Significance Regression: A Statistical Approach to Biased Linear Regression and Partial Least Squares

Tyler R. Holcomb   Håkan Hjalmarsson *

Manfred Morari †

Control and Dynamical Systems 210-41

California Institute of Technology

Pasadena CA 91125

Keywords: biased regression, PLS, multivariable regression, significance regression, collinearity

CIT-CDS Technical Memo 93-002

May 12, 1993

Abstract

This paper first examines the properties of biased regressors that proceed by restricting the search for the optimal regressor to a subspace. These properties suggest features such biased regression methods should incorporate. Motivated by these observations, this work proposes a new formulation for biased regression derived from the principle of statistical significance. This new formulation, significance regression (SR), leads to partial least squares (PLS) under certain model assumptions and to more general methods under various other model assumptions. For models with multiple outputs, SR will be shown to have certain advantages over PLS. Using the new

*on leave from Linkoping University, Sweden

†Author to whom correspondence should be addressed: phone (818)356-4186, fax (818)568-8743, e-mail mm@imc.caltech.edu

formulation a significance test is advanced for determining the number of directions to be used; for PLS, cross-validation has been the primary method for determining this quantity. The prediction and estimation properties of SR are discussed. A brief numerical example illustrates the relationship between SR and PLS.

1 Introduction

This paper studies the linear regression model

$$y = Xr + e, \tag{1}$$

where $X \in \Re^{n_s \times n_i}$ and $y \in \Re^{n_s}$ are known, $r \in \Re^{n_i}$ is an unknown vector, and $e \in \Re^{n_s}$ is an unobservable error vector. Unless otherwise stated, $\mathcal{E}\left(e\right) = 0$ and $\mathcal{E}\left(ee^T\right) = \sigma_e^2 I$ script E where $\mathcal{E}\left(\cdot\right)$ denotes the expectation. Throughout this paper the variables that comprise X will be call the "inputs" and y will be called the "output." This is merely a convenient nomenclature; these variables can equivalently be called "explanatory variables" and the "dependent variable." This paper examines both the prediction problem (determining a $\tilde{b}$ that can be used to predict future y's from future X's) and the estimation problem (determining a $\tilde{b}$ that is close to r in some well-defined sense). In particular, this paper investigates linear regression methods wherein the search for $\tilde{b}$ is confined to a subspace of $\Re^{n_i}$. Section 4 extends the results of the paper to multiple output problems.

For the model in equation 1 the minimum-variance unbiased estimate of r has long been known to result from ordinary least-squares (OLS) regression, namely

$$\tilde{r} = (X^T X)^{-1} X^T y. \tag{2}$$

Equation 2 assumes $(X^T X)$ is non-singular; this assumption is maintained throughout the paper. The variance of $\tilde{r}$, $\mathrm{Var}(\tilde{r}) = \sigma_e^2 (X^T X)^{-1}$, can be unacceptably large, especially if collinearities are present in the input data. The variance can be reduced if the search for the estimate of the regression vector, $\tilde{b}$, is restricted to a subspace of $\Re^{n_i}$. Let the orthonormal columns of $W \in \Re^{n_i \times n_w}$, where $n_i \geq n_w$, span the allowable range for the regression vector estimate. That is, let

$$\tilde{b} = \arg \min_{b \in \mathrm{Range}(W)} ||y - Xb||^2. \tag{3}$$

The regressor that satisfies the search constraints is

$$\tilde{b} = W(W^T X^T X W)^{-1} W^T X^T y. \tag{4}$$

For convenience, we will call this class of regressors "restriction regressors." A variety of widely-used regression methods belong to this class including stepwise regression [Draper and Smith, 1966], Principal Components Regression (PCR) [Joliffe, 1982], and Partial Least Squares (PLS) [Wold et al., 1984]. Additionally, several restriction regressors have been put forward recently including Continuum Regression [Stone and Brooks, 1990] and a related method due to Lorber and coworkers [Lorber et al., 1987]. All of these methods have been successful on specific examples. However, many of these methods have heuristic motivations and are difficult to describe, analyze, or compare using classical statistical methods. Important progress has recently been made developing a statistical view for some of these restriction regressors. In particular, Friedman has developed a Bayesian viewpoint that compares restriction regressors via a "shrinkage" analysis [Frank and Friedman, 1992] and Helland has derived a particular restriction regressor from the maximum likelihood principle [Helland, 1992]. Still our understanding of restriction regressors (and PLS in particular) remains incomplete. Additionally this lack of a statistical foundation has led many of these methods, particularly PLS-related approaches, to rely on cross-validation to answer questions traditionally analyzed by hypothesis testing.

This paper begins by reviewing properties of restriction regressors. Next, a new restriction regression method, Significance Regression (SR), is developed from the classical concept of statistical significance. This method rests directly on a testable null hypothesis and leads to PLS. The SR method is next extended to problems with multiple outputs. For a generalization that emphasizes incrementally building meaningful subspaces of the inputs and outputs (*i.e.* factor analysis), PLS is similar to SR. For a generalization that emphasizes building "significant" regressors, SR produces a new regression method. Lastly, two numerical examples are presented to illustrate the similarities and differences between PLS and SR.

2 Properties of "Restriction Regressors"

Before discussing the properties of restriction regressors, we define two performance measures: the Mean Squared Error (MSE) and the PRedicted Error Sum of Squares (PRESS).

The MSE is defined as

$$\mathrm{MSE}(\tilde{b}) = \mathcal{E}\left(\mathrm{Tr}(\ (\tilde{b}-r)(\tilde{b}-r)^T\ )\right) \tag{5}$$

while the PRESS is defined as

$$\mathrm{PRESS}(\tilde{b}) = (\tilde{y}-y)^T(\tilde{y}-y) \tag{6}$$

where $\tilde{y} = X\tilde{b}$. Typically the X and y used for evaluating the PRESS are different from the X and y used to create $\tilde{b}$. To increase understanding of the PRESS, consider computing the PRESS using a new set of data, X_{new}. Then

$$\mathcal{E}\left(\mathrm{PRESS}(\tilde{b})\right) = \mathcal{E}\left(\mathrm{Tr}(\ (\tilde{b}-r)^T X_{new}^T X_{new}(\tilde{b}-r)\ )\right) + \mathcal{E}\left(\mathrm{Tr}(e^T e)\right) \tag{7}$$

In this light, the PRESS and MSE are clearly related. As shown by Gruber [Gruber, 1990] this relation can be made precise by means of Theobald's Theorem [Theobald, 1974]. Briefly, if for two estimates $\tilde{b}_1$ and $\tilde{b}_2$ the difference

$$\mathcal{E}\left((\tilde{b}_1-r)(\tilde{b}_1-r)^T\right) - \mathcal{E}\left((\tilde{b}_2-r)(\tilde{b}_2-r)^T\right) \tag{8}$$

is a non-negative definite matrix, then $\mathcal{E}\left(\mathrm{PRESS}(\tilde{b}_1)\right) \geq \mathcal{E}\left(\mathrm{PRESS}(\tilde{b}_2)\right)$. Additionally, one can see that using equation 7 involves assuming future inputs will be "similar" to past inputs. If $X_{new}^T X_{new}$ is descriptive of future X's, then equation 7 is clearly a measure of predictive performance. However, if the future inputs will have markedly different characteristics, or if the point of computing $\tilde{b}$ is to estimate r, then the PRESS may give misleading indications. Thus, the suitability of the PRESS or MSE will depend on the application. See, for example, chapters 12 and 13 of Ljung [Ljung, 1987].

Throughout the remainder of this section, assume that $W \in \Re^{n_i \times n_w}$ has been chosen independent of the errors and that the columns of W are orthonormal. Additionally, define $W^\perp$ such that $[W\ W^\perp]^T[W\ W^\perp] = I$ and $r = Wq_1 + W^\perp q_2$, where $q_1 \in \Re^{n_w}$ and $q_2 \in \Re^{n_i - n_w}$. One outstanding property of restriction regressors is

$$\mathrm{Var}(\tilde{r}) \geq \mathrm{Var}(WW^T\tilde{r}) \geq \mathrm{Var}(\tilde{b}). \tag{9}$$

One expects that since $\tilde{b}$ can only vary over a subspace of $\Re^{n_i}$ that the variance of a restriction regressor would be less than the variance of the OLS regressor. However, even when $\tilde{r}$ is projected into Range(W) the variance of the resulting regressor still dominates the variance of the corresponding restriction regressor. This key fact, which is the source of much of the MSE and PRESS advantage of restriction regressors, is proven in appendix C.1. When the bias is "small," the variance advantage of restriction regressors can lead to dramatic improvements over OLS.

