Approaching regression methods through symmetry arguments.

Inge S. Helland *

Abstract

For collinear data we consider regression methods which are equivariant under the rotation group in the $x$-space; in fact, this seems to cover nearly all methods that have been proposed. It is argued that the regression parameter along orbits of the rotation group always can be selected in an optimal way, so any freedom in the choice of method should be confined to the orbit index. Via a Pitman type estimator a first order approximation for the estimated parameter along orbits is found, and principal component regression, partial least squares regression and ridge regression appear as the natural methods under various assumptions. Some light is thrown on the connection between these methods and on the possibility for improvement.

KEY WORDS: Collinearity; Partial Least Squares Regression; Principal Component Regression; Ridge Regression; Rotation; Symmetry.

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

*Department of Mathematics, University of Oslo, P.O.Box 1053 Blindern, N-0316 Oslo, Norway. E-mail: ingeh@math.uio.no
paper is an attempt to introduce some general theory into this area. Explicitly, we will discuss regression methods, but the approach that we use, seems to have potential for extension to other fields.

A rather heuristic statement first: When the number of parameters in a statistical model is 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. An additional argument is that 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.

A natural group to look at in many multivariate situations is the rotation group. In many regression problems it is natural to consider the group of rotations in the $x$-space. We start by characterizing the orbits of this group in the parameter space and the degree of non-exactness of the same group. Next we derive - using an approximation which is valid for reasonably large $n$ - the minimum risk equivariant estimator. It turns out that both ridge regression, principal component regression and partial least squares regression appear in a natural way in this setting.

2 The regression model under rotational symmetry.

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

$$
\begin{pmatrix}
\Sigma_{xx} & \sigma_{xy} \\
\sigma_{xy} & \sigma_y^2
\end{pmatrix}
$$

(1)

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 $\beta$, and we will be interested in prediction problems, assuming a linear conditional expectation $E(y|x) = \alpha + \beta'x$ and a linear predictor $\hat{\alpha} + \hat{\beta}'x$. From this one can see that a natural loss function (cp. Theorem 1 in Helland and Almøy, 1994) is

$$L(\theta; \hat{\beta}) = (\hat{\beta} - \beta)'\Sigma(\hat{\beta} - \beta). \quad (2)$$

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 $\hat{\beta}_{LS} = (X'X)^{-1}X'y$, where $X$ 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 $X'X$ may have some extremely small eigenvalues.

Most of the solutions that have been proposed for these situations: 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, so a natural endeavour 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 orthogonal $p \times p$ matrices $Q$ with determinant $+1$. In the sample space, the group $G$ is given by the transformations $(X, y) \rightarrow (XQ, y)$, which induces the group $\hat{G}$ in the parameter space consisting of the transformations

$$(\Sigma, \beta, \sigma^2_y) \rightarrow (Q'\Sigma Q, Q'\beta, \sigma^2_y). \quad (3)$$

The parametric function $\beta = \beta(\theta)$ is trivially seen to be invariantly estimable: $\beta_1 = \beta_2$ implies $Q'\beta_1 = Q'\beta_2$ for all $Q$. An estimator $\hat{\beta}$ is equivariant if we also have $\hat{\beta} \rightarrow Q'\hat{\beta}$ when $(X, y) \rightarrow (XQ, y)$. For equivariant estimators the loss (2) will be invariant. Since the parameter space is closed under the transformations in $\hat{G}$, 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 estimator that
minimizes 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 \( \mathcal{G} \). When estimating orbit indices (maximally invariant parameters under \( \mathcal{G} \)), 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.

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

We start by decomposing the covariance matrix \( \Sigma \), assuming if necessary that the model reduction has already taken place. Let

\[
\Sigma = \sum_{k=1}^{q} \lambda_k P_k,
\]

where the \( P_k \) are projection matrices upon orthogonal eigenvector spaces \( V_k \) (called *strata*, cp. Nelder, 1965), and where all the \( \lambda_k \) are positive and different (otherwise strata could have been combined). Without loss of generality we can take the \( \lambda_k \)'s in decreasing order. Obviously, \( q \leq p \).

We will assume that \( \Sigma \) has full rank \( p \), which is equivalent to

\[
\sum_{k=1}^{q} \dim(V_k) = \sum_{k=1}^{q} \text{tr}(P_k) = p,
\]

or again to the requirement that the direct sum of the spaces \( V_k \) equals the full \( p \)-dimensional Euclidean space \( \mathbb{R}^p \).

