Qeios

On Optimal Linear Prediction

Inge S. Helland
Department of Mathematics, University of Oslo

Corresponding author: Inge S. Helland, ingeh@math.uio.no

Abstract

The main purpose of this article is to prove that, under certain assumptions in a linear prediction setting, optimal methods based upon model reduction and even an optimal predictor can be provided. The optimality is formulated in terms of the mean square prediction error. The optimal model reduction turns out, under a certain assumption, to correspond to the statistical model for partial least squares discussed by the author elsewhere, and under certain specific conditions, a partial least squares type predictor is proved to be good compared to all other predictors. In order to prove some of the results of this article, techniques from quantum theory are used. Thus, the article is based upon a synthesis of three cultures: mathematical statistics as a basis, algorithms introduced by chemometricians and used very much by applied scientists as a background, and finally, mathematical techniques from quantum theory to complete some of the proofs.

1. Introduction

There exists a large number of different statistical methods for the linear prediction of a single variable y from p variables $x_1, \ldots, x_p$. The user of statistics is often left to choose the method that is familiar to him or her, or the method for which he/she has access to the relevant software. When p is small, multiple linear regression is the method to choose, but in many practical applications, p is large, often larger than the number n of units where data is available.

For this situation, partial least squares (PLS) regression is a method that is emerging and recommended also by some statisticians. The method was developed by chemometricians and has grown popular among very many applied researchers. It was linked to a statistical model by Helland[1], see also Helland[2][3], Næs and Helland[4], and Helland and Almøy[5]. The model was generalized to the case of multivariate y and tied to Dennis Cook's envelope model in Cook et al.[6]. More recently, Cook and Forzani[7] have studied the asymptotics of PLS regression as n and p tend to infinity and have given strong evidence that this should be the method of choice in the case of abundant regression, where many of the predictors x_i contribute information about the response y.

https://doi.org/10.32388/I2OXKX

For the present article, the motivation for PLS regression and related methods that is advocated in Helland et al.[8] is particularly relevant. Here the random x model is the point of departure, and a reduced model is approached through the group of rotations of the eigenvectors for the x-covariance matrix together with scale transformations for the regression coefficients. Below, this group and the orbits of the group will play a fundamental role.

A completely different area where multiple linear regression may be one of the building blocks is machine learning, an important part of artificial intelligence. There is a large literature on machine learning and also a growing literature on the connections between artificial intelligence and statistical modeling. Of special interest for the present article is that there recently have been several investigations related to links between machine learning and quantum-mechanical models, see the review article by Dunjko and Briegel[9]. Given these investigations and the strong link between machine learning and statistics, it is strange that there has up to now been very little published research on possible links between statistical modeling and quantum mechanics. It is one purpose of the present article to discuss such a link. See also parts of the book by Helland[10].

After that book was finished, I have written several articles on the foundation of quantum theory, and some of these have been published in leading physics journals. My final approach towards the foundation is now given in Helland[11][12][13]. This is also part of the basis for the present article.

The plan of the article is as follows: In Section 2, I summarize the theory proposed in Helland[11], giving an alternative foundation of quantum theory. In the rest of the article, this is used in a statistical setting. In Section 3, a corresponding brief introduction to statistical inference is sketched, a setting where statisticians have the choice between two model reductions, as specified by the parameters θ and η. In Section 4, the setting is specified to linear prediction, and a specific model reduction θ is defined in relation to a chosen dimension m, the model reduction which, according to my earlier articles on this topic, gives the statistical model corresponding to partial least squares (PLS) regression. This model is elaborated on in Section 5, where it is shown to correspond to a concrete quantum-mechanical setting, giving an operator A^{θ} corresponding to θ, an operator defined on a Hilbert space $\mathcal{K}$.

A main aim of this paper is to show that the PLS model is optimal in a very concrete sense, and for this purpose, a general model reduction η with dimension m is defined in Subsection 5.2. Then, in Section 6, the first optimality theorem, Theorem 6, is stated, interpreted, and proved. Section 7 discusses estimation under the model which under certain assumptions is proved to be optimal. In Section 8, a quantum theory for data is presented. In Section 9, a possible basis for discussing the optimality of PLS-type regression compared to other methods like ridge regression is discussed. The discussion is completed by giving concrete criteria in Sections 10 and 11, and in Section 12, some concluding remarks are given.

Readers who are only interested in the optimality properties of partial least squares regression may concentrate on the theorems of the last few sections.

2. On quantum foundation

The fundamental notion in my approach towards quantum foundations is that of a theoretical variable connected in a given context to an observer or to a communicating group of observers. In Helland[11][12][13], these variables were mostly physical variables. Later in the present article, they will be statistical parameters or observables relative to some model. The essence of the theory turns out to be the same.

Theoretical variables are divided into *accessible* and *inaccessible* variables. In Helland[11], a physical variable was said to be accessible if the observer(s) in principle in some future can obtain as accurate values of this variable as he/she/they wish to. In this article, I will first let the variables be statistical parameters, and let the inaccessible variables be parameters that are too extensive to be estimable with the available data, and the accessible variables be parameters that can be estimated. Again, the theory from Helland[11] can be adapted. From a mathematical point of view, I require that if λ is an accessible variable and $\theta = f(\lambda)$ for some function f, then θ is also accessible.

One strong postulate from Helland[11], however, is crucial for the theory there. I assume that there is a big inaccessible variable ϕ, varying in a space Ω_ϕ, and I assume that all accessible variables can be seen as functions of this ϕ. In the present article, the inaccessible parameter ϕ will be given a very concrete definition.

Given this postulate in Helland[11], the theory is completely rigorous from a mathematical perspective. The following two theorems are then derived:

Theorem 1. *Consider a context where there are two different related maximal accessible variables θ and η. Assume that both θ and η are real-valued or real vectors, taking at least two values. Make the following additional assumptions:*

(i) On one of these variables, θ, there can be defined a transitive group of actions G with a trivial isotropy group and with a left-invariant measure ν on the space Ω_θ.

(ii) There exists a unitary multi-dimensional representation $U(\cdot)$ of the group behind the group actions G such that for some fixed $|\theta_0\rangle$ the coherent states $U(g)|\theta_0\rangle$ are in one-to-one correspondence with the values of g and hence with the values of θ.

Then there exists a Hilbert space $\mathcal{H}$ connected to the situation, and to every (real-valued or vector-valued) accessible variable there can be associated a symmetric operator on $\mathcal{H}$.

For conditions under which a symmetric operator is self-adjoint/ Hermitian, see Hall[14].

Theorem 2. *Assume that the functions $\theta(\cdot)$ and $\eta(\cdot)$ are permissible with respect to a group K acting on Ω_ϕ. Assume that K is transitive and has a trivial isotropy group. Let $T(\cdot)$ be a unitary representation of K such that the coherent states $T(t)|\psi_0\rangle$ are in one-to-one correspondence with t. Then for any transformation $t \in K$ the operator $T(t)^\dagger A^\theta T(t)$ is the operator corresponding to θ' defined by $\theta'(\phi) = \theta(t\phi)$.*

In addition, if θ and η are different, but related through a transformation k of Ω_ϕ, there is a unitary operator $S(k)$ such that $A^\eta = S(k)^\dagger A^\theta S(k)$.

The two theorems require some definitions.

Definition 1. *The accessible variable θ is called maximal if the following holds: If θ can be written as $\theta = f(\psi)$ for a function f that is not surjective, the theoretical variable ψ is not accessible. In other words: θ is maximal under the partial ordering defined by $\alpha \leq \beta$ iff $\alpha = f(\beta)$ for some function f.*

Definition 2. *Let θ and η be two maximal accessible variables in some context, and let $\theta = f(\phi)$ for some function f. If there is a transformation k of Ω_ϕ such that $\eta(\phi) = f(k\phi)$, we say that θ and η are related (relative to this ϕ). If no such k can be found, we say that θ and η are non-related relative to the variable ϕ.*

Definition 3. *The accessible variable θ is called permissible with respect to the group K acting on Ω_ϕ if the following holds: $\theta(\phi_1) = \theta(\phi_2)$ implies $\theta(t\phi_1) = \theta(t\phi_2)$ for all group elements $t \in K$.*

The point of Definition 3 is that when $\theta(\cdot)$ is permissible, it can be shown[15] that K induces a group G on Ω_θ such that $k \in K$ is mapped to $g \in G$ by

$$g\theta(\phi) = \theta(k\phi). \qquad (1)$$

This mapping between groups is a homomorphism.