Since $\tilde{b}$ is computed with the constraint $\tilde{b} \in \text{Range}(W)$, one expects $\tilde{b}$ to be a biased estimate of r. However, in addition to being unable to account for any of the $r \in \text{Range}(W^{\perp})$, $\tilde{b}$ has an additional bias due to trying to "stretch" in Range(W) to account for the "missing" r. This "stretching" occurs because the least squares estimator when the search space is restricted to a subspace of $\Re^{n_i}$ is not the projection of $\tilde{r}$ into the search space. Consider for the moment the case $\sigma_e^2 = 0$. Then

$$\begin{aligned} \tilde{b} &= \arg \min_{b \in \text{Range}(W)} (y - Xb)^T (y - Xb) && (10) \\ &= W(W^T X^T X W)^{-1} W^T X^T X r, && (11) \end{aligned}$$

which is different from the projection of r into the search space, $WW^T r$. In general $\mathcal{E}\left(\tilde{b}\right) = W(W^T X^T X W)^{-1} W^T X^T X r$. Thus, there will typically be two bias terms.

$$\begin{aligned} \mathcal{E}\left(\tilde{b} - r\right) &= (W(W^T X^T X W)^{-1} W^T X^T X - I) r && (12) \\ &= (W(W^T X^T X W)^{-1} W^T X^T X - I)(W q_1 + W^{\perp} q_2) && (13) \\ &= W(W^T X^T X W)^{-1} W^T X^T X W^{\perp} q_2 - W^{\perp} q_2. && (14) \end{aligned}$$

The first bias term is due to the "stretching " of $\tilde{b}$ to account for variations in the output attributable to q_2 that can be partially described from Range(W), while the second bias term is the direct contribution of q_2, the portion of r that lies outside Range(W). Notice that the two bias terms in equation 14 have complementary range spaces. Thus if $q_2 \neq 0$ then the two bias terms can never "cancel out"; however for the special case of PCR, $W^T X^T X W^{\perp} = 0$.

Consider the situation where $\tilde{b}$ has been constructed and a new input vector x_{new} is available. Computing the PRESS using x_{new} leads to

$$\begin{aligned} \mathcal{E}\left((r^T x_{new} + e_{new} - \tilde{b}^T x_{new})^2\right) &= \left(x_{new}^T (W(W^T X^T X W)^{-1} W^T X^T X - I) r\right)^2 \\ &\quad + x_{new}^T W (W^T X^T X W)^{-1} W^T x_{new} \sigma_e^2 + \sigma_e^2 && (15) \\ &= \left(x_{new}^T (W(W^T X^T X W)^{-1} W^T X^T X W^{\perp} q_2 - W^{\perp} q_2)\right)^2 \\ &\quad + x_{new}^T W (W^T X^T X W)^{-1} W^T x_{new} \sigma_e^2 + \sigma_e^2. && (16) \end{aligned}$$

Clearly, when $q_2 = 0$ then the bias terms vanish and $\tilde{b}$ is superior to $\tilde{r}$. For prediction, even if there is a large bias term, the prediction bias will still tend to be small if x_{new} is not collinear with the bias. Thus restriction regressors are attractive for prediction problems even when one can not assure that the bias is "small." One should note that requiring $x_{new} \in \text{Range}(W)$ does not assure unbiased prediction; there will often still be bias arising from the "stretching" term.

Next, consider $\tilde{b}$ as a point estimator for r. In particular, evaluate

$$\begin{aligned}
\text{MSE}(\tilde{b}) &= \text{Tr}\left(\mathcal{E}\left((r-\tilde{b})(r-\tilde{b})^T\right)\right) && (17)\\
&= \|(W(W^TX^TXW)^{-1}W^TX^TX - I)r\|^2 && (18)\\
&\quad + \sigma_e^2\text{Tr}\left(W(W^TX^TXW)^{-1}W^T\right)\\
&= \|(W(W^TX^TXW)^{-1}W^TX^TXW^{\perp}q_2 - W^{\perp}q_2\|^2 && (19)\\
&\quad + \sigma_e^2\text{Tr}\left(W(W^TX^TXW)^{-1}W^T\right).
\end{aligned}$$

As in the point prediction problem, two bias terms associated with q_2 appear. For the restriction estimator to have an MSE advantage over OLS, q_2 must be "small." Thus a successful restriction regressor (in the MSE sense) must strive to satisfy $r \in \text{Range}(W)$.

Move next to the problem of estimating an interval within which $r^T x_{new}$ lies with a certain probability. If one assumes that e is normally distributed, then one can use $\tilde{r}$ in the classical manner and declare that $\tilde{r}^T x_{new} - r^T x_{new}$ has a zero-mean normal distribution with variance $x_{new}^T(X^TX)^{-1}x_{new}\sigma_e^2$. Using, $\tilde{b}$, one can easily see that $r^T x_{new} - \tilde{b}^T x_{new}$ has a normal distribution with mean $x_{new}^T W(W^TX^TXW)^{-1}W^TX^TX - I)r$ and variance $x_{new}^T W(W^TX^TXW)^{-1}W^T x_{new}\sigma_e^2$. Since r enters explicitly into the mean of the distribution, this can not be used directly for a practical interval predictor. However if one is willing to conjecture that the bias is "small," then one can assume $r^T x_{new} - \tilde{b}^T x_{new}$ has a zero-mean normal distribution with variance $x_{new}^T W(W^TX^TXW)^{-1}W^T x_{new}\sigma_e^2$. Under this assumption (which is equivalent to $q_2 = 0$), one has a smaller prediction interval using $\tilde{b}$ than when using $\tilde{r}$ since $(X^TX)^{-1} \geq W(W^TX^TXW)^{-1}W^T$, as shown in the appendix C.1.

Lastly, turn to the problem of constructing a region of $\Re^{n_i}$ within which r lies with a certain probability. The classical approach is based on the observation that if e is normally distributed then

$$\frac{(\tilde{r}-r)^T X^T X(\tilde{r}-r)}{\sigma_e^2} \tag{20}$$

has a χ^2 distribution with n_i degrees of freedom. One can adapt this approach to restriction regressors in a variety of ways, but the resulting confidence regions are generally more conservative. This outcome is not surprising. The traditional confidence bounds provide a region which contains r with a certain probability. A restriction regressor may choose a "better" point estimator from this region than the OLS point estimate, but the restriction regression process typically does not alter the original confidence region.

Returning to the issue of building the biased regressor, the difficulty with equation 4 is that one must specify W. One approach for specifying W is to choose W so as to optimize the MSE:

$$W^{\mathrm{MSE}} = \arg\min_{W\in\Re^{n_o\times n_i}} \mathrm{MSE}(\tilde{b}). \tag{21}$$

The solution is $W^{\mathrm{MSE}} = \frac{r}{||r||}$ [Bibby and Toutenburg, 1977]. This "optimal" solution is interesting but not very useful because r is unknown, leaving the question of how to choose W unresolved. Still this "optimal" solution, as well as the properties discussed above, highlight the key feature one desires from a restriction regressor: that $q_2 = {W^{\perp}}^T r = 0$.

Instead of computing the MSE-optimal W directly, one could try to build W incrementally by some other criterion in the hope of constructing a W such that $r \in \mathrm{Range}(W)$. Consider for the moment an additional vector w such that $W^T w = 0$ and $||w|| = 1$. One then asks: does using the search space $\mathrm{Range}([W|w])$ lead to a $\tilde{b}$ with an MSE less than or equal to the MSE of $\tilde{b}$ using the search space $\mathrm{Range}(W)$? The answer is "yes" if

$$(w^T r)^2 \geq \mathrm{Tr}\left([W|w]\left(([W|w]^T X^T X)[W|w]\right)^{-1}[W|w]^T - W(W^T X^T X W)^{-1} W^T\right)\sigma_e^2. \tag{22}$$

The left hand side clearly reflects the bias removed from $\tilde{b}$ and the right hand side reflects the variance added to $\tilde{b}$ by including $\mathrm{Range}(w)$ in the search space. Although equation 22 is not directly useful because r appears explicitly, it points the way: directions for which $w^T r = 0$ should not be included in the search space, and directions for which $w^T r$ is "large" and the variance is "small" should be included in the search space. These two observations provide the motivation for the particular restriction regressor developed in this paper.

