The Annals of Statistics
1996, Vol. 24, No. 2, 816–824

PARTIAL LEAST SQUARES ALGORITHM YIELDS SHRINKAGE ESTIMATORS

BY CONSTANTINOS GOUTIS

Universidad Carlos III de Madrid

We give a geometric proof that the estimates of a regression model derived by using partial least squares shrink the ordinary least squares estimates. The proof is based on a sequential construction algorithm of partial least squares. A discussion of the nature of shrinkage is included.

1. Introduction. Partial least squares is a class of regression estimation methods initially developed by Wold (1966, 1973) that has recently been increasingly popular among chemists and other scientists as a technique for treating highly collinear data. It is almost routinely applied in spectroscopy, where one aim is to predict a chemical composition from a near infrared reflectance spectrum. The philosophy is to reduce the data to a manageable form before solving the prediction problem. For a treatment of this and related methods in multivariate calibration setups, see Martens and Næs (1989). Some papers examining partial least squares include Helland (1988) and Næs and Martens (1985), and a thorough review can be found in Frank and Friedman (1993).

In this paper we give a geometric proof that the coefficients derived by partial least squares shrink the ordinary least squares coefficients. Though shrinkage features of partial least squares estimates have been discussed [Sundberg (1993); Frank and Friedman (1993)], these papers do not contain any proof concerning shrinkage effects. Frank and Friedman (1993) give an analysis in terms of the eigendirections of the predictor sample covariance matrix and show that there are some directions in which partial least squares increases the projected length of the ordinary least squares solution. We examine whether the overall length of the vector of the partial least squares coefficients is less than that of ordinary least squares.

The paper is organized as follows. In Section 2 we describe the setup of the problem and a picture illustrating the geometry. Section 3 presents our main result and its proof, and we conclude with a discussion.

Received July 1993; revised December 1994.

AMS 1991 *subject classifications*. Primary 62J07; secondary 62F10.

Key words and phrases. Biased estimation, chemometrics, collinearity, data compression methods, ordinary least squares, regression.

2. Setup and geometry.

We suppose that the data consist of a $n \times p$ matrix $\mathbf{X}$ of rank r and a $n \times 1$ vector $\mathbf{y}$. The columns of $\mathbf{X}$ represent "explanatory" variables, whereas $\mathbf{y}$ is a vector of values of the "response" variable. Each row of $\mathbf{X}$ and the corresponding element of $\mathbf{y}$ represents an observation. The goal is to predict a future value of y_f corresponding to known values x_{fj} of the given variables. A prediction formula, closely related to the standard linear model, is

$$\hat{y}_f = \sum_{j=1}^{p} x_{fj} \hat{\beta}_j \tag{1}$$

and the coefficients $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ can be considered as estimates of parameters.

Both ordinary and partial least squares are equivariant under orthogonal transformations, so we can transform the problem to the canonical form [see, e.g., Scheffé (1959), page 21]. Since lengths of vectors are preserved under orthogonal transformations, the estimates in the original form will have the same length as in the canonical form. We find orthogonal matrices $\mathbf{P}$ and $\mathbf{Q}$ of appropriate dimensions so that $\mathbf{X} = \mathbf{P}'\mathbf{D}\mathbf{Q}$, where $\mathbf{D} = (d_{ij})$ is a $n \times p$ matrix with $d_{ii} > 0$ for $i = 1, 2, \dots, r$ and $d_{ij} = 0$ otherwise, and use $\mathbf{D}$ and $\mathbf{Py}$ instead of $\mathbf{X}$ and $\mathbf{y}$, respectively. To compute $(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$, we can ignore all but the first r rows and columns of $\mathbf{D}$. Hence, we may assume without loss of generality that $n = p = r$ and $\mathbf{X}$ is a diagonal matrix with positive diagonal elements. Furthermore, we can take $\mathbf{y}$ to be the least squares fitted values and $\hat{\boldsymbol{\beta}} = \mathbf{X}^{-1}\mathbf{y}$ to be the least squares coefficients.