From these two theorems, much of the formalism of quantum theory follows. Some of the discussion in Helland[11] is concentrated on the case where the accessible variables take a finite number of values, where it is indicated that the groups G and K and the transformation k can be explicitly constructed. As a consequence, we find from the assumed model:

- Every accessible variable has a self-adjoint operator connected to it.
- The set of eigenvalues of the operator is equal to the set of possible values of the variable.
- An accessible variable is maximal if and only if all eigenvalues of the corresponding operator are simple.
- The eigenvectors can, in the maximal case, be interpreted in terms of a question together with its answer. Specifically, this means that in a context with several variables, a chosen maximal variable θ may be identified with the question 'What will θ be if we measure it?' and a specific eigenvector of A^θ, corresponding to the eigenvalue u may be identified with the answer '$\theta = u$'.
- In the general case, eigenspaces have the same interpretation.
- The operators of related variables are connected by a unitary similarity transformation.

For the discussion of the present article, it is sometimes interesting to let the accessible variables be continuous parameters, real-valued or vector-valued. Then G, K and k need to be defined. However, continuous parameters may be approximated by finite-valued parameters.

Going back to the physical theory, two examples of continuous maximal accessible variables are the position and momentum of a particle. They are maximal by Heisenberg's inequality. In general, different maximal accessible variables are also called - following Niels Bohr - complementary. Complementary variables play a major role in Helland[10].

The general mathematical quantum theory of continuous variables is complicated. Most textbooks concentrate on the case of discrete variables. In Helland[10], the quantum theory of position and momentum is approached by first discretizing the theoretical variable position, and such an approach is always possible. This approach can also be used when considering the probabilities of quantum mechanics, as given by Born's formula.

So consider two discrete theoretical variables θ and η, and let these both be thought of as maximal accessible variables for some observer C. For simplicity, let θ and η take a finite number of values. Then by Theorem 1, we can find self-adjoint operators/ matrices A^{θ} and A^{η} connected to these variables, and possible state vectors are eigenvectors of these matrices.

Let v be a normalized eigenvector of A^{θ} corresponding to the eigenvalue v. Then according to the theory above, if θ is maximal as an accessible variable, this eigenvalue is simple, and the state vector v can be given the following interpretation: We have measured the variable θ and obtained the value $\theta = v$. And by the maximality of the variable, this is the maximal information we can get from the relevant physical system at some given time.

Now assume that we later want to measure a complementary variable η on the same system, and we ask: What is the probability of obtaining the answer $\eta = u$? This event may then be interpreted by the normed eigenvector u of A^{η} which corresponds to the single eigenvalue u. And Born's formula says:

$$P(\eta = u|\theta = v) =|u \cdot v|^2. \qquad (2)$$

In this article, a more general form of Born's formula is sometimes needed; see Helland[13] and Hall[14]. Let now θ and η be continuous parameters, varying in some r-dimensional space S. Let A^{θ} be the operator in the Hilbert space $\mathcal{H}$ which, according to Theorem 1, is associated with the parameter θ. Then A^{θ} has a spectral decomposition

$$A^{\theta} = \int_S v\, dE^{\theta}(v), \qquad (3)$$

where $\{E^{\theta}\}$ is a projection-valued measure, a set of projection operators satisfying

$$\int_S d\, E^{\theta}(v) = I, \qquad (4)$$

a resolution of the identity.

Later in this article, I will work with Hilbert spaces having a finite basis. This simplifies the discussion, and it corresponds to the version of quantum mechanics that one finds in most textbooks. From a statistical point of view, it corresponds to parameters having a discrete set of values, which may seem unusual. It is useful to know, however, that a theory with continuous parameters may always be approximated by a theory with discrete parameters.

With a discrete orthonormal basis $v_1, v_2, \ldots$ of the Hilbert space $\mathcal{H}$, (3) simplifies to

$$A^\theta = \sum_{k=1}^{k_n} \theta_k \, v_k v_k^\dagger, \qquad (5)$$

assuming that A^θ is orthogonal in this basis and has eigenvalues θ_j, the possible values of θ. And (4) is just

$$\sum_{k=1}^{k_n} v_k \, v_k^\dagger = I. \qquad (6)$$

Let us now assume that we have some knowledge of the parameter θ, either in the form of some prior or posterior distribution, or in the form of a confidence distribution (see Schweder and Hjort[16]). Let the density of this distribution be $p_\theta(u)$. Then, in the continuous case, in the language of quantum mechanics, this knowledge can be expressed in the form of a density operator

$$\rho^\theta = \int_S p_\theta \, (v) dE^\theta(v). \qquad (7)$$

This is, in general, a positive operator with trace 1.

Also, of course, a similar definition can be given for the density operator ρ^η connected to the parameter η.

The general Born formula can now be given in the form

$$P(\theta \in C|\rho^\eta) = \mathrm{trace}\big(\rho^\eta \Pi_C^\theta\big), \qquad (8)$$

where, for any Borel set $C \subseteq S$, we have defined the projection operator

$$\Pi_C^\theta = \int_C d \, E^\theta(v). \qquad (9)$$

As corollaries, we have the formulas

$$\mathrm{E}(\theta|\rho^\eta) = \mathrm{trace}\big(\rho^\eta A^\theta\big), \qquad (10)$$

and for any integrable function f, we have

$$\mathrm{E}(f(\theta)|\rho^\eta) = \mathrm{trace}\left(\rho^\eta f\big(A^\theta\big)\right), \qquad (11)$$

where

$$f\big(A^\theta\big) = \int_S f \, (v) dE^\theta(v). \qquad (12)$$

All these formulas are easier to formulate and to understand from the point of view of discrete parameters. Given that the operator A^{θ} is orthogonal in the basis $\{v_k\}$, they read:

$$\rho^{\theta} = \sum_{k=1}^{k_n} p\,(\theta = \theta_k) v_k v_k^{\dagger}. \qquad (13)$$

Again, a similar definition can be given for the density operator ρ^{η} connected to the parameter η, but then a different basis $\{u_j\}$ must be used.

The general Born formula is again in the form

$$P(\theta \in C|\rho^{\eta}) = \mathrm{trace}\big(\rho^{\eta} \mathit{\Pi}_C^{\theta}\big), \qquad (14)$$

where we have defined the projection operator

$$\mathit{\Pi}_C^{\theta} = \sum_{k \text{ such that } \mathrm{\theta_k \in C}} v_k v_k^{\dagger}. \qquad (15)$$

As corollaries, we again have the formulas

$$\mathrm{E}(\theta|\rho^{\eta}) = \mathrm{trace}\big(\rho^{\eta} A^{\theta}\big), \qquad (16)$$

and for any function f, we have

$$\mathrm{E}(f(\theta)|\rho^{\eta}) = \mathrm{trace}\left(\rho^{\eta} f\big(A^{\theta}\big)\right), \qquad (17)$$

where

$$f\big(A^{\theta}\big) = \sum_{k=1}^{k_n} f\,(\lambda_k) v_k v_k^{\dagger}. \qquad (18)$$

There is also a version of Born's formula for the prediction of future data. Assume continuous data, that we have a statistical model with density $p(x|\theta)$, and define first

$$A^{x} = \int_S p\,(x|u) dE^{\theta}(u) \quad \left[A^{x} = \sum_{k=1}^{k_n} p\,(x|\theta = \lambda_k) v_k v_k^{\dagger}\right]. \qquad (19)$$

Then from (11) [(17)] it follows that

$$\mathrm{E}_{\theta}\mathrm{E}_{x}(\psi(x)|\rho^{\eta}) = \mathrm{E}_{\theta}\left(\int_x \psi\,(x) p(x|\square|\theta)|\rho^{\eta}\right) = \int_x \psi\,(x) \mathrm{trace}(\rho^{\eta} A^{x}) \qquad (20)$$

for any integrable function ψ.