3 Significance Regression for Scalar Output Problems

This section develops significance regression (SR) for scalar output problems. The first subsection develops a method for evaluating the left-hand side (LHS) of equation 22. The result is a test statistic which forms the basis for SR. The second subsection derives the algorithm for generating W; this is the heart of the SR method. The next subsection presents a hypothesis testing approach for determining n_d, the number of columns to use in W. The final subsection discusses additional properties of SR.

3.1 Testing the significance of a given w

As revealed in equation 22, one would like to evaluate if a particular subspace, described by w, is orthogonal to r. If the two vectors are perpendicular, then one should not include such a w in the search space Range(W) This criterion translates directly into the null hypothesis script H

$$\mathcal{H}_0^1 : \qquad < r, w > \, = 0 \tag{23}$$

where $< r, w >$ is the vector inner product, $r^T w$. For this null hypothesis, a natural test statistic is

$$\tau(w, y) = \frac{< \tilde{r}, w >^{\,2}}{\text{Var} < \tilde{r}, w >} \tag{24}$$

where

$$\begin{aligned} \text{Var} < \tilde{r}, w > &= \mathcal{E}\left((< \tilde{r}, w > \, - \mathcal{E}\,(< \tilde{r}, w >))^2 \right) && (25) \\ &= \mathcal{E}\left(< (X^T X)^{-1} X^T e, w >^{\,2} \right) && (26) \\ &= w^T (X^T X)^{-1} w \sigma_e^2. && (27) \end{aligned}$$

This $\tau(w, y)$ leads to simple test for $\mathcal{H}_0^1$. If w was chosen independently of y and the errors (elements of e) are normally distributed, then $\tilde{r}$ is normally distributed and $\tau(w, y)$ has a χ^2 distribution with one degree of freedom. If the noise variance σ_e^2 is unknown, then the unbiased estimate $\tilde{\sigma}_e^2 = \frac{1}{n_s - n_i}(y - X\tilde{r})^T (y - X\tilde{r})$ may be used instead; for normal errors, $\tilde{\sigma}_e^2$ arises from a χ^2 distribution. Since $< \tilde{r}, w >^{\,2}$ varies as the error vector e projected into Range(X), while $\tilde{\sigma}_e^2$ varies as the error vector e projected into the orthogonal complement of Range(X), the two terms are independent. If $\tilde{\sigma}_e^2$ is used in equation 24 then $\tau(w, y)$ is associated with the Snedecor's F-distribution. Throughout this paper $\tilde{\sigma}_e^2$ can be used in place of σ_e^2 and the relevant distributions modified in the obvious manner.

Using these distributions one can use classical significance testing procedures to evaluate $\mathcal{H}_0^1$ and to identify w that should be excluded from W. This technique is currently used successfully in the context of principal component regression. In PCR, the W is formed from a subset of the eigenvectors of X^TX; thus, any w considered for inclusion in W is independent of y. One chooses a significance threshold and includes in W only those eigenvectors w for which $\tau(w,y)$ rejects $\mathcal{H}_0^1$. This is identical to the PCR approach in which one uses only principal components that are "significant" for prediction [Massy, 1965].

3.2 The Significance Regression method

The previous subsection discussed how to use $\tau(w,y)$ to choose from among a set of prespecified w's. However, equation 22 revealed that one should seek w for which "r is large and the variance is small." One can do this directly by computing w that cause $\tau(w,y)$ to be "large." In this sense the "most significant subspace" is described by

$$w_1^{opt} = \arg \max_{w \in \Re^{n_i}} \tau(w,y). \tag{28}$$

Equation 24 reveals that w_1^{opt} will not be unique; multiplying any w by a scalar will not affect the value of $\tau(w,y)$. Still, the necessary condition for an unconstrained extremum

$$\nabla_w \; \tau(w,y)|_{w_1^{opt}(y)} = 0 \tag{29}$$

must be met. Computing the gradient of $\tau(w,y)$ gives

$$\begin{aligned} \nabla_w \tau(w,y) &= \nabla_w \frac{(w^T\tilde{r})^2}{w^T(X^TX)^{-1}w\sigma_e^2} && (30) \\ &= \frac{2\tilde{r}(w^T\tilde{r})w^T(X^TX)^{-1}w\tilde{\sigma}_e^2 - 2(w^T\tilde{r})^2(X^TX)^{-1}w\sigma_e^2}{(w^T(X^TX)^{-1}w\tilde{\sigma}_e^2)^2} && (31) \end{aligned}$$

and applying equation 29 gives

$$w_1^{opt}(y) = X^TX\tilde{r} = X^Ty. \tag{32}$$

As discussed in section 2, the goal is to determine a W such that $r \in \text{Range}(W)$. Therefore W may need to have more than one column. Consistent with section 2, one searches for w that maximize $\tau(w,y)$ and are orthogonal to the current $W_{i-1} = [w_1^{opt}|\ldots|w_{i-1}^{opt}]$. Then the ith "significant subspace" is described by

$$w_i^{opt} = \arg \max_{w \in \text{Range}(I - W_{i-1}W_{i-1}^T)} \tau(w,y). \tag{33}$$

Invoking the necessary condition for a constrained extremum yields

$$(I - W_{i-1}W_{i-1}^T) \quad \nabla_w \ \tau(w, y)|_{w_i^{opt}(y)} = 0. \tag{34}$$

As shown in appendix C.2, the PLS loading vectors satisfy equation 34. Thus one can use PLS to find n_d "significant vectors". Here we employ Helland's algorithm [Helland, 1988].

Algorithm 1 (Significance Regression for Scalar Output Problems)

$$\tilde{r} = \quad (X^T X)^{-1} X^T y \tag{35}$$

$$W_0 = \quad [0 \cdots 0]^T, \quad W_0 \in \Re^{n_i} \tag{36}$$

$$\text{DO} \quad i = \quad 1, n_d$$

$$v = \quad (I - W_{i-1}W_{i-1}^T)(X^T X)^i \tilde{r} \tag{37}$$

$$= \quad (I - W_{i-1}W_{i-1}^T)(X^T X)^{i-1} X^T y \tag{38}$$

$$w_i^{opt}(y) = \quad \frac{v}{||v||} \tag{39}$$

$$W_i = \quad [w_1^{opt} | w_2^{opt} | \cdots | w_i^{opt}] \tag{40}$$

$$\text{END DO.}$$

$$\tilde{b} = \quad W_{n_d}(W_{n_d}^T X^T X W_{n_d})^{-1} W_{n_d}^T X^T y \tag{41}$$

3.3 Choosing n_d

The above developments have assumed that n_d is known; however, in practice n_d needs to be determined. In the PLS context, the most popular method is cross-validation [Wold, 1978]. For any given n_d, $\tilde{b}$ is computed with a subset of the available data, and the PRESS is computed for that $\tilde{b}$ using the remainder of the the data. Next the PRESS for different n_d's is compared to determine the "best" value of n_d. $\tilde{b}$ is then recomputed using all available data and the "best" n_d. As discussed in section 2, the PRESS is an intuitively appealing measure when one is building predictors, but may not necessarily be the best measure for evaluating estimates of r. Clearly practitioners would benefit from the development of additional techniques for choosing n_d.

As shown in section 2, a useful condition for any restriction regressor to satisfy is $q_2 = 0$. This leads directly to the null hypothesis

$$\mathcal{H}_0^{2,i} : \qquad < r, w > = 0 \ \ \forall w \in \text{Range}(I - W_{i-1}W_{i-1}^T) \tag{42}$$

for evaluating if $n_d = i - 1$. Let $\tau_i^{opt}(y) = \tau(w_i^{opt}, y)$. If

$$\Pr\left\{\ t \le \tau_i^{opt}(y)\ \right\} \ge \alpha_{thresh}, \tag{43}$$

where α_{thresh} is some pre-specified significance threshold and t is drawn from the distribution for $\tau_i^{opt}(y)$ when $\mathcal{H}_0^{2,i}$ holds, then $\mathcal{H}_0^{2,i}$ can be rejected and $n_d \ge i$. Since W_{i-1} depends on y, evaluation of this distribution can be involved. However as shown below, when the elements of e are independently identically normally distributed $\tau_1^{opt}(y)$ has a χ^2 distribution with n_i degrees of freedom and the distribution of $\tau_i^{opt}(y)$ can be effectively approximated by a χ^2 distribution with $n_p = n_i - i + 1$ degrees of freedom.