The first question we ask, is what transformations in the rotation group that conserve the matrix \( \Sigma \).
Proposition 1.

The following are equivalent:
(a) $Q'\Sigma Q = \Sigma$.
(b) $QP_j = P_jQ$ 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 the vector space $V_k$.

The proof of this is given in Appendix 1.

The next question is simpler: Which transformations conserve the regression vector $\beta$? From $Q'\beta = \beta$ follows that $\beta$ is an eigenvector of $Q$ with eigenvalue 1, or a more useful characterization: $Q$ is a rotation around the axis determined by $\beta$.

Combining these two results, we get:

Corollary 1.

The transformation $Q$ conserves $(\Sigma, \beta)$ 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 $\beta$ has a non-zero component $\beta_k$ we have that $Q_k$ is a rotation around the axis given by $\beta_k$.

Since $\sigma^2$ is unaffected by the transformations, this corollary characterizes the transformations that conserve $\theta = (\Sigma, \beta, \sigma^2)$, 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 $Q'\Sigma_1Q = \Sigma_2$ instead of the previous $Q'\Sigma Q = \Sigma$. 

5
Theorem 1.

Two parameter values \( \theta_1 = (\Sigma_1, \beta_1, \sigma_{y,1}^2) \) and \( \theta_2 = (\Sigma_2, \beta_2, \sigma_{y,2}^2) \) 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 \( P_k^{(2)} = Q P_k^{(1)} Q \), \((k = 1, \ldots, q)\). In particular, the set of dimensions of \( \{V_k^{(2)}\} \) must match the dimensions of \( \{V_k^{(1)}\} \).
4. For the same \( Q \) we have \( \beta_2 = Q \beta_1 \), or equivalently, \( P_k^{(2)} \beta_2 = Q (P_k^{(1)} \beta_1) \), \((k = 1, \ldots, q)\).
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

1. For the ordered set of strata: Their relative orientation, their dimensions and the corresponding eigenvalues \( \lambda_k \).
2. The norms of the projected regression coefficients \( \gamma_k = ||P_k \beta|| \).
3. \( \sigma_y^2 \).

Proof.

From Theorem 1 it is only left to prove that it is enough to characterize the \( \beta \)-dependence of the orbits by the norms \( \gamma_i \) 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 \( \beta \) within each stratum is arbitrary, and only its norm is fixed.
4 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

\[
S = S_{xx} = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})',
\]

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

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

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, 1998; the essential result is also given in Berger, 1980), the minimum risk estimator, given the orbit, will then be

\[
E_{\text{post}}(\Sigma_{xx})^{-1}E_{\text{post}}(\sigma_{xy}),
\]

where \(E_{\text{post}}\) is the corresponding a posteriori expectation. Using then 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: Proposition 2.

Fix an orbit of the rotation group in the above situation. Then the minimum risk estimator for \(\beta\), given the orbit, is given by

\[
\begin{align*}
&\left(\int Q'\Sigma Q \exp\left[\frac{n}{2\sigma^2} (2\beta'Qs - \beta'QSQ'\beta) - \frac{n}{2} \text{tr}(Q'\Sigma^{-1}QS)\right]d\gamma_Q\right)^{-1} \\
&\left(\int Q'\Sigma \beta \exp\left[\frac{n}{2\sigma^2} (2\beta'Qs - \beta'QSQ'\beta) - \frac{n}{2} \text{tr}(Q'\Sigma^{-1}QS)\right]d\gamma_Q\right),
\end{align*}
\]

where \((\Sigma, \beta)\) correspond to a fixed point on the orbit and \(d\gamma_Q\) is Haar measure indexed by the orthogonal matrix \(Q\). Also, \(\sigma^2 = \sigma_y^2 - \beta'\Sigma\beta = \sigma_y^2 - \sum_k \lambda_k \gamma_k^2\) is the residual variance, a constant of the orbit.
To calculate explicitly the integrals of (5) seems to be possible only in special cases. For instance, when $\Sigma = S = I$, it follows from the last section in Helland (1998) that

$$
\hat{\beta} = \frac{I_{\frac{n}{2}}(\frac{3}{\sigma^2} ||\beta|| \cdot ||s||) ||\beta||}{I_{\frac{n}{2}-1}(\frac{3}{\sigma^2} ||\beta|| \cdot ||s||)||s||},
$$