These formulas, in particular (2), (8), and (14), define quantum probabilities. It can be shown that quantum probabilities do not satisfy Kolmogorov's axioms. The law of total probability does not hold; there is an additivity of probability amplitudes (as expressed by the vectors in (2)), not of probabilities. Also, the probability of the intersection of two events may depend on the ordering of the events.

There are many approaches towards the proof of Born's formula in the literature, see Campanella et al.[17]. In Helland[10][13], it is related to an observer C who believes in the likelihood principle and, in addition, has certain ideals that he looks up to. These ideals are modeled by a higher concrete or abstract actor D which is seen by C to be perfectly rational, as expressed by

The Dutch Book Principle. *No choice of payoffs in a series of bets shall lead to a sure loss for the bettor.*

The situation for an observer C as described above is called a rational epistemic setting in Helland[10].

Theorem 3. [10][13]. *Assuming a rational epistemic setting, the Born formula holds. The probabilities can be thought of as calculated by the actor D*

It is shown in Helland[10][11][12][13] that a complete foundation of quantum theory can be based upon the above 3 Theorems.

3. An addition to statistical inference theory; two statisticians

Consider two different statisticians A and B. For the purpose of this article, let B be a very experienced statistician, and let him be a Bayesian with a very open mind. Let A be some statistician who is inspired by some ideals; without much loss of generality, we assume that this can be modeled by B. As a basis for their joint thinking, let there be a concrete statistical problem with a data set $\mathcal{X}$. We start with the problem as formulated in the mind of B. He thinks of a large parameter ϕ varying in some space Ω_{ϕ}. But instead of starting with only a concrete prior distribution, he assumes some symmetry in the space Ω_{ϕ} as expressed by a transitive group K acting on this space. For simplicity, he also assumes that K has a trivial isotropy group. Then there is a one-to-one correspondence between the elements $t \in K$ and the points $\phi \in \Omega_{\phi}$.

Under certain mathematical conditions, this structure induces a measure ν on Ω_{ϕ}. Concretely, we will assume that the group K is what is called proper and locally compact. Wijsman[18] called such a group proper if every inverse image of compact sets under the function $(t, \phi) \mapsto (t\phi, \phi)$ is compact. He then proved the following general theorem (cp. Theorem 1 in Helland[11]):

Theorem 4. *The right-invariant measure ν on Ω_{ϕ} exists if the action of K on Ω_{ϕ} is proper and the group is locally compact. There also exists a left-invariant measure.*

In fact, in this situation where there is a one-to-one correspondence between K and Ω_{ϕ}, we have a simpler result: There always exists a left-invariant Haar measure on K (if the group is a locally compact Hausdorff topological group); in this case, we can let this be an invariant measure on Ω_{ϕ}.

B will now take ν as his prior. This may be an improper prior, but as shown by Taraldsen and Lindqvist[19], such priors may, under some well-defined conditions, give proper posterior distributions. More specifically, a necessary and sufficient condition is that $f(x) = \int f(x|\phi)\nu(d\phi)$ is finite for (almost) all x, where $f(x|\phi)$ is the density of B's data model.

Now A sees B as his ideal, and he considers B to be perfectly rational. But, to him, the big parameter space Ω_ϕ is too extensive. He will consider a smaller parameter θ, a function of ϕ. However, in the initial stage, he does not quite know how to choose θ.

Consider a group G acting on θ. In a similar way as above, we can assume that the action of G on Ω_θ is proper, and that the group is locally compact. Then there exists a right-invariant measure μ on Ω_θ, and if necessary, A can take μ as his prior.

In the next section, I will make this more concrete by considering a linear prediction problem with a large number of predictor variables $x_1, \dots, x_p$, and let β, which now can be seen as a function of B's large parameter ϕ, be the theoretical regression parameter with respect to all these variables.

Go back to the decision situation for A as described above. Assume that he has the choice between two parametric functions to estimate, θ and η, and assume that both these are maximal for A in the sense that he is not able to estimate a larger parametric function ξ such that θ can be seen as a function of ξ, say.

4. A setting for linear prediction

Consider a statistical setting with a large number p of possible predictor variables $x = (x_1, \dots, x_p)'$ and a response y. Assume that these variables have a joint distribution, and that we have observed n samples from this distribution.

This introduces the following parameters: $\Sigma_{xx} = \text{Cov}(x)$, $\sigma_{xy} = \text{Cov}(x, y)$, $\sigma^2 = \text{Var}(y|x) = \text{Var}(y) - \sigma_{xy}'\Sigma_{xx}^{-1}\sigma_{xy}$ and $\beta = (\beta_1, \dots, \beta_p)' = \Sigma_{xx}^{-1}\sigma_{xy}$. Let the collection of these parameters be denoted by ϕ.

Let us assume that we know a new vector x_{new} with the same distribution as x, and want to predict the y corresponding to this vector. By well-known statistical theory, see Hastie et al.[20], the best linear predictor, if β is known, is given by $\hat{y} = \beta \cdot x_{new}$.

Here, without loss of generality, I have assumed that y and all the vectors x are centered to zero expectation.

Go back to the statistician A from the previous section. He has data X, y, consisting of n samples from the above distribution, and wants to estimate β. Since p is large and n may be moderate, the above set of parameters may be too large for him. He may consider two estimators $\hat{\theta}$ and $\hat{\eta}$, both based upon parameter reduction.

Specifically, the estimator $\hat{\theta}$ is based on the following model reduction.

Let $d_1, \dots, d_p$ be the normalized eigenvectors of Σ_{xx}, and consider the decomposition

$$\beta = \sum_{j=1}^{p} \gamma_j \, d_j. \qquad (21)$$

In agreement with the PLS model in Helland et al.[8] and the envelope model of Cook et al.[6], fix a number m, and consider estimation/prediction under the hypothesis:

H_m: *There are exactly m nonzero terms in (21).*

There are two mechanisms by which this number of terms can be reduced: 1) Some γ's are zero at the outset. 2) There are coinciding eigenvalues of Σ_{xx}, and then the eigenvectors may be rotated in such a way that there is only one in the relevant eigenspace that is along β.

Considering H_m as a model reduction, it is shown in Helland[1] and Cook et al.[6] that it can be formulated in the following equivalent way: Let $\theta = \theta_m$ be defined by the Krylov set $\sigma_{xy}, \Sigma_{xx}\sigma_{xy}. \Sigma_{xx}^2\sigma_{xy}, \ldots, \Sigma_{xx}^{m-1}\sigma_{xy}$, then m is the smallest number such that β is a linear function of θ_m.

For the purpose of this article, however, we will define $\theta = (\gamma_1 d_1, \ldots, \gamma_m d_m)$, with all $\gamma_i \neq 0$, and define the model under H_m:

$$\beta = \beta_m = \beta(\theta) = \sum_{j=1}^{m} \gamma_j \, d_j. \qquad (22)$$

Note that (21) is invariant under permutations of the terms, so we might as well take the non-trivial terms to be the first m terms.

The model reduced by the hypothesis H_m is equivalent to the PLS model of Helland[1], and is a special case of the envelope model of Cook[21].

It is interesting that this model reduction may be connected to a particular group K acting on the parameter β, also involving Σ_{xx}, that is, a group on the parameter space Ω_ϕ:

Definition 4. *Let the group K be defined by orthogonal matrices acting on all the vectors d_j in (21), and in addition separate scale transformations of the parameters γ_j: $\gamma_j \mapsto g_j(\gamma_j)$ for some bijective continuous functions g_j on the line.*

The first part is equivalent to orthogonal transformations of Σ_{xx}. It can be induced by rotating the vector x, and in addition by changing the sign of this vector.