The remainder of this subsection assumes that the elements of e are independently identically normally distributed. We know from algorithm 1 that $w_1^{opt}(y) = X^T y \ / \ \|X^T y\|$. Turning attention to the $\tau_1^{opt}(y)$ that results from maximizing the significance criterion,

$$\begin{aligned}
\tau_1^{opt}(y) &= \tau(w_1^{opt}(y), y) && (44)\\
&= \frac{(y^T X(X^T X)^{-1} X^T y)^2}{y^T X(X^T X)^{-1} X^T y \sigma_e^2} && (45)\\
&= \frac{y^T X(X^T X)^{-1} X^T y}{\sigma_e^2} && (46)\\
&= \frac{\left((X^T X)^{\frac{1}{2}} r + (X^T X)^{-\frac{1}{2}} X^T e\right)^T \left((X^T X)^{\frac{1}{2}} r + (X^T X)^{-\frac{1}{2}} X^T e\right)}{\sigma_e^2}. && (47)
\end{aligned}$$

Thus $\tau_1^{opt}(y)$ has a non-central χ^2 distribution with n_i degrees of freedom and non-centrality parameter $r^T X^T X r$. Once again, if $\tilde{\sigma}_e^2$ is used in place of σ_e^2, then $\tau_1^{opt}(y)$ has a non-central F distribution with $(n_i, n_s - n_i)$ degrees of freedom.

When $\mathcal{H}_0^{2,1}$ holds ($r = 0$), then

$$\tau_1^{opt}(y) \ = \ \frac{e^T X(X^T X)^{-1} X^T e}{\sigma_e^2}, \tag{48}$$

and $\tau_1^{opt}(y)$ has a χ^2 distribution with n_i degrees of freedom. If $\tau_1^{opt}(y)$ rejects $\mathcal{H}_0^{2,1}$ then either an unusual e has occurred (*eg.* an outlier) that made $\tau_1^{opt}(y)$ unduly large or the non-centrality parameter $r^T X^T X r$ is "large" relative to the likely values of $\tau_1^{opt}(y)$ under $\mathcal{H}_0^{2,1}$. Thus, $\mathcal{H}_0^{2,1}$ tends to be rejected when the value of $\tau_1^{opt}(y)$ is dominated by the (deterministic) non-centrality parameter rather than the (probabilistic) noise. This "deterministic dominance" is reflected in $w_1^{opt}(y)_{\bar{2}} X^T y \ / \ \|X^T y\|$ because $X^T y = X^T X r + X^T e$: as $r^T X^T X r$ becomes large, $w_1^{opt}(y)$ is less influenced by e. Thus, if a direction exists which strongly refutes $\mathcal{H}_0^{2,1}$, then SR will tend to identify this direction, and the subspace represented by $w_1^{opt}(y)$ will tend to be weakly affected by e.

When $\mathcal{H}_0^{2,1}$ is refuted, one would include $w_1^{opt}(y)$ in W and evaluate the "second-most significant subspace," w_2^{opt}. Once again, the question of computing the distribution for $\tau_2^{opt}(y)$ arises. Strictly speaking $w_1^{opt}(y)$ is not independent of y. However, "$w_1^{opt}(y)$ will tend to be weakly affected by e" so one may compute the distribution assuming w_1^{opt} is independent of y. Let $\text{Range}(W_i^{\perp}) = \text{Range}(I - W_{i-1}W_{i-1}^T)$ and ${W_{i-1}^{\perp}}^T W_{i-1}^{\perp} = I$. Under the assumption that W_{i-1} is independent of e,

$$w_i^{opt}(y) = W_{i-1}^{\perp}\left({W_{i-1}^{\perp}}^T (X^TX)^{-1} W_{i-1}^{\perp}\right)^{-1} {W_{i-1}^{\perp}}^T \tilde{r} \tag{49}$$

and $\tau_i^{opt}(y)$ has a non-central χ^2 distribution (as above) with $n_p = n_i - i + 1$ degrees of freedom and non-centrality parameter $r^T W_{i-1}^{\perp}\left({W_{i-1}^{\perp}}^T (X^TX)^{-1} W_{i-1}^{\perp}\right)^{-1} {W_{i-1}^{\perp}}^T r$. Once again, if $\mathcal{H}_0^{2,i}$ holds, then $\tau_i^{opt}(y)$ has a χ^2 distribution, but if a direction strongly violates the null hypothesis, $w_i^{opt}(y)$ will be relatively unaffected by e.

These observations have tangible implications. If one wishes to evaluate $\mathcal{H}_0^{2,i}$ using $\tau_i^{opt}(y)$, one can consider approximating the distribution of $\tau_i^{opt}(y)$ with a χ^2 distribution with n_p degrees of freedom. In fact, such an approximation is valid in several asymptotic limits. Clearly, as the noise vanishes, the dependence of $w_i^{opt}(y)$ on e vanishes. That is

$$\lim_{\sigma_e^2 \to 0} \text{Span}\left(W_i^{opt}(y)\right) = \text{Range}\left([X^TXr|\ldots|(X^TX)^i r]\right). \tag{50}$$

Thus, when the noise is small enough, the independence assumption is justified. The independence assumption can also be justified when n_s is large. Consider again $w_1^{opt}(y)$ and the condition that the input data is persistently exciting ($\lim_{n_s\to\infty} \frac{1}{n_s} X^TX = V$ for some non-singular V). Then

$$\begin{aligned}
\lim_{n_s\to\infty} w_1^{opt}(y) &= \lim_{n_s\to\infty} \frac{\frac{1}{n_s}X^Ty}{\left\|\frac{1}{n_s}X^Ty\right\|} && (51)\\
&= \lim_{n_s\to\infty} \frac{\frac{1}{n_s}X^TXr + \frac{1}{n_s}X^Te}{\left\|\frac{1}{n_s}X^TXr + \frac{1}{n_s}X^Te\right\|} && (52)\\
&= \frac{Vr}{\|Vr\|}. && (53)
\end{aligned}$$

In this limit, $w_1^{opt}(y)$ is independent of e, and $\tau_2^{opt}(y)$ has a non-central χ^2 distribution with $n_i - 1$ degrees of freedom. One can also show that all $w_i^{opt}(y)$ obey similar limits.

Thus $\tau_i^{opt}(y)$ has a non-central χ^2 distribution with n_p degrees of freedom in the limit of large n_s.

The above arguments have motivated using the independence assumption for computing the distribution of $\tau_{i+1}^{opt}(y)$ and revealed that this assumption will tend to hold when $w_i^{opt}(y)$ is " weakly affected by e." Stated differently, one wishes the non-centrality parameter for the distribution for $\tau_i^{opt}(y)$ to dominate the expected variance. Since a χ^2 random variable is involved, this variance is roughly the degrees of freedeom (the dimension of the search space) $n_p = n_i - i + 1$. Therefore as $r^T W_{i-1}^{\perp} \left({W_{i-1}^{\perp}}^T (X^T X)^{-1} W_{i-1}^{\perp} \right)^{-1} {W_{i-1}^{\perp}}^T r$ becomes comparable to n_p, the independence assumption will be incorrect for $w_j^{opt}(y)$ and $\tau_j^{opt}(y)$ when $j > i$. Consider briefly the extreme case $r = 0$, $X^T X = I$. Then $\tau_1^{opt}(y) \sim \chi^2_{n_i}$. The non-centrality parameter is zero, $w_i^{opt}(y)$ is totally determined by e, and the independence assumption completely breaks down for $w_2^{opt}(y)$: $\tau_2^{opt}(y) = 0$. This simplified example illustrates a larger point: as the independence assumption begins to break down, the earlier directions "steal" variance from later directions, and the correct distributions of $\tau_i^{opt}(y)$ for later directions will have smaller tails than the distributions computed using the independence assumption. When using the independence assumption with equation 43 to choose n_d, a test using the independence assumption will choose an n_d less than or equal to the n_d determined using that same test with the distribution that properly accounts for the dependence of $w_i^{opt}(y)$ on e.