The partial least squares method determines the coefficients by choosing subspaces of the column space of $\mathbf{X}$ sequentially and then projecting $\mathbf{y}$ onto these subspaces. The fitted values and parameter estimates at step a of the sequence will be denoted by $\hat{\mathbf{y}}_a$ and $\hat{\boldsymbol{\beta}}_a$, respectively. The algorithm, under the canonical form of $\mathbf{X}$ and $\mathbf{y}$, can be described as follows [see also Helland (1988) or Stone and Brooks (1990)]:

At the first step, one searches for a $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ to maximize the sample covariance of $\mathbf{t}_1$ and $\mathbf{y}$ for a fixed $\mathbf{w}_1'\mathbf{w}_1$. The solution is $\mathbf{t}_1 \propto \mathbf{X}^2\mathbf{y}$ or $\mathbf{w}_1 \propto \mathbf{X}\mathbf{y}$. The normalization constraint $\mathbf{w}_1'\mathbf{w}_1 = 1$ is often used, but we will take

$$\mathbf{w}_1'\mathbf{w}_1 = K_1^2 \equiv \mathbf{y}'\mathbf{X}^{-2}\mathbf{y}. \tag{2}$$

The fitted values are given by $\hat{\mathbf{y}}_1 = \mathbf{t}_1(\mathbf{t}_1'\mathbf{t}_1)^{-1}\mathbf{t}_1'\mathbf{y} \equiv \mathbf{P}_1\mathbf{y}$, where $\mathbf{P}_1$ is the projection matrix $\mathbf{t}_1(\mathbf{t}_1'\mathbf{t}_1)^{-1}\mathbf{t}_1'$.

The subspaces of subsequent steps are determined by finding arrays orthogonal to the previous ones, to maximize the sample covariance with $\mathbf{y}$. More precisely, at step a, $\mathbf{t}_a$ has the form $\mathbf{t}_a = \mathbf{X}\mathbf{w}_a$ and maximizes $\mathbf{t}_a'\mathbf{y}$ subject to $\mathbf{t}_1'\mathbf{t}_a = \mathbf{t}_2'\mathbf{t}_a = \cdots = \mathbf{t}_{a-1}'\mathbf{t}_a = 0$ and a normalizing constraint

$$\mathbf{w}_a'\mathbf{w}_a = K_a^2 \equiv (\mathbf{y} - \hat{\mathbf{y}}_{a-1})'\mathbf{X}^{-2}(\mathbf{y} - \hat{\mathbf{y}}_{a-1}). \tag{3}$$

So by letting $\mathbf{P}_s = \mathbf{t}_s(\mathbf{t}'_s\mathbf{t}_s)^{-1}\mathbf{t}'_s$, we obtain

$$\mathbf{t}_a \propto \mathbf{X}^2\left(\mathbf{I} - \sum_{s=1}^{a-1}\mathbf{P}_s\right)\mathbf{y}, \tag{4}$$

$$\mathbf{w}_a \propto \mathbf{X}\left(\mathbf{I} - \sum_{s=1}^{a-1}\mathbf{P}_s\right)\mathbf{y} \tag{5}$$

and $\hat{\mathbf{y}}_a = \sum_{s=1}^{a}\mathbf{P}_s\mathbf{y}$. Since the span of $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_a$ is a subspace of the column space of $\mathbf{X}$, it follows immediately that $\|\hat{\mathbf{y}}_a\| \le \|\mathbf{y}\|$ [see Denham (1991), page 74].

The geometry of the model can be seen in Figure 1, to which we will refer throughout the paper. A similar picture and an extensive discussion appears in Phatak, Reilly and Penlidis (1992). Due to limited drawing ability, we have taken $r = 3$. The ellipsoid $\mathscr{E}_0$ has axes with lengths proportional to the diagonal elements of $\mathbf{X}$. The fitted values using least squares are represented by the point A which lies on $\mathscr{E}_0$.