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

In general we will use the following approximation: For large or moderate $n$ the value of the exponential term in both the numerator and the denominator in (5) will be strongly dominated by the value at the $Q = Q_0$ for which the exponent is a maximum, or by values attained close to this $Q_0$. As a first approximation we can therefore write the MRE-solution as

$$
\hat{\beta} \approx (Q_0^T \Sigma Q_0)^{-1} Q_0^T \Sigma \beta = Q_0^T \beta.
$$

The matrix $Q_0$ can depend on all of $\Sigma, \beta, \sigma^2, S, s$ and $s^2$, but in agreement with the results of the previous section, the resulting $\hat{\beta}$ can only depend upon data and on maximal-invariant parameters.

Our task is now to find the orthogonal matrix (with positive determinant) which minimizes

$$
f(Q) = \text{tr}(Q^T \Sigma^{-1} QS) + \frac{1}{\sigma^2} \beta^T QS \beta - \frac{2}{\sigma^2} \beta^T Q s.
$$

(6)

**Theorem 2.**

The optimal rotation matrix $Q_0$ is given as follows:

First diagonalize $\Sigma$ and $S$; that is, find orthogonal matrices $U$ and $V$ such that $U^T \Lambda U = \Sigma$ and $V^T L V = S$ for diagonal $\Lambda$ and $L$. In these diagonal matrices let $\{\lambda_k\}$ be taken as decreasing, and let the eigenvalues $\{l_i\}$ of $S$ have the ordering as determined in the proof of the Theorem (in the most common cases, this will also be decreasing). Then we can take $Q_0 = V^T O U$, where $O$ is blockdiagonal with blocks corresponding to strata, and gives the following rotation within strata:

Let $P_k$ be the projection upon stratum $V_k$, and let $\beta_k = P_k \beta$, such that $||\beta_k||^2 = \gamma_k^2$. Then, for each $k$, let $\nu_k$ be determined by

$$
||P_k(S + \nu_k I)^{-1} s||^2 = \gamma_k^2,
$$

(7)
and let \( Q_0 \) in each stratum \( V_k \) rotate \( \beta_k \) to \( P_k(S + \nu_k I)^{-1}s \).

In particular, this gives \( \hat{\beta} = Q_0\beta \) as a vector with component \( P_k(S + \nu_k I)^{-1}s \) in each stratum where \( \gamma_k \neq 0 \) and component 0 elsewhere.

The proof of the Theorem is given in Appendix 2.

Note that what we have done in this Section is in fact to maximize a likelihood, but only along a fixed orbit of the parameter group. The motivation for doing this, is given by equation (5).

5 Estimation of stratum parameters; relation to various regression methods.

Recall that from Section 2, the orbit indices, i.e., the maximal invariants with respect to the rotation group in the \( x \)-space, are given by the stratum dimensions \( p_k \) and eigenvalues \( \lambda_k \), the norm of the regression parameters \( \gamma_k = \| P_k \beta \| \) and the \( y \)-variance \( \sigma_y^2 \). The estimation of these parameters will depend upon which assumptions we make about the model. By varying these assumptions, the present framework will allow us to throw new light upon several different regression methods, in particular those in common use for collinear data, where most methods have been proposed using rather ad hoc arguments.

First of all, the distinction between different models disappear if we choose the simplest possible estimate in each stratum, namely estimating each \( \gamma_k^2 \) by \( \| P_k S^{-1}s \|^2 \), which leads to \( \hat{\nu}_k = 0 \) and to the ordinary regression estimate \( S^{-1}s \). Note that this also results from choosing the \( Q = Q_0 \) which minimizes just the last terms in (6). This least squares solution has many good properties, but when \( p \) is large compared to \( n \), some shrinkage may be wanted to reduce the prediction error, and the \( \gamma_k^2 \)-estimates above may be too large. The demand for shrinkage in \( \hat{\beta} \) is here very natural in light of the loss function (2). Other problems, for instance those in classification, where the natural loss function is different, may require other approaches to the solution.
In principle there are two ways to achieve such shrinkage, either putting $\gamma_k = 0$ for some strata, or use a non-zero estimate for $\nu_k$.

With scarce data it is also natural to try to keep the number of parameters in the model reasonably low. In the statistical literature one very common procedure is to reduce the number of $x$-variables included, which can either be looked upon as a choice of conditioning, or an assumption that some components of $\beta$ - on the ordinary scale - can be taken as zero. For some regression applications this is definitively the most meaningful way to reduce the model. Various criteria like AIC or BIC may then help in the choice of model.