As discussed, SR violates the independence assumption essential for the derivation of the results of section 2 (properties of restriction regressors). However, as shown above, the w_i^{opt} which SR prefers to include in W_{n_d} will tend to approximate the assumption that w_i^{opt} is independent of e. Thus section 2 is approximately valid for SR for W consisting of w_i^{opt} where the non-centrality parameter dominates the variance, allowing one to apply the results of section 2 to SR restriction regressors.

The significance test developed here should not be viewed as a replacement for cross-validation, but as a complement. Often the two approaches will give similar determinations of n_d. However, the cross-validation techniques and significance tests rest on different assumptions and have varying computational needs. Significance tests will tend to impose less computational burden, but cross-validation will tend to be less impacted if the data deviate from the noise assumptions. Moreover, other approaches for determining n_d can be developed from the viewpoint developed in this paper.

3.4 Some properties of SR

Because of the statistical basis for SR, one can investigate properties for this regression method beyond those discussed in section 2. The remainder of this subsection discusses further results specific to SR. Computing the expectation value of $W_i^{opt}(y)$ is involved. However, for the sake of determining the expected value of the search space, one can use the results in appendix C.2 and state

$$\begin{aligned} \mathcal{E}\left(\text{Range}(W_i)\right) &= \text{Span}\left(\mathcal{E}\left([w_1^{opt}|\ldots|w_i^{opt}]\right)\right) & (54)\\ &= \text{Span}\left(\mathcal{E}\left([X^TX\tilde{r}|\ldots|(X^TX)^i\tilde{r}]\right)\right) & (55)\\ &= \text{Span}\left([X^TXr|\ldots|(X^TX)^ir]\right). & (56)\end{aligned}$$

Thus, $W_i^{opt}(y)$ provides an unbiased estimate of the "true" search space. Alternatively, consider the behavior of $W_i^{opt}(y)$ as n_s is increased. Assume that the input data is persistently exciting, that is $\lim_{n_s\to\infty}\frac{1}{n_s}X^TX = V$ for some non-singular V. For any w,

$$\begin{aligned} \lim_{n_s\to\infty}\tau(w,y) &= \lim_{n_s\to\infty}\frac{(w^T\tilde{r})^2}{w^T(X^TX)^{-1}w\sigma_e^2} & (57)\\ &= \lim_{n_s\to\infty}\frac{n_s(w^T\tilde{r})^2}{w^T(\frac{1}{n_s}X^TX)^{-1}w\sigma_e^2} & (58)\\ &= \begin{cases}\infty & \text{if } w^Tr\neq 0\\ 0 & \text{if } w^Tr = 0\end{cases} & (59)\end{aligned}$$

When n_s is large enough, $\tau(w,y)$ will be large enough to overcome any given threshold for "significance" for all directions where $w^Tr \neq 0$. This means that if the criterion in equation 43 is used to determine n_d, then for n_s sufficiently large $r \in \text{Range}(W_{n_d})$ and $\tilde{b}$ is an unbiased estimator of r.

Beyond the above asymptotic result, one can make other statements about bias. Obviously, $\tilde{b}$ is an unbiased estimate of r whenever $\mathcal{H}_0^{2,n_d+1}$ is true. Moreover, SR strives to choose W_{n_d} so that $r \in \text{Range}(W_{n_d})$, so SR regressors will tend to have the advantages discussed in section 2 for restriction regressors when the bias is "small." In fact, empirical work such as [Negiz and Cinar, 1992] and [Mejdell, 1990] has shown that assuming the prediction bias is "small" can be a good assumption for SR, so SR may yield smaller prediction intervals than one would compute using classical methods, as discussed in section 2.

To see further benefits of using a restriction regressor directly derived from a statistical foundation, consider the heteroscedastic case, that is $\mathcal{E}\left(ee^T\right) = \sigma_e^2 P$. The SR method begins by computing the minimum-variance unbiased estimator and its variance. Thus, the additional error information is naturally incorporated into the procedure. PLS does *not* make use of this additional information, so PLS is not equivalent to SR in this case. If one draws draws an analogy to generalized least squares and rescales the data $X_{rescaled} = P^{-\frac{1}{2}}X$ and $y_{rescaled} = P^{-\frac{1}{2}}y$, then performing PLS on the scaled data is equivalent to SR. Thus PLS rests on the assumption of homoscedasticity.

Lastly, we touch upon the vital but difficult issue of scaling. Let z_i be the vector of inputs for the i-th input; that is, let $X = [z_1|\ldots|z_{n_i}]$. Then $w_1^{opt} = X^T y = [z_1^T y \; z_2^T y \ldots z_{n_i}^T y]^T$. The "most significant" vector is formed from the covariances between the individual inputs and the output. Herein one can see the effect of scaling; if one of the z_i is multiplied by a large constant, then that input will figure much more prominently in the "most significant" vector. Clearly, one would like to mollify such effects. A common method is autoscaling: subtracting the mean from each z_i and dividing the resulting values by the standard deviation of that z_i. Autoscaling clearly removes the scaling effect discussed here. Moreover, when autoscaling is used, the "most significant" vector is formed from the correlation coefficients between the individual inputs and the output. This observation provides a heuristic motivation for using autoscaling when the inputs are uncorrupted by measurement noise.

4 Multiple Output Problems

The development to this point has dealt with scalar output problems. This section generalizes the results of section 3 to problems with multiple outputs. Sections 4.1 and 4.2 work with vector output problems of the form

$$Y = XR + E, \tag{60}$$

where $Y \in \Re^{n_s \times n_o}$ is known, $R \in \Re^{n_i \times n_o}$ is an unknown regression matrix, and $E \in \Re^{n_s \times n_o}$ is an unobservable matrix of errors. For simplicity of development, further assume that the elements of E are zero-mean, independent, and homoscedastic random variables: $\mathcal{E}(E) = 0$, $\mathcal{E}\left(E^T E\right) = n_s \sigma_e^2 I$ and $\mathcal{E}\left(EE^T\right) = n_o \sigma_e^2 I$. The independence and homoscedasticity assumptions can be readily relaxed. The first subsection examines the generalization of SR for the situation where one is interested in incrementally building subspaces of the inputs and outputs that capture "useful" variations; such problems are sometimes referred to as "factor analysis" problems. PLS will be seen to be very similar to the SR method for "factor analysis." The second subsection develops the SR regressor for vector output problems. The third subsection discusses further generalization to tensor data problems.

For the scalar output case the estimate of the regression vector, $\tilde{b}$, was restricted to linear combinations of a set of mutually orthogonal vectors, namely the columns of W. For the vector generalization, we restrict the estimate of the regression matrix, $\tilde{B}$, to linear combinations of a set of mutually orthogonal matrices. This orthogonality is described via the natural inner product for spaces of matrices, the tensor inner product. This inner product is defined as $< A, B > = \mathrm{Tr}(AB^T)$ and is the inner product that defines the matrix Frobenius norm: $\sqrt{< A, A >} = ||A||_F$.

The results of section 2 generalize to the vector output case in a straight-forward manner. Thus one is interested in the null hypothesis

$$\mathcal{H}_0^3 : \qquad < R, S > = 0 \tag{61}$$

where $S \in \Re^{n_i \times n_0}$. For this null hypothesis, a natural test statistic is

$$\begin{aligned} \tau(S, Y) &= \frac{< \tilde{R}, S >^2}{\mathrm{Var} < \tilde{R}, S >} &(62)\\ &= \frac{\mathrm{Tr}(\tilde{R}S^T)}{\mathrm{Tr}\,(S^T(X^TX)^{-1}S)\,\sigma_e^2}. &(63)\end{aligned}$$

Analogous to the section 3, $\mathcal{H}_0^3$ and the resulting test statistic provide the basis for the multivariable SR methods.

