number p of classification variables is rather large, also compared to n.

We also assume a multinormal model, the standard point of departure in discriminant analysis: In sample 1 we have n independent observations which are $N(\mu_1, \Sigma)$, and having mean $\bar{x}_1$; in sample 2 the n observations are $N(\mu_2, \Sigma)$ and have mean $\bar{x}_2$. The standard way of classifying a new observation x is to allocate it to class 1 iff

$$x^t S^{-1}(\bar{x}_1 - \bar{x}_2) - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)^t S^{-1}(\bar{x}_1 - \bar{x}_2) > 0,$$

where S is the pooled covariance matrix from the $2n$ observations.

When p is large or if we have other sources of collinearity, we may want to replace $S^{-1}(\bar{x}_1 - \bar{x}_2)$ by some other observator vector b. Thus x is classified to be of class 1 iff

$$[x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)]^t b > 0,$$

and the probability of misclassification, given the samples, is given by

$$\frac{1}{2}\Phi\left(\frac{(\mu_2 - \frac{1}{2}(\bar{x}_1 + \bar{x}_2))^t b}{\sqrt{b^t \Sigma b}}\right) + \frac{1}{2}\Phi\left(\frac{(\frac{1}{2}(\bar{x}_1 + \bar{x}_2) - \mu_1)^t b}{\sqrt{b^t \Sigma b}}\right),$$

where Φ is the standard normal cumulative distribution function.

We will now simplify notation by assuming $\mu_1 = -\mu_2 = \mu$. Also, it is natural to assume that $\frac{1}{2}(\bar{x}_1 + \bar{x}_2)$ is independent of b, which is true for the original choice of b. Integrating over the probability distribution of this variable, using the fact that $\mathrm{E}(\Phi(\nu - u)) = \Phi(\nu/\sqrt{1+\tau^2})$ when $u \sim N(0, \tau)$, we arrive at a

natural loss function for the classification problem:

$$L(\theta, b) = \Phi\left(-\frac{\mu^t b}{\sqrt{b^t \Sigma b}}\sqrt{\frac{2n}{2n+1}}\right), \tag{17}$$

where $\theta = (\Sigma, \mu)$.

Nearly everything that was done in Section 4 can now be repeated here. In particular, the decomposition (8) of Σ will hold and will have the same interpretation. Under a rotation determined by the ortogonal matrix Q we have

$$(\Sigma, \mu) \to (Q^t \Sigma Q, Q^t \mu),$$
$$(x, b) \to (Q^t x, Q^t b).$$

The loss function L is invariant under the rotation group, and the group is non-transitive in the parameter space $\{(\Sigma, \mu)\}$ with orbit index given by the ordered set of strata (their relative orientation, their dimensions and the corresponding eigenvalues λ_k) and by the norm of the projected components $\gamma_k = \|P_k \mu\|$. This is proved in the same way as Corollary 2.

As a function of b, the loss function (17) takes its minimal value at all vectors of the form $\kappa \Sigma^{-1} \mu$ where $\kappa > 0$. A general feature is that $L(\theta, b)$ is independent of the norm of b. Similarly, there is an invariance of L as a function of θ under scale transformations $(\Sigma, \mu) \to (a^2 \Sigma, a\mu)$. This implies that the estimation problem as such - regarding b as some estimator - is invariant under a larger group than the rotation group. A scale change in b does not change the classification rule, however. In this Section we will therefore stick to the rotation group, which also will help illuminating the connections to the regression problem.