Several geometrical arguments and concepts are independent of the coordinate system and for them we will use purely geometrical language and notation. In other words, we differentiate between a vector and its coordinates, since the vector is a geometric object whereas the coordinates depend on the basis and have different significance in different bases. The norm of an array will be the square root of the sum of squares of its elements.

We will use two systems of bases and coordinates in the drawn space. The first basis is the natural one consisting of equal unit vectors along the axes of $\mathscr{E}_0$. We will refer to the coordinates with respect to this basis as *spherical coordinates* or simply *coordinates*. We will also use the half axes of $\mathscr{E}_0$ as basis and refer to the respective coordinates as *elliptical coordinates*. The relation between elliptical coordinates $\mathbf{w}$ and spherical coordinates $\mathbf{t}$ is $\mathbf{t} = \mathbf{X}\mathbf{w}$. The terminology is justified since all vectors with elliptical coordinates $\mathbf{w}$ such that $\mathbf{w}'\mathbf{w} = d^2$ lie on an ellipse of the form $\mathbf{t}'\mathbf{X}^{-2}\mathbf{t} = d^2$ instead of a sphere. Abusing the language somewhat, we will refer to d as the radius of the ellipse. Therefore, the point A has elliptical coordinates equal to $\hat{\boldsymbol{\beta}}$ and spherical coordinates equal to $\mathbf{y}$ and $\mathscr{E}_0$ has radius K_1. As a convention, we consider the various ellipsoids as surfaces rather than solids. Hence a point will lie "inside," "on" or "outside" an ellipsoid, rather than in the "interior," the "boundary" or the "exterior," respectively.

3. The main result.

We are now ready to show our main result concerning the partial least squares estimates, which can be stated as follows:

THEOREM 1. *For every $a \le r$ we have $\|\hat{\boldsymbol{\beta}}\| \ge \|\hat{\boldsymbol{\beta}}_a\|$.*

PROOF. The proof will proceed by constructing the partial least squares estimates sequentially and showing that the result is true at each step. It will be clear that taking $r = 3$ is sufficient; higher dimensions are essentially the