4.1 Sequentially building "significant factors"

One common objective of multivariable analysis is to develop a lower dimensional description of data in which most of the "useful" information has been preserved. For example, Principal Components Analysis has long been used to identify a small set of "loading vectors" that encompass the greatest portion of the variance of a set of data. More recently, several practitioners have recommended PLS when one is interested in variances of input data that explain variances of dependent variables. [Martens and Næs, 1989, Geladi and Kowalski, 1986]. The approach has been found particularly effective for multivariable stochastic process control. [Kresta et al., 1991, Piovoso et al., 1992]

To evaluate vectors $w \in \Re^{n_i}$ and $c \in \Re^{n_0}$ using $\tau(S, Y)$, one parameterizes the matrix being evaluated to be rank one, that is $S = wc^T$. Equation 63 then becomes

$$\tau(w, c, y) = \frac{(w^T \tilde{R} c)^2}{w^T (X^T X)^{-1} w \; c^T c \; \sigma_e^2} \tag{64}$$

which, if normal errors are assumed, arises from a normal distribution for any given w and c. Next, one solves for the optimal w and c. The wc^T parameterization yields $\nabla_S(\cdot) = \begin{bmatrix} \nabla_w(\cdot) \\ \nabla_c(\cdot) \end{bmatrix}$. Solving

$$\nabla_w \; \tau(w, c, Y)|_{S^{opt}(Y)} = 0 \tag{65}$$

for $w^{opt}(y)$ one finds

$$w^{opt} = X^T X \tilde{R} c^{opt}(y) = X^T Y c^{opt}(y), \tag{66}$$

which corresponds to the formula for PLS (see appendix B, equation 93). Solving

$$\nabla_c \; \tau(w, c, Y)|_{S^{opt}(Y)} = 0 \tag{67}$$

and dropping the arguments results in

$$\tilde{R}^T w^{opt} = \frac{w^{opt^T} \tilde{R} c^{opt}}{c^{opt^T} c^{opt}} c^{opt}. \tag{68}$$

Substituting w^{opt} from equation 66 one finds that c^{opt} satisfies

$$\tilde{R}^T X^T X \tilde{R} c^{opt} = \sigma_e^2 \; \tau|_{S^{opt}(Y)} \, c^{opt}. \tag{69}$$

Thus finding the eigenvector of maximum eigenvalue in equation 69 yields the c^{opt} needed for equation 66. Since $\mathcal{E}\left(\tilde{R}^T X^T X \tilde{R}\right)$ and $\mathcal{E}\left(Y^T Y\right)$ have the same eigenvectors, one can see that in using the "most significant subspace" one is selecting the c that explains the greatest variance in the output data. This is in contrast to PLS, where c is chosen as the eigenvector of $Y^T X X^T Y$ with maximum eigenvalue.

To find additional "significant vectors," one repeats the process enforcing the orthogonality constraint $< S_i^{opt}(Y), S_j^{opt}(Y) > = 0$. Due to the wc^T parameterization, $< S_i^{opt}(Y), S_j^{opt}(Y) > = 0$ if $< w_i^{opt}(Y), w_j^{opt}(Y) > = 0$ or $< c_i^{opt}(Y), c_j^{opt}(Y) > = 0$. The method developed below relies on the two assumptions: $n_i \geq n_o$ and $< w_i^{opt}(Y), w_j^{opt}(Y) >$

$= 0\ \forall i \neq j$. An equivalent method for the case $n_i < n_o$ can be developed by invoking the alternate second assumption $< c_i^{opt}(Y), c_j^{opt}(Y) > \; = 0\ \forall i \neq j$.

Any matrix $W_{i-1} \in \Re^{n_i \times n_w}$ can be factored as $W_{i-1} = Q_{\rm QR} R_{\rm QR}$ where $Q_{\rm QR} \in \Re^{n_i \times n_i}$ is an orthogonal matrix and $R_{\rm QR} \in \Re^{n_i \times n_w}$ is an upper triangular matrix. A property of the QR decomposition is that if one partitions $Q_{\rm QR} = [Q_{\rm QR,1}, Q_{\rm QR,2}]$ such that $Q_{\rm QR,1}$ contains the first n_w columns of $Q_{\rm QR}$, then the columns of $Q_{\rm QR,2}$ describe the null space of W_{i-1}^T. Let the columns of W_{i-1} be $w_1^{opt}(Y)$ through $w_{i-1}^{opt}(Y)$ and $W_{i-1}^{\perp} = Q_{\rm QR,2}$. For this problem the necessary condition for a constrained extremum is

$$W_{i-1}^T \;\; \nabla_w \; \tau(w,c,Y)|_{S^{opt}(Y)} = 0, \tag{70}$$

which gives rise to the eigenvector equation

$$\sigma_e^2 \tau q = {W_{i-1}^{\perp}}^T X^T X \tilde{R} \tilde{R}^T W_{i-1}^{\perp} \; q \tag{71}$$

where $q \in \Re^{n_p}$ and the arguments have been suppressed. The q which is the eigenvector of maximum eigenvalue in equation 71 yields $w_i^{opt}(Y) = W_{i-1}^{\perp} q$. The full algorithm is

Algorithm 2 (Identifying SR factors)

$$\tilde{R} \quad = \quad (X^T X)^{-1} X^T Y \tag{72}$$

$$W_0 = \quad [0 \ \ldots \ 0]^T, \quad W_0 \in \Re^{n_i} \tag{73}$$

$$\text{DO } i \quad = \quad 1 \text{ to } n_d$$

$$\text{Perform} \quad QR \text{ factorization of } W_{i-1}$$

$$W_{i-1} \Rightarrow Q_{\rm QR}, \; R_{\rm QR} \tag{74}$$

$$W_{i-1}^{\perp} \quad = \text{last } n_p = n_i - i + 1 \text{ columns of } Q_{\rm QR} \tag{75}$$

$$q \quad = \text{eigenvector of maximum eigenvalue of}$$

$${W_{i-1}^{\perp}}^T X^T X \tilde{R} \tilde{R}^T W_{i-1}^{\perp} \tag{76}$$

$$w_i^{opt}(Y) \quad = W_{i-1}^{\perp} q \;/ ||W_{i-1}^{\perp} \; q|| \tag{77}$$

$$c_i^{opt}(Y) \quad = \tilde{R}^T w_i^{opt} \tag{78}$$

$$W_i \quad = [w_1^{opt} | w_2^{opt} | \cdots | w_i^{opt}] \tag{79}$$

$$\text{END} \quad \text{DO}$$

The only difference between this algorithm and PLS is in the specification of w_i^{opt}. SR chooses vectors that explain the greatest variance in $Y^T Y$. To see this, use the alternate but equivalent second assumption $< c_i^{opt}(Y), c_j^{opt}(Y) > = 0 \; \forall i \neq j$. Under this assumption, the condition for a constrained extremum gives rise to the eigenvector equation

$$\sigma_e^2 \tau q = {C_{i-1}^{\perp}}^T \tilde{R}^T X^T X \tilde{R} C_{i-1}^{\perp} \; q \tag{80}$$

where $C_{i-1}^{\perp}$ spans the space of allowable $c_i^{opt}(Y)$. Since

$$\mathcal{E}\left({C_{i-1}^{\perp}}^T \tilde{R}^T X^T X \tilde{R} C_{i-1}^{\perp}\right) = \mathcal{E}\left({C_{i-1}^{\perp}}^T Y^T Y C_{i-1}^{\perp}\right), \tag{81}$$

SR computes the vector in Range$(C_{i-1}^{\perp})$ that explains the greatest variance in $Y^T Y$. PLS chooses vectors using the slightly different criterion $Y^T X X^T Y = \tilde{R}^T (X^T X)^2 \tilde{R}$, so PLS can be viewed as a very close approximation to the SR method for building "significant factors."