Theorem 5. *If and only if the bijective continuous functions g_j are such that $g_j(0) = 0$, the orbits of the group K are determined by: a given m and the hypothesis H_m.*

Proof. If and only if $g(0) = 0$, the group on γ defined by $\gamma \mapsto g(\gamma)$ has two orbits: 1) the single value $\gamma = 0$; 2) the set of all γ such that $\gamma \neq 0$. Going to the whole group K, this implies an orbit where $p - m$ of the γ_k's are zero and m of the γ_k's are non-zero. That is, exactly the hypothesis H_m. ■

Definition 5. *Define the group G acting on θ by orthogonal transformations of the vectors d_j in (22) and in addition separate linear scale transformations of the parameters γ_j: $\gamma_j \mapsto \alpha_j \gamma_j$ with $\alpha_j > 0$.*

Taking into account that the changes of sign $\gamma_j \mapsto -\gamma_j$ may also be obtained by orthogonal transformations of the d_i's, this implies that the group G is transitive, and it also has a trivial isotropy group. The elements $g \in G$ are then in one-to-one correspondence with the values of θ.

5. A quantum-mechanical setting related to the model reduction θ

5.1. The group

Let θ in any case be a function of the nonzero parameters $\gamma_1, \dots, \gamma_m$ and the Σ_{xx}-eigenvectors $d_1, \dots, d_m$, all normalized: $\theta = (\gamma_1 d_1, \dots, \gamma_m d_m)$. The elements of the group G are given by 1) a matrix O with orthonormal columns such that $(d_1, \dots d_m) \mapsto O(d_1, \dots, d_m)$; 2) positive scalars α_j giving scale transformations $\gamma_j \mapsto \alpha_j \gamma_j$.

The (right-)invariant measure of the scale transformation $\gamma \mapsto \alpha\gamma$ is given by $\mu(d\gamma) = d\gamma/\gamma$ on $\{\gamma : \gamma > 0\}$. Negative signs of γ may be tackled through a sign change of d, so this implies that μ can be extended to the whole line except $\gamma = 0$. The (right-)invariant measure on the m-dimensional rotation group is given by the uniform measure σ on the m-dimensional sphere in $\mathbb{R}^p$, and the change of sign by $\nu(+) = \nu(-) = 1/2$. This determines the measure ν on Ω_θ.

Theorem 1 gives, in general, a Hilbert space and, in particular, an operator A^θ on this Hilbert space under certain conditions for the case when we, in addition, have a complementary parameter η. From the proof of Theorem 1 in Helland[11], the Hilbert space can be taken to be $\mathcal{K} = L^2(\Omega_\theta, d\nu)$. One of the conditions behind the theorem is the existence of a transitive group G acting upon θ. This can be taken as the group defined above.

5.2. Another model reduction

Again, consider a statistical setting with a large number p of possible predictor variables $x = (x_1, \dots, x_p)'$ and a response y. Assume that these variables have a joint distribution and that we have observed n samples from this distribution.

Again, this introduces the parameters: $\Sigma_{xx} = \mathrm{Cov}(x)$, $\sigma_{xy} = \mathrm{Cov}(x, y)$, $\sigma^2 = \mathrm{Var}(y|x) = \mathrm{Var}(y) - \sigma_{xy}'\Sigma_{xx}^{-1}\sigma_{xy}$ and $\beta = (\beta_1, \dots, \beta_p)' = \Sigma_{xx}^{-1}\sigma_{xy}$. Let again the collection of these parameters be denoted by ϕ, varying in some space Ω_ϕ.

There are many ways to perform a model reduction in a prediction context. Assume that the statistician A also considers another reduction η based upon the same inaccessible parameter ϕ, so $\eta = \eta(\phi)$.

More specifically, I assume: Fix some number m, for $r = 1, \dots, m$ let $\eta_r(\cdot)$ be a p-dimensional vector function defined on Ω_ϕ, and put $\eta(\phi) = (\eta_1(\phi), \dots, \eta_m(\phi))$. For linear prediction, let the reduced regression parameter be $\beta'_m = \beta(\eta, \phi')$ for some function $\beta(\cdot)$, where ϕ' is chosen so that $(\eta(\phi), \phi')$ is in one-to-one correspondence with ϕ. I will suppose that β_m can be estimated under the hypothesis

$H_m{}'$: $\beta'_m = \beta(\eta, \phi')$ *is estimable, but maximally so: If* $\eta = f(\xi)$ *for some function* f *which is not surjective, then* $\beta(\xi, \phi')$ *is not estimable.*

This should be compared to the hypothesis H_m that was made in connection to the specific reduction θ of Section 4. Note that I assume that η also, in relation to the regression coefficient β, has just m vector components, and that both θ and η can be seen as maximal accessible parameters for A. Let M be a fixed group acting on Ω_ϕ which transforms such sets of m p-dimensional vectors into other sets of m p-dimensional vectors.

Then, assuming some fixed value θ_1 of θ, we can first find a $\phi_1 \in \Omega_\phi$ such that $\theta(\phi_1) = \theta_1$. Given some fixed value η_2 of η and a $\phi_2 \in \Omega_\phi$ such that $\eta(\phi_2) = \eta_2$, then either ϕ_1 and ϕ_2 lie on the same orbit of M, or they belong to different orbits. In the first case, there is a bijective function f such that $\phi_2 = f(\phi_1)$. In the second case, there is an element $k \in M$, a $\phi_3 \in \Omega_\phi$ and a bijective function f such that $\phi_2 = f(\phi_3)$ and $\phi_3 = k\phi_1$. Since bijective functions in Ω_ϕ imply equivalent model reductions $\eta(\phi)$, this means that one can without loss of generality assume a transformation k such that $\eta_2 = \eta(\phi_2) = \theta(k\phi_1)$ while $\theta_1 = \theta(\phi_1)$. Since θ_1 and η_2 were arbitrarily chosen, this implies that θ and η are related as defined in Definition 2. The crucial assumptions are that both parameters have the same dimension and are defined as functions on the same space Ω_ϕ.

But this, together with Theorem 7, implies that the conditions of Theorem 1 are satisfied. There exist operators A^θ and A^η in the same Hilbert space $\mathcal{H}$ corresponding to the two model reductions. Note that in this case, A^θ was defined as an operator on $\mathcal{K} = L^2(\Omega_\theta, \nu)$ already in Section 5. Without loss of generality, we can assume that $\mathcal{K} \subseteq \mathcal{H}$, and the operator A^θ is then extended to $\mathcal{H}$ if necessary.

We assume now that one set of assumptions behind Born's rule is satisfied. But first, we just study the two parametric functions $\beta(\theta)$ and $\beta(\eta)$ (strictly speaking $\beta(\eta, \phi')$) in relation to the true regression parameters β.

5.3. Construction of the necessary operators

By assumption, we now have two maximal variables θ and η, and these are both functions of the total parameter ϕ. There is a natural group G acting on θ. If we were able to construct a representation $U(G)$ with the properties given in (iii) of Theorem 1 in Section 2, all the assumptions of that theorem would have been satisfied. We would have proved the existence of a Hilbert space $\mathcal{H}$ and operators A^θ and A^η in that Hilbert space associated with the two theoretical variables.

Instead, we will choose another route. In Helland[11], it was proved that in the case of finite-valued theoretical variables, the assumptions (ii) and (iii) of Theorem 1 are automatically satisfied. We can always approximate continuous theoretical variables (even matrix-valued variables as in this case) by variables that take a finite number of values. For a concrete such construction, see Subsection 5.3 in Helland[10]. For variables taking a finite number of values, we also have a simpler spectral theorem, which will simplify our discussion here.