The situation that we have in mind in the present paper, however, is different. If the $x$-variables are, say, absorption of light at a large number of different wavelengths, and $y$ is some chemical quantity, it may be less meaningful to assume that only a specified set of wavelengths are correlated to $y$. It may be more reasonable to assume a symmetry of the situation, which can be approximated by the rotational symmetry which is the point of departure of the present paper. Note that this does not imply any symmetry properties of the model parameters, only that the model is closed under rotation and that the loss function (2) is invariant.

Then also different ways to reduce the model become relevant. In particular, it is crucial that the reduced model also is invariant under the rotation, and this can be achieved by making any model reduction on the orbit index parameters.

A natural first assumption then is that the number of strata is reasonably low. A natural second assumption may be that some of the stratum components of $\beta$ vanish, i.e., that $\gamma_k = 0$ for some $k$.

A well known difficult situation with unstable estimators arises if $\gamma_k \neq 0$ for strata with $\lambda_k \approx 0$, but perhaps $\gamma_k = 0$ for some other strata. Fortunately, this situation seems to be rare in well designed studies. So assume $\gamma_k = 0$ for some $k$'s with small $\lambda_k$. If we again take the estimate $\hat{\gamma}_k^2 = \|P_k S^{-1} s\|^2$ for other strata, this leads to a version of principal component regression. Other versions of PCR can be obtained in related ways.

Another solution might be to make restrictions on the parameters $\nu_k$ instead. One might assume that all these are equal: $\nu_1 = \ldots = \nu_q = \nu$ and estimate $\nu$ by ridge trace or in other ways. This leads to $\hat{\gamma}_k^2 = \|P_k (S + \nu)^{-1} s\|^2$, but more
importantly the ridge estimator: \( \hat{\beta} = (S + \hat{\nu})^{-1} s \).

One feature of the regression estimators found in this way should be noted: The dimensions of the strata do not enter anywhere, just the number of strata and the norm of the regression component in each stratum. The reason can be seen from Theorem 2: In each stratum the regression component was rotated in a unique way dictated by the criterium to be minimized.

This leads to the final regression method considered here, and perhaps to the cleanest set of model assumptions made up to now: Just assume that exactly \( m \) \((m \leq q)\) of the strata have \( \gamma_k \neq 0 \). This implies that for some set of indices \( \{ k_j \} \) the regression coefficient can be written as a linear combination

\[
\beta = \sum_{j=1}^{m} e_{k_j} \gamma_{k_j},
\]

where \( e_k \) is a unit vector in stratum \( V_k \), which means that it is an eigenvector of \( \Sigma \) corresponding to the eigenvalue \( \lambda_k \). In other words

\[
\beta = R \gamma,
\]

where \( R \) is a \( p \times m \) matrix whose columns are eigenvectors of the covariance matrix of the \( x \)-variables.

This is just the crucial assumption needed for the partial least squares population model, which has been discussed in a series of papers (Helland, 1988, 1990, 1992; Næs and Helland, 1993).

The corresponding PLS sample estimator, basically found by replacing population (co)variances by sample (co)variances in a natural recursive representation of the PLS \( \beta \), is very much used and have been generalized in many different directions in the chemometrical literature. Typically, the order is found by cross validation. It has recently been shown that the PLS estimator cannot be optimal in any reasonable statistical sense (Frank and Friedman, 1993; Butler and Denham, 1999). The discussion above, where the PLS population model seems to come out in a reasonable way from model reduction under symmetry, seems to suggest that this lack of optimality must be due to inefficient estimation under the model. The task of improving the estimates seems to be non-trivial, however. The full maximum likelihood under the PLS model was found in Helland (1992), is cumbersome to use, and
has turned out to be of rather poor efficiency in some limited simulations (Almøy, private communication).

In the spirit of the present paper it seems more reasonable to use the Minimum Risk Equivariant estimator on orbits as it was found in the previous section, and then use the maximum likelihood of orbit index (maximal invariant) in the sample space to estimate the orbit index in the parameter space. In principle this procedure is always possible, and it is used for another common estimation problem in Helland (1999). In the present case, the maximal invariant from \((X, y)\) under rotation is found by using a singular value decomposition of \(X\), i.e., \(X = ZDV\) with \(D\) diagonal and the other matrices orthogonal. Then the maximal invariant in the sample space is \((Z, D, y)\). Since only asymptotic joint distribution of eigenvectors and eigenvalues can be found, it is impossible to give the likelihood explicitly in this case. It may be possible to get some information from the asymptotic likelihood, but this is outside the scope of the present paper.