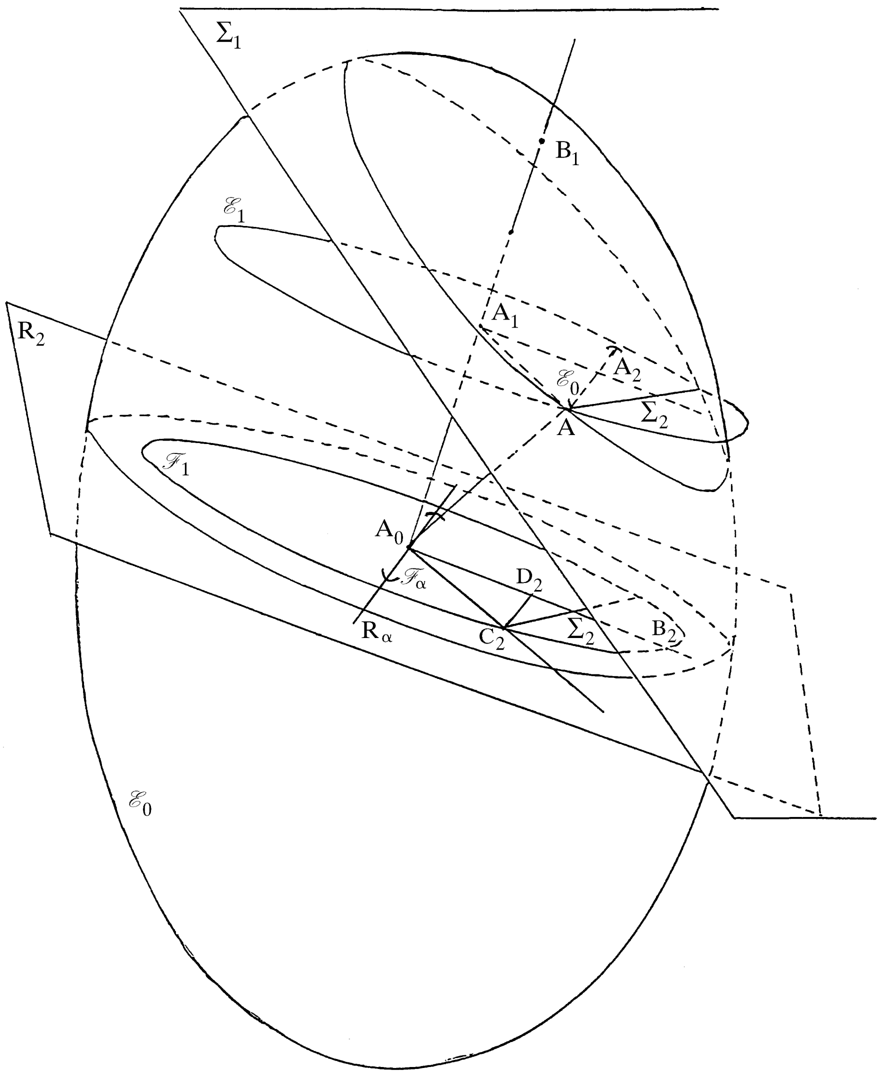

FIG. 1. *Geometry of partial least squares.*

same. Taking $r = 2$ would not be enough since there are no intermediate steps between $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}}$. We will take without loss of generality $\hat{\boldsymbol{\beta}}_a \neq \hat{\boldsymbol{\beta}}$ for $a < r$.

For the first step, we note that all vectors with an endpoint in a hyperplane perpendicular to $\overrightarrow{A_0A}$ have the same inner product with $\overrightarrow{A_0A}$. The standardization (2) restricts the vectors with coordinates $\mathbf{t}_1$ to have endpoints on the ellipsoid $\mathscr{E}_0$ and the particular radius K_1 is chosen so that the vector with coordinates $\mathbf{y}$ lies on $\mathscr{E}_0$. The vector with an endpoint on $\mathscr{E}_0$ maximizing this inner product is $\overrightarrow{A_0 B_1}$ (with coordinates $\mathbf{t}_1 = K_1 \mathbf{X}^2 \mathbf{y}/\|\mathbf{X}\mathbf{y}\|$ and elliptical coordinates $\mathbf{w}_1 = K_1 \mathbf{X}\mathbf{y}/\|\mathbf{X}\mathbf{y}\|$). The point B_1 is the intersection of $\mathscr{E}_0$ and a hyperplane (not drawn in Figure 1) tangent to $\mathscr{E}_0$ and perpendicular to $\overrightarrow{A_0 A}$. The coordinates $\mathbf{t}$ of points of this hyperplane satisfy

$$\mathbf{t}_1' \mathbf{X}^{-2} \mathbf{t}_1 = \mathbf{t}_1' \mathbf{X}^{-2} \mathbf{t} \tag{6}$$

because it is tangent to $\mathscr{E}_0$, or

$$\mathbf{y}' \mathbf{t}_1 = \mathbf{y}' \mathbf{t} \tag{7}$$

because it is perpendicular to $\overrightarrow{A_0 A}$. The fitted values after one step are obtained by projecting $\overrightarrow{A_0 A}$ onto $\overrightarrow{A_0 B_1}$. If A_1 is the projection of A, the elliptical coordinates of A_1 are equal to the coefficients $\hat{\boldsymbol{\beta}}_1$, whereas the spherical coordinates are the fitted values $\hat{\mathbf{y}}_1$.