Concretely, I will approximate θ and η by finite-valued parameters θ_t and η_t (taking k_t values, say.) This can always be done in such a way that $\theta_t \to \theta$ and $\eta_t \to \eta$ uniformly as $t \to \infty$. By the discussion in Helland[11], the assumptions of Theorem 2 are also satisfied for such finite-valued operators, and the consequences for ordinary textbook quantum mechanics follow. In particular, we have a spectral decomposition

$$A^{\theta_t} = \sum_{r=1}^{k_t} \theta_{rt} \otimes \left[v_{rt} v_{rt}^{\dagger}\right], \qquad (23)$$

where θ_{rt} and v_{rt} are the eigenvalues and eigenvectors of A^{θ_t}. Here, θ_{rt} are the different values of θ_t. A Cartesian product is needed since θ_t is a matrix. From now on in this subsection, I will drop the index t, and just treat the parameters θ and η as finite-valued, say taking K values.

Note that each γ_j can be seen as a function of θ: $\gamma_j = \sqrt{\left|\gamma_j d_j\right|^2} \operatorname{sign}(\gamma_j)$, where the sign is determined as follows: Since $\gamma_j d_j = (-\gamma_j)(-d_j)$, each pair (γ_j, d) is counted twice in θ. We can let one of these repetitions correspond to a positive γ_j, the other to a negative γ_j.

In the following, we will not need a theory involving A^{θ}, but only A^{ξ}, where $\xi = \xi(\gamma_1, \dots \gamma_m)$ is a scalar function of the γ-parameters, thus also a function of θ. By (18), this can be found to be

$$A^{\xi} = \sum_{r=1}^{K} \xi\,(\theta_r) v_r v_r^{\dagger}, \qquad (24)$$

where θ_r and v_r are the eigenvalues and eigenvectors of A^{θ}.

6. An optimality theorem for model reduction

In this Section and the next one, we will go back to the case of continuous parameters again.

Let us assume that $\theta = \theta(\phi)$ is the PLS model reduction with a fixed number m of relevant components as described in Section 4, and let $\eta = \eta(\phi)$ be another m-dimensional model reduction as described in Subsection 5.2. Here, ϕ is the parameter of the full model. Assume that there is a continuous group M acting upon Ω_ϕ which transforms the specific sets of m p-dimensional vectors into similar sets of p-dimensional vectors.

The purpose of this Section is to investigate when the PLS model gives the best model reduction for prediction when p is large. Seen from an asymptotic point of view, there are several criteria in the literature for when PLS regression performs well in a prediction setting[5]. Cook and Forzani[7] indicated that PLS performs well in abundant regression where many predictors contribute information about the response. In this article, I will make exact computations and formulate a relatively concrete criterion.

Assumption A. *Let θ with m components be the PLS model assumption, let η denote another m-dimensional model reduction, and let β be the true regression coefficient. Assume that, relative to the distribution of x*

$$\mathrm{Cov}\big(\big[\big(\beta(\eta)-\beta(\theta)\big)\cdot x\big],[(\beta(\eta)+\beta(\theta)-2\beta)\cdot x]\big)>0 \qquad (25)$$

Note that if $\beta(\theta)$ is close to the true regression coefficient, this is guaranteed to hold; see later.

I will prove the following Theorem:

Theorem 6. *Let (x,y) have a joint distribution with all second order parameters given by the parameter ϕ. Assume that all variables have expectation 0 and that the x-covariance matrix Σ_{xx} is positive definite. Make the Assumption A. Then the m-dimensional reduction of ϕ given by the PLS-model is better than the m-dimensional reduction given by η, in the sense that $\mathrm{E}_{(x,y)}(y-\beta(\cdot)\cdot x)^2$ is minimized. Conversely, if the PLS model gives a better prediction than η, then Assumption A must hold.*

Proof. As a point of departure, let η be any m-dimensional model reduction satisfying Assumption A, and let $\beta_m=\beta(\eta,\phi')$, where $(\eta(\phi),\phi')$ is in one-to-one correspondence with the inaccessible parameter ϕ. As shown in Subsection 5.2, we can write $\eta(\phi)=\theta(k\phi)$ for some $k\in M$, and since M is a continuous group, it is meaningful to let k approach the identity, that is, let $\eta(\phi)\rightarrow\theta(\phi)$. Furthermore, looking at prior distributions in the parameter space, the distribution of θ gives a stationary point for the distribution of η.

Since the eigenvectors d_r of Σ_{xx} form a basis for $\mathbb{R}^p$, we have

$$\beta_m(\eta)=\sum_{r=1}^{p}\delta_r\,(\eta)d_r. \qquad (26)$$

The δ_r's are functions of η, and may be seen as close to some γ_r's when η is close to θ. Note that the terms in (26) can be permuted, so without loss of generality, we can let the first m terms correspond to the PLS solution (22). If the hypothesis H_m holds, the γ_r's for $r=m+1,\ldots,p$ are zero.

Let $\beta(\eta)=\beta(\theta)+e(\phi)$. Define $\tau\big(\eta(\phi)\big)=\mathrm{E}(y-\beta(\eta)\cdot x)^2$. Then

$$\tau\big(\eta(\phi)\big)=\mathrm{E}(y-\beta(\theta)\cdot x)^2-2\mathrm{E}(y-\beta(\theta)\cdot x)(e\cdot x)+e'\Sigma_{xx}e. \qquad (27)$$

The cross-term here may be written

$$\sigma_{xy}{}'e-\beta(\theta)'\Sigma_{xx}e=\big(\beta-\beta(\theta)\big)'\Sigma_{xx}\big(\beta(\eta)-\beta(\theta)\big), \qquad (28)$$

So

$$\tau\big(\eta(\phi)\big)=E_{(x,y)}(y-\beta(\theta)\cdot x)^2+F(\phi)=\tau\big(\theta(\phi)\big)+F(\phi), \qquad (29)$$

where $\beta(\theta)$ is given by (22) and

$$F(\phi) = (\beta(\eta) + \beta(\theta) - 2\beta)'\Sigma_{xx}\left(\beta(\eta) - \beta(\theta)\right), \qquad (30)$$

where β is the true regression vector. Comparing this with (25, concludes the proof of Theorem 6. Since all calculations are exact, there is an if and only if here. ■

Corollary 1. *Under the hypothesis H_m of Section 4, the PLS regression model always gives the best model reduction for linear prediction.*

Proof. Under H_m we have $\beta = \beta(\theta)$, and (30) is positive for all $\eta \neq \theta$. ■

Corollary 2. *Assume that* $\mathrm{Var}\left(\left(\beta - \beta(\theta)\right) \cdot x\right) < \frac{1}{4}\mathrm{Var}\left(\left(\beta(\eta) - \beta(\theta)\right) \cdot x\right)$*. Then the PLS model will give better linear predictions than the model reduction η.*

Proof. (30) can be written

$$F(\phi) = \left(\beta(\eta) - \beta(\theta)\right)'\Sigma_{xx}\left(\beta(\eta) - \beta(\theta)\right) - 2\left(\beta - \beta(\theta)\right)'\Sigma_{xx}\left(\beta(\eta) - \beta(\theta)\right). \qquad (31)$$

By a version of Schwarz' inequality, this is guaranteed to be positive if $\left(\beta - \beta(\theta)\right)'\Sigma_{xx}\left(\beta - \beta(\theta)\right) < \frac{1}{4}\left(\beta(\eta) - \beta(\theta)\right)'\Sigma_{xx}\left(\beta(\eta) - \beta(\theta)\right)$. ■

7. Estimation

Let there now be data (X, y), and consider an estimator $\widehat{\beta_a} = \hat{\beta}(X, y)$ of the regression vector β under the hypothesis H_m. In general, we will seek estimators based upon a θ corresponding to a relevant components, where $a \geq m$, with m specified by the hypothesis. In the PLS case, this means that $\widehat{\beta_a}$ is based upon a steps in the PLS algorithm. In general, it means that we can write $\widehat{\beta_a} = \sum_{r=1}^{a} \hat{\gamma}_r \, \widehat{d_r}$ corresponding to $\beta = \sum_{r=1}^{a} \gamma_r \, d_r$, which holds under H_m.

Note that, by the previous Section, if m is chosen such that H_m holds, or if $\beta(\theta)$ is sufficiently close to the true regression parameter β, then the PLS model is best in the sense of giving the best linear prediction among all m-dimensional model reductions. In this Section, we will try to find estimators based upon a relevant components that give as good predictions as possible. In practice, a must be chosen by cross-validation, or by some independent training set of observations.