A very natural question to ask when we have a situation like this with many similar methods, is if it is possible to find a method which is ‘best’ in some sense. The present approach may point at several directions for further research towards this aim. One possible way to proceed might be through the quantities \(\nu_k\). One way to look upon the PCR and the PLS models, is that one assumes \(\nu_k = +\infty\) for some strata and \(\nu_k = 0\) for the rest (specified through the size of eigenvalues for PCR, unspecified for PLS). For the ridge regression model we have that \(\nu_k = \nu\) is constant. From this perspective it is of course natural to search for other ways to choose \(\{\nu_k\}\).

In fact, by taking the expectation over \(y\) of the loss function (2) (perhaps with \(\Sigma\) replaced by \(S\)), fairly managable expressions result from which one might hope to get some help in the process of choosing \(\{\nu_k\}\). There is a difficulty with this approach, however: In the present setting, it is really the orbit index parameters \(\gamma_k = \|P_k\beta\|\) which should be estimated. The link to \(\nu_k\) is data-dependent; in particular, it depends upon \(y\) through \(s\); see equation (7).
6 Final remarks.

The purpose of this paper has been to illustrate how rotational symmetry can be utilized in studying regression models, and which estimators that emerge from this. From the discussion above, both ridge regression, principal component regression and partial least squares regression seem to belong to the reasonably 'optimal' rotation-equivariant family of regression methods, and the relation between these methods seem to have been clarified to some extent by using rotational symmetry. Other methods, like the continuum regression of Stone and Brooks (1990) do not in general belong to the family. However, as shown by Sundberg (1993), one step continuum regression is closely related to ridge regression, and can thus be taken as member of the family.

It is sometimes said that in practical terms the difference between the different regression models for collinear data is so small that it does not matter much which method you use. The difficulty with such statements is that even though they are based upon experience, by its nature, experience will always be limited. The fact that the most well known regression methods belong to the same 'optimal' class, as discussed in this paper, is encouraging, but further research in this area will be needed. Comparison of regression methods by simulation was done in Almøy (1996); see also references there. In Helland and Almøy (1994) we also made analytical comparisons of various regression methods. The calculations made there are based on the loss function (2), and may complement what was found in the present paper.

The method used in this paper is also in principle applicable to several multivariate methods, where the problem with many parameters and relatively scarce data may be as acute as in the regression case. We hope to be able to discuss some of this in a later publication.

References.


**Appendix 1: Proof of Proposition 1.**

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

$$\sum_{k=1}^{q} \lambda_k (QP_k - P_k Q) = 0. \quad (8)$$

Multiplying (8) from the right by $P_j$ and from the left by $P_k$, respectively, gives

$$\lambda_j Q P_j = \Sigma Q P_j, \quad (9)$$

$$\lambda_k P_k Q = P_k Q \Sigma. \quad (10)$$

Next, assume $j \neq k$ and multiply (9) from the left by $P_k$ and (10) 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

$$P_k Q P_j = 0 \quad (j \neq k). \quad (11)$$

Inserting this into (9) and (10) 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.

**Appendix 2: Proof of Theorem 2**

Let first $L = \text{diag}(l_1, l_2, \ldots, l_p)$ with $l_1 \geq \ldots \geq l_p > 0$, assuming that $S$ has full rank $p$. Defining $O = U Q V'$ (which is orthogonal), and taking $\alpha = U \beta$ and $t = V s$, the task is to minimize

$$\sigma^2 \text{tr}(\Lambda^{-1} O L O') + (Q \alpha - L^{-1} t)' L (Q \alpha - L^{-1} t) - t' L^{-1} t. \quad (12)$$