Now $\mathbf{X}\hat{\boldsymbol{\beta}}_1 = c\mathbf{t}_1$, where $c = (\mathbf{t}_1'\mathbf{t}_1)^{-1}\mathbf{t}_1'\mathbf{y}$; hence $0 \le c \le 1$. Note that c is nonnegative, since if it were negative, we would have $\mathbf{t}_1'\mathbf{y} < 0 < -\mathbf{t}_1'\mathbf{y}$ and $\mathbf{t}_1$ could not be the maximizing array. It follows that

$$\|\hat{\boldsymbol{\beta}}_1\|^2 = c^2 \mathbf{t}_1' \mathbf{X}^{-2} \mathbf{t}_1 = c^2 \mathbf{y}' \mathbf{X}^{-2} \mathbf{y} = c^2 \|\hat{\boldsymbol{\beta}}\|^2, \tag{8}$$

hence $\|\hat{\boldsymbol{\beta}}\| \ge \|\hat{\boldsymbol{\beta}}_1\|$, that is, the partial least squares estimates after one step shrink the least squares estimates. Geometrically this can be seen by noting that the radius of an ellipsoid concentric to $\mathscr{E}_0$ going through A_1 is necessarily smaller than or equal to that of $\mathscr{E}_0$. Hence the sum of the squared elliptical coordinates of A_1 is less than or equal to the sum of the squared elliptical coordinates of A.

In subsequent steps we will need the hyperplane Σ_1, perpendicular to $\overrightarrow{A_0 A}$, that goes through A. The hyperplane Σ_1 is parallel to the one given by (6) or (7) and separates the whole space into two half spaces. The distance of the point A_0 from Σ_1 is equal to $\|\mathbf{y}\|$; hence all points with coordinates $\mathbf{t}$ such that $\|\mathbf{t}\| \le \|\mathbf{y}\|$ lie in the half space containing A_0. In particular, since $\|\hat{\mathbf{y}}_s\| \le \|\mathbf{y}\|$, the points representing the fitted values of partial least squares at any step $s = 2, 3, \dots, a$ lie in the half space containing A_0.

Using a defining equation analogous to (6) rather than (7), since the hyperplanes are parallel and $\hat{\mathbf{y}}_1 = \mathbf{P}_1 \mathbf{y} = c\mathbf{t}_1$, the coordinates $\mathbf{t}$ of points of Σ_1 satisfy

$$(\mathbf{P}_1 \mathbf{y})' \mathbf{X}^{-2} \mathbf{y} = (\mathbf{P}_1 \mathbf{y})' \mathbf{X}^{-2} \mathbf{t}. \tag{9}$$

The half spaces into which Σ_1 separates the whole space can be defined by (9) with inequalities instead of equality, and since $(\mathbf{P}_1 \mathbf{y})' \mathbf{X}^{-2} \mathbf{y} \ge 0$, the half space containing A_0 is defined by

$$(\mathbf{P}_1 \mathbf{y})' \mathbf{X}^{-2} \mathbf{y} \ge (\mathbf{P}_1 \mathbf{y})' \mathbf{X}^{-2} \mathbf{t}. \tag{10}$$

Hence (10) is true for $\mathbf{t} = \hat{\mathbf{y}}_s$, $s = 2, 3, \dots, a$.

For the derivation of the two-step estimates, the orthogonality requirement $\mathbf{t}_1'\mathbf{t}_2 = 0$ restricts the vector with coordinates $\mathbf{t}_2$ to the $r - 1$ dimensional subspace R_2, perpendicular to $\overrightarrow{A_0B_1}$. The normalization (3) restricts the endpoints of the vector with coordinates $\mathbf{t}_2$ to an ellipsoid concentric to $\mathscr{E}_0$; hence we consider vectors with endpoints on $\mathscr{F}_1$, which lies in R_2 and has radius K_2 given by (3).

The projection of A onto R_2 is C_2 and is on $\mathscr{F}_1$. Hence finding $\mathbf{t}_2$ is equivalent to finding a vector with an endpoint on $\mathscr{F}_1$ maximizing the inner product with $\overrightarrow{A_0C_2}$. Similarly to the one-step case, the desired $\mathbf{t}_2$ is represented by $\overrightarrow{A_0B_2}$, where B_2 is the intersection of $\mathscr{F}_1$ and the $r-2$ dimensional affine in R_2 perpendicular to $\overrightarrow{A_0C_2}$ and tangent to $\mathscr{F}_1$. Projecting the point C_2 or A onto $\overrightarrow{A_0B_2}$ we obtain D_2, and adding $\overrightarrow{A_0A_1}$ and $\overrightarrow{A_0D_2}$ we obtain the vector $\overrightarrow{A_0A_2}$. The point A_2 represents the fitted values after two steps, so its elliptical coordinates are the parameter estimates $\hat{\boldsymbol{\beta}}_2$.