The most common solution to this problem is the sample PLS algorithm with a components. However, this can in principle be improved somewhat by the following argument: The hypothesis H_m is characterized in the parameter space by an algebraic condition $w_{m+1} = 0$, where $w_{\cdot}$ is the PLS weight vector (see, for instance, Helland[1]). This also implies $w_{a+1} = 0$ for $a \geq m$, a restriction in the parameter space. The sample PLS estimator does not necessarily follow a similar restriction; in general, we have $\widehat{w_{a+1}} \neq 0$, which implies that the estimator is outside the relevant parameter space.

An estimator that is inside the parameter space is the maximum likelihood estimator, first proposed in Helland[2] and improved in several articles by Dennis Cook and collaborators; see Cook[21]. However, this estimator does not exist when $p > n$.

A completely different approach is given in Helland et al.[8]. The point of departure is a generalization of the group G of Definition 5 in Section 4 above, seen as acting on the parameter β defined by (21).

Definition 6. *Define the group G acting on β by orthogonal transformations of the vectors d_j in (21) and, in addition, separate linear scale transformations of the parameters γ_j: $\gamma_j \mapsto \alpha_j \gamma_j$ with $\alpha_j > 0$.*

In analogy to Theorem 5, it can be proved that the hypotheses H_m for varying m represent the orbits of the group G. (See Theorem 2 in Helland et al.[8].)

An important concept in connection with group transformations of statistical models is that of *equivariance*. This requires first that we start with a group G_0 as acting on the sample space, and then introduce the group G on the parameter space Ω_θ by

$$P^{g\theta}(A) = P^{\theta}(g_0^{-1}A). \qquad (32)$$

Then an estimator $\hat{\beta}$ of the parameter β is equivariant under the group G_0 if it transforms under the group in the same way as the corresponding parameter: $\widehat{g(\beta)} = g_0(\hat{\beta})$ for all $g_0 \in G_0$.

In the present case, we can let $\widehat{d_j}$ $(j = 1, \ldots, p)$ be the eigenvectors of the sample covariance matrix $(n-1)^{-1}X'X$, define $\widehat{\gamma_j}$ by the sum $(n-1)^{-1}X'y = \sum_{j=1}^{p} \widehat{\gamma_j}\, \widehat{d_j}$, and then define the sample group G_0 acting on the $\widehat{\gamma_j}$'s and the $\widehat{d_j}$'s in analogy to Definition 6.

It is shown in Theorem 3 and Theorem 4 of Helland et al.[8] that both the PLS estimator and the principal component estimator with the usual ordering of the eigenvectors are equivariant under the group G.

A main theorem from Helland et al.[8] is

Theorem 7. *In a statistical model, let the parameter $\eta(\phi)$ be estimated by $\hat{\eta}(z)$. Let the loss function be given by $B(z) \parallel \hat{\eta}(z) - \eta(\phi) \parallel^2$, and assume that this loss function is invariant under the parametric group G together with the corresponding data group G_0. Let G be transitive with right invariant measure ν. Then the best equivariant estimator for η in terms of expected loss is given by the Bayes estimator with prior ν if this Bayes estimator exists.*

In the current situation, we will estimate β by some data vector $\hat{\beta}$, and we can, in principle, use the invariant loss function $\parallel \hat{\beta} \parallel^{-2} \parallel \hat{\beta} - \beta \parallel^2$. The relevant prior for β under H_m is found by first letting $(d_1, \ldots, d_m)$ have an invariant distribution under orthogonal transformations, and then letting the positive scale parameters $\gamma_1, \ldots, \gamma_m$ have a joint improper density

$$\frac{1}{\gamma_1}\dots\frac{1}{\gamma_m}. \qquad (33)$$

Unfortunately, the Bayes estimator does not exist in this case. The criterion for existence given by Taraldsen and Lindqvist[19] (see Section 3) is not satisfied: The relevant integral involving the scale parameters as these parameters tend towards 0 is infinite. In Helland et al.[8] this is solved by proposing a near-optimal solution, Bayes PLS, where the prior density $1/\gamma$ is replaced by the density $1/(\gamma)^{1-\epsilon}$, where ϵ is a small positive number. A corresponding modified group G, having such an invariant measure, is then discussed.

The near-optimal predictor Bayes PLS does well compared to other methods, such as ordinary PLS, as shown in a simulation study by Helland et al.[22]. The main disadvantage is that it involves heavy calculation. An R program for Bayes PLS has been written by Solve Sæbø.

8. Quantum theory from data

A version of Born's formula that can be used for estimators of parameters was given by (20) in Section 2.

Go back to the statistician A. Assume that he has done a statistical analysis on the dataset $\mathcal{X} = (X, y)$ and has found an estimate $\hat{\theta}$ of the parameter θ as defined in this article. Assume that this estimate is found by a partial least squares procedure like Bayes PLS with a components. In this case, the estimator $\hat{\theta}$ has a probability distribution depending on the parameter θ for some fixed $m < a$ as defined in this article, so the assumptions behind (20) hold if the density operator ρ^{η} can be defined and the assumptions leading to Born's formula are satisfied.

Now introduce the more experienced statistician B, and assume that he is interested in another model reduction η as described in Subsection 5.2. By a version of Theorem 1, there exists an operator A^{η} associated with η and a corresponding resolution of the identity. Furthermore, assume that B has a probability distribution of η, either a prior, or from the dataset a posterior distribution or a confidence distribution. From this, one can construct a density operator ρ^{η} for η, see Section 2. Given this ρ^{η}, he is interested in 1) the conditional probability distribution of A's θ and 2) the conditional distribution of A's estimator $\hat{\theta}$.

The basic assumptions behind Born's formula must hold: A and B must believe in the likelihood principle, and A must have ideals that can be modeled by what he considers to be a perfectly rational superior being. This can be identified by the experienced statistician B plus some theoretical statistical ideals that they both share. The probabilities of Born's formula must be seen as calculated by this superior being, in the language of Section 3, by the experienced statistician B.

9. A first condition for optimal linear prediction by PLS

In the statistical literature, there are several methods proposed for linear prediction of a variable y from many predictors, possibly related. One example is ridge regression with some given ridge parameter. In this section, I will investigate, in principle, when PLS-like methods are optimal in some sense in this large class of methods. In this section, I will fix a number m and assume that the hypothesis H_m (see Section 4) holds.

Assume that we want to find a good predictor of y from a p-dimensional x based upon n data X, y. For simplicity, let all data variables be centered to zero expectation.

In the Theorem below, I consider either ordinary PLS regression or Bayes PLS. The criterion used is mean square prediction error, where we take expectation over the variables in the data set, the future x and y data and the PLS parameter θ.