In general, consider a loss function L of the form $f(b^t \mu \mu^t b / b^t \Sigma b)$, where f is an increasing function. With b as an estimator, this gives formally an estimation problem which is invariant under the rotation group. In this general problem we cannot give an explicit formula for the Pitman type estimator, in the way we did in Theorem 2, but the estimator can be found in principle by minimising

$$\int f(b^t Q^t \mu \mu^t Q b / b^t Q^t \Sigma Q b) l(Q, \Sigma, \mu, \{x_{1i}\}, \{x_{2i}\}) d\gamma(Q), \tag{18}$$

where

$$l = \exp\{-\frac{1}{2}\mathrm{tr}(Q^t\Sigma Q)^{-1}[\sum_i (x_{1i} - Q^t\mu)(x_{1i} - Q^t\mu)^t + \sum_i (x_{2i} + Q\mu)(x_{2i} + Q\mu)^t]\},$$

and where again $d\gamma(\cdot)$ is Haar measure for the rotation group, the elements of this group being identified by orthonormal matrices Q.

Assuming that f is differentiable (which it is in the concrete application we have in mind), we find that the equation from which (the direction of) b can be determined takes the form:

$$\int f'(r(\cdot))\,\frac{l(\cdot)}{b^tQ^t\Sigma Qb}Q^t[\mu\mu^t - r(\cdot)\Sigma]Qb\; d\gamma(Q) = 0, \tag{19}$$

with

$$r(\cdot) = \frac{b^tQ^t\mu\mu^tQb}{b^tQ^t\Sigma Qb}.$$

Similarly to the proof of Proposition 2 we have that every parameter dependence is through $Q^t\Sigma Q = \sum_i \lambda_i (Q^te_i)(Q^te_i)^t$ and $Q^t\mu = \sum_i \mu_i (Q^te_i)$, where e_i are the eigenvectors of Σ and μ_i are the corresponding components of μ. Also, when μ has a vanishing component for some stratum, it follows from equation (19) that the corresponding component of b must vanish, and by further inspection: The solution can not depend upon the eigenvalue λ_i of that stratum. Hence the argument of the proof of propositon 2 and the model reduction argument of Section 6 can be repeated in verbatim for the present situation, and we conclude:

For the standard discriminant analysis situation, if we decide to do a Pitman type estimation on orbits of the rotation group, which is the optimal choice there, the only model reduction that can affect the discrimination rule, is given by:

(1) Fix the number m of relevant component, i.e., the number of strata of Σ with nonzero component γ_i of μ.

(2) Possibly also: Fix some symmetric function of $((\lambda_1, \gamma_1, \ldots, (\lambda_m, \gamma_m))$, where λ_j now is the eigenvalue corresponding to stratum j.

The most natural general procedure, having no special information about the problem, is to skip step (2) here. In a similar way as in Section 7 above, this leads to the population version of the partial least squares algorithm for the discriminant analysis problem. Again, of course, the sample algorithm implied by the above argument is different than the sample PLS-algorithm. A necessary part of this procedure will be to solve equation (19) for each dimension. It is an open question if such a procedure can be made practical.

10 Discussion.

One way to look upon the present paper, is that it is a companion paper to Helland (2000b), which is a survey paper on the application of group theory to statistics. This author is convinced that theoretical statistics needs more legs to stand on than just probability theory. Symmetry is a very simple and natural concept which is illuminating for many different aspects of statistics, both theoretical and applied. Group theory gives the mathematics needed to make symmetry considerations precise. In the present paper we have concentrated on the rotation group and on the orbit aspect of the group. Several other applications of group theory to statistics are discussed in Helland (2000b).

Another way to look at the paper is that it is a companion paper to Helland (2000c), an invited survey paper written for chemometricians on what is known from a mathematical statistical point of view on partial least squares regression. Since the present paper is written for mathematical statisticians, it gives an opportunity to be more precise and prove some definite results on the population version of the algorithm. And not least: It gives an opportunity to state some open problems on which more work is needed, work that probably can only be done by the community of mathematical statisticians. This is a field where there has been done very much - in fact useful - work using a lower level of precision, and it is a field full of applications, which can be seen by looking at the two chemometrical journals and several related journals.