Using an argument similar to the one-step case, it follows that D_2 lies inside $\mathscr{F}_1$. Let $\mathscr{E}_1$ be the parallel displacement of $\mathscr{F}_1$, that is, the ellipsoid $\{E: \overrightarrow{A_0E} = \overrightarrow{A_0A_1} + \overrightarrow{A_0F},\ F \in \mathscr{F}_1\}$. Since $\overrightarrow{A_0C_2}$ is parallel to $\overrightarrow{A_1A}$, the normalization (3) is exactly the one that guarantees that A will lie on $\mathscr{E}_1$, which has radius K_2. Since $\mathscr{E}_1$ is the parallel displacement of $\mathscr{F}_1$, it follows that A_2 lies inside $\mathscr{E}_1$, so its coordinates $\hat{\mathbf{y}}_2$ satisfy

$$\text{(11)} \qquad \mathbf{y}'(\mathbf{I}-\mathbf{P}_1)\mathbf{X}^{-2}(\mathbf{I}-\mathbf{P}_1)\mathbf{y} \ge \hat{\mathbf{y}}_2'(\mathbf{I}-\mathbf{P}_1)\mathbf{X}^{-2}(\mathbf{I}-\mathbf{P}_1)\hat{\mathbf{y}}_2.$$

Furthermore, since $\overrightarrow{A_0D_2}$ is perpendicular to $\overrightarrow{A_0A_1}$, it follows that

$$\text{(12)} \qquad \mathbf{P}_1\hat{\mathbf{y}}_2 = \mathbf{P}_1\mathbf{y}.$$

Substituting $\mathbf{t} = \hat{\mathbf{y}}_2$ in (10) and using (11) and (12) we obtain

$$\text{(13)} \qquad \mathbf{y}'\mathbf{X}^{-2}\mathbf{y} \ge \hat{\mathbf{y}}_2'\mathbf{X}^{-2}\hat{\mathbf{y}}_2;$$

hence A_2 lies inside $\mathscr{E}_0$. The last statement is equivalent to $\|\hat{\boldsymbol{\beta}}\| \ge \|\hat{\boldsymbol{\beta}}_2\|$.

Now let Σ_2 be the hyperplane through A perpendicular to $\overrightarrow{A_1A}$ and $\overrightarrow{A_0C_2}$. Following arguments similar to the ones that led to (9), it is also parallel to a hyperplane tangent to $\mathscr{F}_1$ at the point B_2, so its coordinates satisfy

$$\text{(14)} \qquad (\mathbf{P}_2\mathbf{y})'\mathbf{X}^{-2}\mathbf{y} = (\mathbf{P}_2\mathbf{y})'\mathbf{X}^{-2}\mathbf{t}.$$

The hyperplane Σ_2 appears in Figure 1 as two lines—one going through A and the other going through C_2. It separates the space into two half spaces. From the inequality $\|\hat{\mathbf{y}}_s - \hat{\mathbf{y}}_1\| \le \|\mathbf{y} - \hat{\mathbf{y}}_1\|$ for $s > 1$, and the fact that the distance of A_1 from Σ_2 is equal to $\|\mathbf{y} - \hat{\mathbf{y}}_1\|$, it follows that the half space containing A_1 and A_0 also contains the partial least squares points at subsequent steps. For the coordinates $\mathbf{t}$ of all points in this half space, and hence for $\mathbf{t} = \hat{\mathbf{y}}_s$, $s = 3, \dots, a$,

$$\text{(15)} \qquad (\mathbf{P}_2\mathbf{y})'\mathbf{X}^{-2}\mathbf{y} \ge (\mathbf{P}_2\mathbf{y})'\mathbf{X}^{-2}\mathbf{t}.$$