Theorem 8. *Let $p > n$, and let $\hat{\beta}$ be an arbitrary estimator of β. Then, for each a such that $m \leq a < p$, letting $\widehat{\beta_a}$ be constructed from PLS estimation with a components, assuming the hypothesis H_m, we have*

If $\hat{\beta}$ is sufficiently far from $\widehat{\beta_a}$, more concretely if

$$\mathrm{E}_\theta \mathrm{E}_{X,y}\big(\hat{\beta} - \widehat{\beta_a}\big)' \Sigma_{xx} \big(\hat{\beta} - \widehat{\beta_a}\big) > 4\mathrm{E}_\theta \mathrm{E}_{X,y}\big(\beta - \widehat{\beta_a}\big)' \Sigma_{xx} \big(\beta - \widehat{\beta_a}\big), \qquad (34)$$

where β is the true regression coefficient, then we have

$$\mathrm{E}_\theta \mathrm{E}_{X,y} \mathrm{E}_{x,y}\big(y - \widehat{\beta_a} \cdot x\big)^2 < \mathrm{E}_\theta \mathrm{E}_{X,y} \mathrm{E}_{x,y}\big(y - \hat{\beta} \cdot x\big)^2. \qquad (35)$$

Proof. Let $\mathrm{E} = \mathrm{E}_\theta \mathrm{E}_{X,y} \mathrm{E}_{x,y}$. In analogy with the calculations of Section 6, we have

$$\mathrm{E}\big(y - \hat{\beta} \cdot x\big)^2 = \mathrm{E}\big(y - \widehat{\beta_a} \cdot x\big)^2 + F, \qquad (36)$$

where

$$\begin{aligned} F &= \mathrm{E}\big(\hat{\beta} + \widehat{\beta_a} - 2\beta\big)' \Sigma_{xx} \big(\hat{\beta} - \widehat{\beta_a}\big) \qquad (37) \\ &= \mathrm{E}\big(\hat{\beta} - \widehat{\beta_a}\big)' \Sigma_{xx} \big(\hat{\beta} - \widehat{\beta_a}\big) - 2\mathrm{E}\big(\beta - \widehat{\beta_a}\big)' \Sigma_{xx} \big(\hat{\beta} - \widehat{\beta_a}\big) \qquad (38) \\ &\geq \mathrm{E}\big(\hat{\beta} - \widehat{\beta_a}\big)' \Sigma_{xx} \big(\hat{\beta} - \widehat{\beta_a}\big) \\ &\quad - 2\sqrt{\mathrm{E}\big(\beta - \widehat{\beta_a}\big)' \Sigma_{xx} \big(\beta - \widehat{\beta_a}\big) \cdot \mathrm{E}\big(\hat{\beta} - \widehat{\beta_a}\big)' \Sigma_{xx} \big(\hat{\beta} - \widehat{\beta_a}\big)} \qquad (39) \end{aligned}$$

by a variant of Schwarz's inequality.

Inspecting the inequality (39), it follows that $F > 0$, and hence (35) holds if (34) is satisfied. ∎

The criterion (34) will be simplified considerably under reasonable assumptions in Section 11. But first, we discuss the model reduction problem further.

10. On the optimality of the PLS model under model reduction

Let us go back to the situation with two different model reductions θ and η, both corresponding to reductions to dimension m, as specified with the hypothesis H_m of Section 4 and the hypothesis H'_m of Subsection 5.2. We are interested in conditions under which the PLS model is best in terms of mean square prediction error. Go back to Section 6.

Theorem 9. *Assume that*

$$4\mathrm{E}_\theta\big(\beta-\beta(\theta)\big)'\Sigma_{xx}\big(\beta-\beta(\theta)\big)<\mathrm{E}_\theta\big(\beta(\eta)-\beta(\theta)\big)'\Sigma_{xx}\big(\beta(\eta)-\beta(\theta)\big). \qquad (40)$$

Then

$$\mathrm{E}_\theta\mathrm{E}_{(x,y)}(y-\beta(\theta)\cdot x)^2<\mathrm{E}_\theta\mathrm{E}_{(x,y)}(y-\beta(\eta)\cdot x)^2. \qquad (41)$$

Proof. Repeat the proof of Theorem 6 and of Corollary 2 of Section 6 with the expectation over θ taken in all equations. ■

We want to study the criterion (40) more closely. Since $\beta(\theta)=\sum_{j=1}^{m}\gamma_j\, d_j$, the left-hand side is just

$$4\mathrm{E}_\theta\sum_{j=m+1}^{p}\gamma_j^2\,\lambda_j=4\sum_{j=m+1}^{p}\gamma_j^2\,\lambda_j, \qquad (42)$$

assuming that the γ_j's are independent, where here $\{\lambda_j\}$ are the irrelevant eigenvalues of Σ_{xx}, those not affected by the model reduction θ.

The right-hand side of the inequality (40) is bounded below by

$$\sum_{j=1}^{m}\mathrm{E}_\theta\left(\zeta_j-\gamma_j\right)^2\lambda_j, \qquad (43)$$

where we have evaluated $\beta(\eta)$ in terms of the p eigenvectors d_j of Σ_{xx}:

$$\beta(\eta)=\sum_{j=1}^{p}\zeta_j\, d_j. \qquad (44)$$

Our aim is to find a criterion under which the PLS model reduction is better in some sense than any other model reduction. This means that the parameters ζ_j in (43) are completely arbitrary.

Let now the basic parameter θ have some probability distribution, which implies a probability distribution of $\gamma_1,\ldots,\gamma_m$. Then the criterion (40) is satisfied over θ for a model reduction η if

$$\mathrm{E}_\theta\sum_{j=1}^{m}\left(\zeta_j-\gamma_j\right)^2\lambda_j>4\sum_{j=m+1}^{p}\gamma_j^2\,\lambda_j. \qquad (45)$$

For each j we have that $\mathrm{E}_{\gamma_j}\left(\zeta_j-\gamma_j\right)^2\geq\mathrm{E}_{\gamma_j}\left(\mu_j-\gamma_j\right)^2$, where $\mu_j=\mathrm{E}_{\gamma_j}(\gamma_j)$. So, taking a lower bound on the left-hand side of (45), we see that the criterion (40) is satisfied for every possible reduction η if

$$\mathrm{E}_\theta \sum_{j=1}^{m} \left(\gamma_j - \mu_j\right)^2 \lambda_j > 4 \sum_{j=m+1}^{p} \gamma_j^2 \lambda_{j\cdot j} \qquad (46)$$

The probability distribution of θ will depend on the situation. As a first tentative situation, let us first assume a probability distribution of γ_j which is close to the right-invariant measure $\mu(d\gamma) = d\gamma/\gamma$ under the group G. (See subsection 5.1.) This measure gives an improper distribution, and under a proper distribution close to this distribution, the lefthand side of (46) can be made arbitrarily large. This indicates that under such circumstances, it will be easy to satisfy the criterion (40) (for any m).

Imagine now a situation similar to the one sketched in Section 3: An experienced statistician B who is an open-minded Bayesian, and a younger statistician A. Assume that B in some way has made himself a joint prior or posterior of the parameter η, and we are interested in his conditional distribution of θ, given this distribution of η. We can think of A as connected to the PLS-parameter θ, that η is maximally accessible to B, and θ is maximally accessible to A. These two parameters are complementary in this situation.

For simplicity, we approximate the parameters by finite-valued parameters as in Subsection 5.3. Then the symmetry conditions of Theorem 1 of Section 2 are automatically satisfied; we have a Hilbert space $\mathcal{H}$, and operators A^θ and A^η in this Hilbert space. There is also a resolution of the identity connected to the parameter η, and from this, a density operator ρ^η describing the information on η that the statistician B has. By the Born formula, this statistician also has a probability distribution of A's parameter θ, and by the theory of Section 2, we get for any scalar function f of θ:

$$\mathrm{E}_\theta(f(\theta)|\rho^\eta) = \mathrm{trace}\left(\rho^\eta f\left(A^\theta\right)\right) \qquad (47)$$

Let now $f(\theta) = \sum_{j=1}^{m} \left(\gamma_j - \mu_j\right)^2 \lambda_j$. Considering the criterion (46), we find

Theorem 10. *Assume that*

$$\sum_{j=1}^{m} \lambda_j \,\mathrm{trace}\left(\rho^\eta \sum_{r=1}^{K} \left(\gamma_{jr} - \mu_{jr}\right)^2 v_{jr} v_{jr}^\dagger\right) > 4 \sum_{j=m+1}^{p} \gamma_j^2 \lambda_j, \qquad (48)$$

where the discrete θ is assumed to take K values, γ_{jr} are the corresponding values of γ_j, v_{jr} are the eigenvectors of A^θ, and $\mu_{jr} = \mathrm{E}_{\gamma_{jr}}\left(\gamma_{jr}\right)$. Then the criterion (40) for optimality of the PLS model is satisfied, and (41) holds.

The simplest case is when B has a non-informative prior on η. Then $\rho^\eta = K^{-1} I$, and in the sum over r in (48) we can let K tend to infinty and use the law of large numbers.