The solution proposed in this paper relies heavily on the concept of model reduction. This is a concept for which a proper theoretical statistical theory is

lacking. I have a strong feeling that more can be done towards creating at least elements of such a theory. At present model reduction is just one of several tools that are used nearly daily by applied statisticians and other people doing statistical modelling, a tool that has intuitive appeal to many people, and a tool which is completely avoided by theoretical statisticians. In general, taking such concepts into use - both in theory and in practice - may be one way to escape the danger that statistics - in principle an art that should include every aspect of inductive inference from empirical data - shall degenerate into a purely deductive science.

The main result of this paper is that model reduction under rotational symmetry under reasonably simple conditions is very close to implying the population partial least squares model both for the regression and for the classification case. My main hope is that this result will inspire some start of a closer contact between scientific communities which so far have had nearly completely separate developments.

References.

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

Berger, J.O. (1980) *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

Bickel, P.J. (1984) Parametric robustnes: Small biases can be worthwhile. *Ann. Statistics* **12**, 864-879.

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

Desai, U.B., P. Debajyoti and R.D. Kirkpatrick (1985) *Int. J. Control* **42**, 821-838.

Frank, I.E. and J.H. Friedman (1993) A statistical view of some chemometric tools. *Technometrics* **35**, 109-135.

Gottfries, J., K. Blennow, A. Wallin and C.G. Gottfries (1995) Diagnosis of dementias using partial least squares discriminant analysis. *Dementia* **6**, 83-88.

Gradshteyn, I.S. and I.M. Ryzhik (1994) *Tables of Integrals, Series, and Products.* 5th ed. Academic Press, Boston.

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.

Helland, I.S. (2000c) Some theoretical aspects of partial least squares regression. Submitted to Chemometrics and Intelligent Laboratory Systems.

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.

Hjort, N.L. (1991) Estimation in moderately misspecified models. *Statistical Research Report* No. 8/1991, Department of Mathematics, University of Oslo.

Lehmann, E.L. and G. Casella (1998) *Theory of Point Estimation.* Springer, New York.

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

Nelder, J.A. (1965) The analysis of randomized experiments with orthogonal block structure. I. Block structure and null analysis of variance. *Proc. Royal Soc.* A **283**, 147-162.

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

Snedecor, G.W. and W.G. Cochran (1989) *Statistical Methods.* 8th ed. Iowa State University Press, Ames.

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

Appendix 1: Proof of Proposition 1.

Assume for a fixed $\Sigma = \sum_{k=1}^{q} \lambda_k P_k$ that $Q^t \Sigma Q = \Sigma$, or equivalently $\Sigma Q = Q\Sigma$ for an orthogonal matrix Q. Thus

$$\sum_{i=1}^{q} \lambda_i (QP_i - P_i Q) = 0. \tag{20}$$

Multiplying (20) from the right by P_j and from the left by P_k, respectively, gives

$$\lambda_j Q P_j = \Sigma Q P_j, \tag{21}$$

$$\lambda_k P_k Q = P_k Q \Sigma. \tag{22}$$

Next, assume $j \neq k$ and multiply (21) from the left by P_k and (22) from the right by P_j, giving $\lambda_j P_k Q P_j = P_k \Sigma Q P_j$ and $\lambda_k P_k Q P_j = P_k Q \Sigma P_j$, respectively. Since the righthand sides of the latter equations are equal by assumption, and since $\lambda_j \neq \lambda_k$, it follows that

$$P_k Q P_j = 0 \quad (j \neq k). \tag{23}$$

Inserting this into (21) and (22) then gives

$$\lambda_j Q P_j = \lambda_j P_j Q P_j \quad \text{and} \quad \lambda_j P_j Q = \lambda_j P_j Q P_j.$$

Since $\lambda_j \neq 0$, it follows that $QP_j = P_j Q$, completing the proof that (a) implies (b). The implication from (b) to (a) is trivial.

Applying (b) to $v \in V_j$ shows that Qv satisfies $P_j Qv = Qv$, so $Qv \in V_j$. Thus the transformation given by Q must conserve all the spaces V_j; hence it can only consist of rotations within each single V_j. Again the opposite implication is trivial.