The patient reader who has followed the arguments up to here should be convinced that the problem is identical in all following steps. Suppose that A_s are the points representing the fitted values at steps $s = 1, 2, \dots, a-1$. At step a, the orthogonality constraints on $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_a$ are equivalent to projecting A onto a space R_a perpendicular to $\overrightarrow{A_1 A_1}$, $\overrightarrow{A_1 A_2}, \dots, \overrightarrow{A_{a-2} A_{a-1}}$ and searching for a vector with coordinates $\mathbf{t}_a$ lying on R_a. Let C_a be the projection of A onto R_a. The normalisation (3) generates an ellipsoid $\mathscr{F}_{a-1}$ of radius K_a, concentric to $\mathscr{E}_0$ but of lower dimension. The desired direction of the vector is, say, $\overrightarrow{A_0 B_a}$, where the hyperplane which is tangent to $\mathscr{F}_{a-1}$ is perpendicular to $\overrightarrow{A_0 C_a}$. Using an argument similar to the one-step case, we can show that the projection of $\overrightarrow{A_0 C_a}$ onto $\overrightarrow{A_0 B_a}$ is inside $\mathscr{F}_{a-1}$. This implies that, after displacing $\mathscr{F}_{a-1}$ by the vector $\overrightarrow{A_0 A_a}$ to obtain $\mathscr{E}_{a-1}$, A_a is inside $\mathscr{E}_{a-1}$, so

$$\begin{aligned} &\mathbf{y}'\left(\mathbf{I} - \sum_{s=1}^{a-1} \mathbf{P}_s\right)\mathbf{X}^{-2}\left(\mathbf{I} - \sum_{s=1}^{a-1} \mathbf{P}_s\right)\mathbf{y} \\ &\qquad \geq \hat{\mathbf{y}}_a'\left(\mathbf{I} - \sum_{s=1}^{a-1} \mathbf{P}_s\right)\mathbf{X}^{-2}\left(\mathbf{I} - \sum_{s=1}^{a-1} \mathbf{P}_s\right)\hat{\mathbf{y}}_a. \end{aligned} \tag{16}$$

Furthermore, the orthogonality constraints on $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_a$ imply

$$\mathbf{P}_s \hat{\mathbf{y}}_a = \mathbf{P}_s \mathbf{y} \tag{17}$$

for $s = 1, 2, \dots a$. Considering the hyperplanes $\Sigma_1, \Sigma_2, \dots, \Sigma_{a-1}$ that are perpendicular to $\overrightarrow{A_0 A}, \overrightarrow{A_1 A}, \dots, \overrightarrow{A_{a-2} A}$, respectively, and using arguments similar to the ones that led to (10) and (15), we obtain

$$(\mathbf{P}_s \mathbf{y})'\mathbf{X}^{-2}\mathbf{y} \geq (\mathbf{P}_s \mathbf{y})'\mathbf{X}^{-2}\hat{\mathbf{y}}_a \tag{18}$$

for $s = 1, 2, \dots, a - 1$. A simple algebraic manipulation of (16)–(18) implies $\mathbf{y}'\mathbf{X}^{-2}\mathbf{y} \geq \hat{\mathbf{y}}_a'\mathbf{X}^{-2}\hat{\mathbf{y}}_a$ and the assertion follows. □

4. Discussion. The fact that partial least squares estimators shrink classifies them to a large class of widely used estimators that perform well with highly collinear data. An immediate result is that their estimation mean squared error will be better than that of ordinary least squares if the true parameter values are small. One can also hope that they will have a smaller prediction mean squared error. Obtaining analytical results concerning the range of parameters for which there will be an improvement seems a hard, if not impossible, task, but for most practical problems with large p one would expect that most of the true coefficients will be small. Hence, achieving better results can reasonably be expected.