As a preliminary step, consider the minimization of the first term here. We assume as in the previous section that $\Sigma$ has $q$ strata, and that the dimensions of
these are \( p_1, p_2, \ldots, p_q \) with \( p_1 + \ldots + p_q = p \), and with corresponding eigenvalues \( \lambda_1 > \ldots > \lambda_q > 0 \). This means that the matrix \( A \) has diagonal elements \( \lambda_k \) with multiplicities \( p_k \). By \( O \) we shall now rotate \( L \) to \( A = OLO' \) such that \( \sum A_{ii}^{-1}A_{ii} \) is as small as possible. This means that the largest term \( A_{pp}^{-1} = \lambda_q^{-1} \) must have as small weight as possible, thereafter \( \lambda_{q-1}^{-1} \) have as small weight as possible from the remaining terms and so on. The way to achieve this, is to let \( A \) be block diagonal in accordance with strata. More specifically:

Let \( O \) be such that the union of the eigenspaces corresponding to \( l_1, \ldots, l_{p_1} \) equals stratum \( V_1 \), the union of the eigenspaces corresponding to \( l_{p_1}, \ldots, l_{p_1+p_2} \) equal to \( V_2 \) and so on until the union of the eigenspace corresponding to the smallest eigenvalues \( l_{p-p_q+1}, \ldots, l_p \) equals the stratum \( V_q \) with smallest \( \Sigma \)-eigenvalue \( \lambda_q \). One possibility is to take \( A = L \), but any rotation within strata will give the same solution.

Now turn to the minimization of the sum of the two relevant terms in (12). Then it is necessary also to divide the second term into strata, since the norms \( \gamma_k \) of the stratum components of the regression coefficient are fixed on the orbits. Write \( i \in S_k \) for the indices in stratum \( k \), i.e., for \( p_1 + \ldots + p_k + 1 \leq i \leq p_1 + \ldots + p_{k+1} \). The second term in (12) will then be:

\[
\sum_{k=1}^{q} \sum_{i \in S_k} l_i (O' \alpha - L^{-1} t)_i^2.
\] (13)

The first thing to do, is to find the optimal rotation within each stratum. The norm of stratum component \( k \) of \( \beta \) is a fixed invariant \( \gamma_k \); hence the norm of the stratum component \( k \) of \( \alpha \) is \( \gamma_k / \sigma \). We need an auxiliary result, which is proved by straightforward use of Lagrange multiplicator:

**Lemma 1.**

Minimization of \( \sum_i k_i (z_i - b_i)^2 \) under the constraint \( \sum i z_i^2 = c^2 \) has the solution

\[
z_i = \frac{k_i b_i}{k_i + \nu},
\]

where \( \nu \) is determined from the constraint.
Applying this to the minimization of (13) by rotating within each stratum, we find
\[ (O'\alpha)_i = \frac{t_i}{l_i + \nu_k}, \]  
where \( \nu_k \) is determined from
\[ \sum_{i \in S_k} \frac{t_i^2}{(l_i + \nu_k)^2} = \gamma_k^2. \]  
Note that, since the left-hand side here decreases from \(+\infty\) to 0 as \( \nu_k \) goes from \(-\min(l_i)\) to \(+\infty\), this equation always has a solution.

Furthermore, the minimal value of (13) will be
\[ \sum_k \nu_k^2 \sum_{i \in S_k} \frac{t_i^2}{l_i(l_i + \nu_k)^2}. \]

The last step of the minimizing procedure is then to see if the ordering of the \( S \)-eigenvalues should be changed in order that the minimum sum of the two terms should be as small as possible, that is, The eigenvalues \( \{l_i\} \) should be ordered such that
\[ \sum_k \sigma_k^{-2} \lambda_k^{-1} \sum_{i \in S_k} l_i + \sum_k \nu_k^2 \sum_{i \in S_k} \frac{t_i^2}{l_i(l_i + \nu_k)^2} \]
should be as small as possible. Since we always have
\[ cs' s^{-1} s < \sum_k \nu_k^2 \sum_{i \in S_k} \frac{t_i^2}{l_i(l_i + \nu_k)^2} < s' s^{-1} s \]
for some nonnegative constant \( c \) depending upon the minimal value of \( \nu_k \) and of the maximal value of \( l_i \), the inclusion of the last term will probably not change the ordering in many cases.

Finally, allow the case where \( S \) is singular, that is, for some strata \( S_k \) there may be indices \( l_i \) with \( l_i = 0 \). A straightforward extension of the Lagrange multiplier argument, using the original expression found from (6) for these indices, show that (14) and (15), and hence the Theorem, also hold when terms with \( l_i = 0 \) are included. However, the last discussion has to be amended, and the solution \( \nu_k = 0 \) is impossible for such strata.

17