4.2 Sequentially building "significant" regressors

We now turn to the problem of building the biased regressor $\tilde{B}$. One can develop SR directly from $\tau(S, Y)$ using tensor operations. However, one can also recast the vector output problem as a scalar output problem and use existing results. To use the scalar output results, the input data matrix and the output data matrix need to be suitably redefined. Normally,

$$Y = \begin{bmatrix} y_1^T \\ \vdots \\ y_{n_s}^T \end{bmatrix}, \qquad X = \begin{bmatrix} x_1^T \\ \vdots \\ x_{n_s}^T \end{bmatrix}, \qquad \text{and } Y = XR + E. \tag{82}$$

To conform to equation 1, Y, E, and R must be transformed into column vectors. Considering the columns of $R = [r_1| \; \cdots \; |r_{n_o}]$, let

$$y_{stacked} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_s} \end{bmatrix} \qquad \text{and } r_{stacked} = \begin{bmatrix} r_1 \\ \vdots \\ r_{n_o} \end{bmatrix} \tag{83}$$

where $y_{stacked} \in \Re^{n_s n_o}$ and $r_{stacked} \in \Re^{n_i n_o}$. Create $e_{stacked}$ from E is the same manner that $y_{stacked}$ was created from Y. Moreover, build $X_{stacked} \in \Re^{n_s n_o \times n_i n_o}$ such that

$$
X_{stacked} = \begin{array}{c} \overbrace{\hspace{8em}}^{n_i \times n_o} \\ \left[\begin{array}{c} n_o \left\{ \begin{array}{cccc} x_1^T & 0 & \dots & 0 \\ 0 & x_1^T & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & x_1^T \end{array} \right. \\ \begin{array}{cccc} x_2^T & 0 & \dots & 0 \\ 0 & x_2^T & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & x_2^T \end{array} \\ \vdots \\ \begin{array}{cccc} x_{n_s}^T & 0 & \dots & 0 \\ 0 & x_{n_s}^T & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & x_{n_s}^T \end{array} \end{array} \right] \end{array} . \tag{84}
$$

Then equation 60 data can be described by

$$
y_{stacked} = X_{stacked}\, r_{stacked} + e_{stacked}. \tag{85}
$$

Equation 85 is consistent with equation 1, so algorithm 1 can be used. After the SR process is completed, $\tilde{b}_{stacked}$ must be "unstacked" in the reverse manner from which $r_{stacked}$ was built in equation 83.

The difference between the SR method for regression and PLS can be easily seen by considering the S used in the two approaches. In PLS, S is constrained to only those matrices that can be described wc^T; thus, S is rank one and has only $n_o + n_i$ free parameters. However in SR regression, S is allowed to be any matrix, and thus has $n_o \times n_i$ free parameters. Generally, S_i^{opt} will be full rank (after all, $S_1^{opt} = \alpha X^T Y$). However to describe a full rank S from the PLS vectors, one must build at least n_o PLS directions, assuming $n_i \geq n_o$. Building a full-rank S_{PLS}^{opt} from the first n_o PLS loading vectors leads to $\tau(S_{PLS}^{opt}, Y) \leq (\tau(S_1^{opt}, Y)$; PLS is sub-optimal for determining the "most significant subspace." Another PLS variant, "one-at-a-time" PLS (OAT-PLS) is closely related to SR. In OAT-PLS one ignores the vector output problem and solves n_o "independent" single-output problems. However, OAT-PLS is also sub-optimal for determining the "most significant subspace." See [Holcomb and Morari, 1993] for further discussion of OAT-PLS and also the properties of PLS for vector output problems.

4.3 Tensor problems

Tensors are generalizations of matrices that allow one to work with a richer class of problems than is encompassed by multivariable regression. As improvements in computer technology have made larger and richer data structures more readily available to practitioners, more researchers have been pondering tensor data descriptions [Sanchez and Kowalski, 1986, Wold et al., 1987]; tensors are particularly interesting for problems where multiple sensors are being used [Carey et al., 1986].

For tensors an appropriate inner product can still be described by $< A, B > = \mathrm{Tr}(AB^T)$. That is, for any tensors $A, B \in \Re^{n_1 \times \ldots \times n_N}$, $< A, B > = \sum_{i_1,\ldots,i_M} a_{i_1,\ldots,i_M} b_{i_1,\ldots,i_M}$. If one desired to build "significant factors," then one would define a tensor S from the outer product of a vector from each of the component vector spaces and optimize equation $\tau(S, Y)$. Such an approach generates "power-law" algorithms familiar to PLS researchers. If one desired to build a tensor restriction regressor, then the tensors can be "stacked" and one can use algorithm 1 and the results of section 3.

5 Simulation Examples

The above development focused on theoretical understanding and derivation. This section presents numerical studies that clarify the relationship between PLS and SR and illustrate the utility of the new results for problems with multiple outputs. In this study, the examples are simulation studies using purely synthetic data. The data are not claimed to correspond to any particular "real world" process; rather, the data were generated to conform to the model assumptions and to illustrate the relative effectiveness of various methods for problems that satisfy the model assumptions. The "real world" successes of PLS (*e.g.* [Martens and Næs, 1989, Mejdell, 1990, Ricker, 1988]) are suggested as evidence of the practical utility of SR since the two methods are closely related.

The regression methods investigated were

- ordinary least squares (OLS),

- partial least squares using cross-validation (PLScv),
- significance regression using cross-validation (SRcv), and
- significance regression using equation 43 and the approximate distribution for $\tau_i^{opt}(y)$ defined in section 3.3 (SR).

All examples had ten inputs and four outputs ($n_i = 10$ and $n_o = 4$). For each case study, one thousand distinct examples were examined to mitigate sampling effects in the numerical results. Each example was generated by the method presented in appendix D. Since both input variances and the values of the regression parameters varied over five orders of magnitude and since there were typically large variances in the input data that had little effect on the output, this exploration shed light on the relative strengths and weaknesses of the four methods for a class of problems that has historically bedeviled OLS.

Two measures were employed to evaluate regressor performance. Since the examples were synthetic, R was known and a point estimate of the MSE could be computed for each example. The measure was

$$RMS_{\text{MSE}} = \sqrt{\frac{\text{Tr}(\ (\tilde{B} - R)(\tilde{B} - R)^T\ )}{\text{Tr}(RR^T)}}. \tag{86}$$

The $\text{Tr}(RR^T)$ term was included to produce a relative error and allow averaging over all one thousand examples.

The second measure was computed based on the PRESS. For each example an additional one hundred samples $(X_{new},\ Y_{new})$ were generated from the identical distribution as the training data, but the Y_{new} were not corrupted by error ($E_{new} = 0$). Then

$$RMS_{\text{PRESS}} = \sqrt{\frac{\text{Tr}(\ (X_{new}\tilde{B} - Y_{new})^T(X_{new}\tilde{B} - Y_{new})\ )}{400}}. \tag{87}$$

Since the data were generated with the constraint

$$\sqrt{\frac{\text{Tr}(Y_{new}^T Y_{new})}{400}} = 1 \tag{88}$$

the RMS_{PRESS} was averaged over the examples without normalization. Note that $n_s \times n_o = 400$ for the test set. Also, for each example, the rank (relative performance) of each estimator was recorded: rank = 1 if no other regressor did better for that example, rank = 2 if one other regressor did better, and rank = 3 if two other regressors did better.

method	$\overline{RMS_{\mathrm{MSE}}}$	$\overline{\mathrm{rank}}$	$\overline{RMS_{\mathrm{PRESS}}}$	$\overline{\mathrm{rank}}$
OLS	860	3.0	0.36	3.0
PLScv	8.0	1.4	0.19	1.4
SRcv	1.1	1.3	0.19	1.4

Table 1: Comparison of PLS and SR using cross-validation over $1,000$ examples of synthetic data.

method	$\overline{RMS_{\mathrm{MSE}}}$	$\overline{\mathrm{rank}}$	$\overline{RMS_{\mathrm{PRESS}}}$	$\overline{\mathrm{rank}}$
OLS	860	3.0	0.36	2.8
SR	0.9	1.5	0.27	1.7
SRcv	1.1	1.2	0.19	1.2

Table 2: Comparison of the approximate significance test using a 90 % threshold and cross-validation over $1,000$ examples of synthetic data.

The average rank with respect to both MSE and PRESS was computed. In both cases, thirty samples were available for training ($n_s = 30$). Where cross-validation was used to determine n_d, ten-way (three-out) cross-validation was employed.

The first case compared SRcv to PLScv. The results are shown in Table 1. Since PLScv and SRcv are similar, one should not be surprised that the two methods had similar results and outperformed OLS in all measures. As discussed in subsection 4.2, PLS is *not* optimal for determining the "most significant subspace" of $\Re^{n_i \times n_o}$. This is reflected in the MSE results; the $\overline{RMS_{\mathrm{MSE}}}$ for PLS was almost eight times that of the $\overline{RMS_{\mathrm{MSE}}}$ for SRcv. This difference did not appear to be crucial for prediction; SRcv and PLScv produced almost identical results as measured by $\overline{RMS_{\mathrm{PRESS}}}$.