Since typically the $\mathbf{X}$ matrix and the $\mathbf{y}$ vector are centered, the shrinkage toward the origin of the transformed vectors translates to shrinkage toward the sample means of the original variables, whereas the intercept remains constant. However, the kind of shrinkage is by no means as obvious as in ridge or principal components regression [cf. Frank and Friedman (1993)].

The direction of the vector of parameter estimates depends on the response variable in a complicated way and from the geometrical picture it can be seen that shrinking does not apply to all parameter components. Indeed one can have components that expand, but this should not be considered a deficiency. If not all components of estimates are shrunken by the same proportion, one can find a coordinate system in which some coordinates expand for some data.

The complicated shrinking nature of partial least squares should not come as a surprise, since the method uses information about variances and covariances of both explanatory and response variables. However the geometry can give us some idea about their behavior [see also Phatak, Reilly and Penlidis (1992)]. Principal components explaining a large proportion of the variance in the explanatory variables will tend to draw all partial least squares fitted values toward them, since tangent planes will tend to be closer to peaks of the ellipsoids. Nevertheless they will have no effect if the corresponding least squares estimates (in the canonical form) are zero or near zero. For highly collinear explanatory variables the lower dimensional ellipsoids will be very small after some steps. This is in common with principal components regression, but involving the response variable tilts the ellipsoids in each step so that the effect will be stronger. On the other hand if the explanatory variables are orthogonal or nearly orthogonal to each other, the ellipsoids are spheres and partial least squares will be nearly equal to least squares. This gives a geometric interpretation to the empirical fact that partial least squares regression “saturates” faster than principal components.

Since the original submission of this paper, we have become aware of another algebraic proof of the shrinkage of the partial least squares coefficients relative to the ordinary least squares estimator [de Jong (1995)]. Indeed de Jong proves the somewhat stronger result that the length of partial least squares coefficients is a nondecreasing function of the step a, that is, $\|\hat{\boldsymbol{\beta}}_a\| \leq \|\hat{\boldsymbol{\beta}}_{a+1}\|$ for every $a = 1, 2, \dots, r-1$.

Acknowledgments. I would like to thank Tom Fearn for suggesting the problem and Mervyn Stone for patiently listening to me, for correcting an earlier version of Figure 1 and for his careful reading of the manuscript. I would also like to thank the Editor, an Associate Editor and two anonymous referees for suggestions that led to an improved version of the paper.

REFERENCES

DE JONG, S. (1995). PLS shrinks. *Journal of Chemometrics* **9** 323–326.

DENHAM, M. C. (1991). Calibration in infrared spectroscopy. Ph.D. dissertation, Univ. Liverpool.

FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35** 109–147.

HELLAND, I. S. (1988). On the structure of partial least squares regression. *Comm. Statist. Simulation Comput.* **17** 581–607.

MARTENS, H. and NAES, T. (1989). *Multivariate Calibration*. Wiley, New York.

NAES, T. and MARTENS, H. (1985). Comparison of prediction methods for multicollinear data. *Comm. Statist. Simulation Comput.* **14** 545–576.

PHATAK, A., REILLY, P. M. and PENLIDIS, A. (1992). The geometry of 2-block partial least squares regression. *Comm. Statist. Theory Methods* **21** 1517–1553.

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York.

STONE, M. and BROOKS, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. Roy. Statist. Soc. Ser. B* **52** 237–269.

SUNDBERG, R. (1993). Continuum regression and ridge regression. *J. Roy. Statist. Soc. Ser. B* **55** 653–659.

WOLD, H. (1966). Nonlinear estimation by iterative least squares procedures. In *Research Papers in Statistics. Festschrift for J. Neyman* (F. N. David, ed.) 411–444. Wiley, New York.

WOLD, H. (1973). Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In *Multivariate Analysis* (P. R. Krishnaiah, ed.) **3** 383–407. Academic Press, New York.

DEPARTAMENTO DE ESTADISTICA Y
ECONOMETRIA
UNIVERSIDAD CARLOS III
DE MADRID
CALLE MADRID 126
28903 GETAFE , MADRID
SPAIN