Theorem 11. *Assume that, as seen by B, the model reduction η is non-informative, and independent of the model reduction θ, and assume that relative to the γ_j distribution as seen by B*

$$\sum_{r=1}^{m} \lambda_j \,\mathrm{E}_{\gamma_j}\big(\gamma_j - \mu_j\big)^2 > 4 \sum_{j=m+1}^{p} \gamma_j^2 \,\lambda_j. \qquad (49)$$

Then the criterion (40) for optimality of the PLS model is satisfied, and (41) holds.

This indicates strongly that H_m gives a good model reduction when the relevant eigenvalues of Σ_{xx} are substancially larger than the irrelevant ones. It is also relavant that the variances of the relevant regression coefficients are fairly large.

Note that the criterion (49) is only connected to the PLS model reduction. If this criterion is satisfied for some m, the PLS model reduction is better than *all* other model reductions. Also, note that the lefthand side of (49) is increasing with m, and the righthand side is decreasing. Thus it seems reasonable that the criterion in most situations is satisfied if m is large enough.

11. On optimal linear prediction by PLS

We are now ready to go back to the situation of Section 9, where the Bayes PLS estimator $\hat{\beta}_a$ with a steps was compared with an arbitrary estimator $\hat{\beta}$. The inequality (34) has the same form as the inequality (40), and the discussion from Section 10 can be carried over. The following can be seen as a main result of this article.

Theorem 12. *Let $p > n$, let $\hat{\beta}$ be an arbitrary estimator of β, and assume that the hypothesis H_m holds for some m. For some a such that $m \le a < p$, let $\widehat{\beta_a}$ be constructed from Bayes PLS estimation with a components. Consider a situation with two statisticians A and B, where A has a prior distribution over the PLS-parameter θ under H_m, and B has a noninformative distribution over the parameters behind the estimator $\hat{\beta}$. Assume that A considers B as a perfectly rational ideal. Then, if in relation to the γ_j distribution as seen by B*

$$\sum_{r=1}^{m} \lambda_j \,\mathrm{E}_{\gamma_j}\big(\gamma_j - \mu_j\big)^2 > 4 \sum_{j=m+1}^{p} \gamma_j^2 \,\lambda_j, \qquad (50)$$

we conclude that

$$\mathrm{E}_{\theta}\mathrm{E}_{X,y}\mathrm{E}_{x,y}\big(y - \widehat{\beta_a}\cdot x\big)^2 < \mathrm{E}_{\theta}\mathrm{E}_{X,y}\mathrm{E}_{x,y}\big(y - \hat{\beta}\cdot x\big)^2. \qquad (51)$$

Proof. Start with the criterion in Theorem 8, and follow arguments as in Section 10 using the assumptions behind Born's formula. The relevant operator $\hat{A}^{\hat{\theta}}$ is the data operator discussed in Section 8. (Compare equations (19) and (20) in Section 2.) ■

Remarks

The result here is formulated for Bayes PLS, since this method gives an estimator for θ which is inside the parameter space. The estimator $\hat{\theta}$ under the ordinary PLS algorithm is not inside the parameter space defined by H_m, in the sense that the algorithm does not stop automatically at step a. Simulations by Helland et al.[22] show also that Bayes PLS in general seems to perform better than ordinary PLS.

However, the result should be seen in relation to the discussion of Helland and Almøy[5], based upon simulations. It should also be compared to the statements of Cook and Forzani[7] concerning when PLS is the best choice.

We should also remark the following: Theorem 12 is formulated in a somewhat artificial setting with two statisticians A and B, in agreement with the general discussion of this article. But the criterion (50) has a meaning over and above this setting, and the conclusion (51) is a general result on properties of two different estimators of β. This indicates that the result must have a validity also above this artificial setting.

12. Conclusions

One purpose of this article has been to find arguments connected to the optimality of PLS type regression under certain conditions. As exact results on the properties of the PLS algorithm are lacking in the literature, in my opinion, all related results, either tentative or building upon exact calculations, are of interest.

Another purpose of the article has been to illustrate how recent results from quantum theory can be used in a statistical setting.

I conjecture that the results of this article can be generalized to the envelope model of Cook[21]. To formulate the precise theorems and to construct the proofs for this general case, however, are open problems.

Finally, I think that it may be of some value to use arguments from different scientific cultures in a theoretical statistical context. In general, communications across scientific borders is, as I see it, one of the prerequisites for real progress in science.

Further applications of quantum theory to statistics are under investigation[23].

References

1. [a, b, c, d]Helland IS (1990). "Partial least squares regression and statistical models". Scand. J. Stat.. 17: 97-114.
2. [a, b]Helland IS (1992). "Maximum likelihood regression on relevant components". J. Roy. Statist. soc. Ser. B. 54: 637-647.
3. ^Helland IS (2001). "Some theoretical aspects of partial least squares regression". Chemom. Intell. Lab. Syst.. 58: 97-107.
4. ^Næs T, Helland IS (1993). "Relevant components in regression". Scand. J. Stat.. 20: 239-250.
5. [a, b, c]Helland IS, Almøy T (1994). "Comparison of prediction methods when only a few components are relevant". J. Amer. Stat. Ass.. 89: 583-591.
6. [a, b, c]Cook RD, Helland IS, Su Z (2013). "Envelopes and partial least squares regression". J. R. Stat. Soc. Ser. B. Stat. Methodol.. 75: 851-877.

7. [a, b, c]Cook RD, Forzani L (2019). "Partial least squares prediction in high-dimensional regression". The Annals of Statistics. 47 (2): 884-908.
8. [a, b, c, d, e, f, g]Helland IS, Sæbø S, Tjelmeland H (2012). "Near optimal prediction from relevant components". Scand. J. Stat.. 39: 695-713.
9. ^Dunjko V, Briegel HJ (2019). "Machine learning & artificial intelligence in the quantum domain: a review of recent progress". Rep. Prog. Phys.. 81: 074001.
10. [a, b, c, d, e, f, g, h]Helland IS (2021). Epistemic Processes. A Basis for Statistics and Quantum Theory. 2. edition. Springer, Berlin.
11. [a, b, c, d, e, f, g, h, i, j, k, l, m]Helland IS (2024a). "An alternative foundation of quantum theory". Foundations of Physics. 54: 3.
12. [a, b, c]Helland IS (2024b). "A new approach toward the quantum foundation and some consequences". Academia Quantum. 1: 1-9.
13. [a, b, c, d, e, f]Helland IS (2024c). "On probabilities in quantum mechanics". APL Quantum. 1: 036116.
14. [a, b]Hall BC (2013). Quantum Theory for Mathematicians. Graduate Texts in Mathematics 267. Springer, Berlin.
15. ^Helland IS (2010). Steps Towards a Unified Basis for Scientific Models and Methods. World Scientific, Singapore.
16. ^Schweder T, Hjort NL (2016). Confidence, Likelihood, Probability. Statistical Inference with Confidence Distributions. Cambridge University Press, Cambridge.
17. ^Campanella M, Jou D, Mongiovi MS (2020). Interpretative aspects of quantum mechanics. In Matteo Campanella's Mathematical Studies. Springer, Cham.
18. ^Wijsman RA (1990). Invariant Measures on Groups and Their Use in Statistics. Lecture Notes - Monograph Series 14. Institute of Mathematical Statistics, Hayward, California.
19. [a, b]Taraldsen G, Lindqvist H (1990). "Improper priors are not improper". The American Statistician. 64 (2): 154-158.
20. ^Hastie T, Tibshirani R, Friedman J (2009). The Elements of Statistical Learning. 2. ed., Springer, Berlin.
21. [a, b, c]Cook RD (2018). An Introduction to Envelopes. Dimension Reduction for Efficient Estimation in Multivariate Statistics. Wiley, Hoboken, NJ.
22. [a, b]Helland IS, Sæbø S, Almøy T, Rimal R (2018). "Model and estimators for partial least squares regression". J. Chemometrics. 32: 1-13.
23. ^Helland IS (2024d). Quantum probabilities for statisticians. Some new ideas. Preprint.