The second test compared cross-validation to using equation 43 with the approximate distribution (F- distribution) developed in section 3.3 operating with a 90% significance criterion. These result are shown in Table 2. The two methods used the identical algo-

method	$\overline{RMS_{\text{MSE}}}$	$\overline{\text{rank}}$	$\overline{RMS_{\text{PRESS}}}$	$\overline{\text{rank}}$
OLS	860	3.0	0.36	2.8
SR	0.9	1.4	0.25	1.7
SRcv	1.1	1.2	0.19	1.2

Table 3: Comparison of the "over-simplified" significance test and cross-validation over $1,000$ examples of synthetic data.

rithm to compute $\tilde{B}$. SR and SRcv have similar $\overline{RMS_{\text{MSE}}}$, even though SRcv used ten times more computations than SR. Thus, in terms of $\overline{RMS_{\text{MSE}}}$ the approximate significance test performed almost as well as cross-validation and was much less computationally demanding. Interestingly, the $\overline{RMS_{\text{MSE}}}$ for SR reported in Table 2 was less than one-eight of the $\overline{RMS_{\text{MSE}}}$ for PLScv reported in Table 1. These numbers are directly comparable since both were generated using the same one-thousand synthetic examples.

In the above example the 90% significance criterion was chosen arbitrarily. Next, the sensitivity of the results to this threshold was investigated. In this simulation an "over-simplified" approximate significance test was employed: reject $\mathcal{H}_0^{2,i}$ if $\tau_i^{opt}(y) > n_p + 1$. For large n_o and n_s this is a crude approximation to the 50% significance threshold. Table 3 shows the results; the "over-simplified" method had similar results to the approximate test with a 90% significance test. For these simulations the results were relatively insensitive to the choice of threshold. In terms of the $\overline{RMS_{\text{PRESS}}}$, cross-validation was clearly superior to the approximate significance test.

These numerical explorations illustrated several points. For the purpose of prediction partial least squares is virtually identical to the significance regression. However, SR was clearly superior for estimation in these problems. Less computationally demanding alternatives to cross-validation can be developed from the classical viewpoint of significance, but more work is needed on these significance tests. In particular, the relationship between desired objective (*e.g.* $\overline{RMS_{\text{MSE}}}$ or $\overline{RMS_{\text{PRESS}}}$) and choice of significance test needs further work. Still, even the current SR approach using approximate significance test outperformed PLS for estimation while using only one-tenth the computational effort.

6 Conclusion

This work developed a novel biased regression technique founded upon the concept of statistical significance. The technique was applied to various linear regression models. For the case of vector inputs and scalar outputs, the method is equivalent to partial least squares (PLS). The study of the heteroscedastic model illustrated a key assumption underlying the PLS: PLS assumes independent homoscedastic errors. For the problem of vector inputs and vector outputs, two different problems were examined. First a significance regression method was developed for "factor analysis," that is to identify "significant subspaces" of the input and output spaces. PLS was almost identical to this method. Next, the significance regression method for multivariable regression was developed; this method was superior to PLS for both maximizing the "significance" objective function, $\tau(S,Y)$, and for reducing the MSE of estimation in the simulation example. Generalizations for tensor data were also briefly discussed.

Significance regression has several advantages. Unlike many of the more common PLS formulations, all of the assumptions are stated at the beginning of the process and the procedure follows algorithmically from these assumptions. If one changes the assumptions, the implications for the regression process are clear. Clearly this formulation allows one to examine the appropriateness of the assumptions being used for the problem at hand.

Additionally, the significance concept underlying the new regression technique can be translated directly into a null hypothesis to drive significance tests; this approach represents an alternative to cross-validation for choosing the number of "significant subspaces." In a simulation example, an approximate significance test required an order of magnitude less computational effort than cross-validation but yielded better performance than PLS as measured by the MSE. More work is needed to improve and understand these significance tests.

Acknowledgments: *Tyler Holcomb is a recipient of a National Science Foundation Graduate Fellowship. This research was supported by the Caltech Consortium in Chemistry and Chemical Engineering. Founding members of the Consortium are E. I. du Pont de Nemours and Company, inc., Eastman Kodak Company, Minnesota Mining and Manufac-*

turing Company, and Shell Oil Company Foundation. Håkan Hjalmarsson was partially supported by the Swedish Institute and the Blanceflor Boncompagni-Ludovisi Foundation during this work.

References

[Bibby and Toutenburg, 1977] Bibby, J. and Toutenburg, H. (1977). *Prediction and Improved Estimation in Linear Estimators.* Wiley.

[Carey et al., 1986] Carey, W. P., Beebe, K. R., Sanchez, E., and Kowalski, B. (1986). Chemometric analysis of multisensor arrays. *Sensors and Actuators*, 9:223–234.

[Draper and Smith, 1966] Draper, N. R. and Smith, H. (1966). *Applied Regression Analysis.* Wiley.

[Frank and Friedman, 1992] Frank, I. E. and Friedman, J. H. (1992). A statistical review of some chemometrics regression tools. Technical report, Dept. of Statistics, Stanford University, Stanford, CA 94305.

[Geladi and Kowalski, 1986] Geladi, P. and Kowalski, B. (1986). Partial least squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17.

[Gruber, 1990] Gruber, M. H. (1990). *Regression Estimators.* Academic Press.

[Helland, 1988] Helland, I. S. (1988). On the structure of partial least squares. *Communications in Statistics - Simulation*, 17(2):1581–607.

[Helland, 1990] Helland, I. S. (1990). Partial least squares and statistical models. *Scandinavian Journal of Statistics*, 17:97–114.

[Helland, 1992] Helland, I. S. (1992). Maximum likelihood regression on relevant components. *Journal of the Royal Statistial Society, Series B: Methodological*, 54(2):637–647.

[Holcomb and Morari, 1993] Holcomb, T. R. and Morari, M. (1993). Pls leads to different algorithms for factor analysis and regression. CDS Technical Memo CIT-CDS 93-003, California Institute of Technology, Pasadena, CA 91125.

[Horn and Johnson, 1985] Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge.

[Höskuldsson, 1988] Höskuldsson, A. (1988). Pls regression methods. *Journal of Chemometrics*, 2:211–228.

[Joliffe, 1982] Joliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, 31:300–303.

[Kresta et al., 1991] Kresta, J., MacGregor, J., and Marlin, T. E. (1991). Multivariate statistical monitoring of process operating performance. *Canadian Journal of Chemical Engineering*, 69(1):35–47.

[Ljung, 1987] Ljung, L. (1987). *System Identification: Theory for the User*. Prentice Hall.

[Lorber et al., 1987] Lorber, A., Wangen, L., and Kowalski, B. (1987). A theoretical foundation for the pls algorithm. *Journal of Chemometrics*, 5:19–31.

[Martens and Næs, 1989] Martens, H. and Næs, T. (1989). *Multivariate Calibration*. Wiley.

[Massy, 1965] Massy, W. F. (1965). Principal components regression in exploratory statistical research. *American Statistical Association Journal*, pages 234–256.

[Mejdell, 1990] Mejdell, T. (1990). *Estimators for Product Composition in Distillation Columns*. PhD thesis, University of Trondheim, The Norwegian Institute of Technology.

[Moler et al., 1990] Moler, C., Little, J., Bangert, S., and Kleinman, S. (1990). *MATLAB User's Guide*. The MathWorks.

[Negiz and Cinar, 1992] Negiz, A. and Cinar, A. (1992). On the detection of multiple sensor abnormalities in multivariate processes. In *Proceedings of the 1992 Automatic Control Conference*, pages 2364–2368.

[Piovoso et al., 1992] Piovoso, M., Kosanovich, K., and Pearson, R. K. (1992). Monitoring process performance in real-time. In *Proceedings of the 1992 Automatic Control Conference*, pages 2359–2364.

[Ricker, 1988] Ricker, N. L. (1988). The use of biased least-squares estimators for parameters in discrete-time pulse response models. *Industrial and Engineering Chemical Research*, 27:343–350.

[Sanchez and Kowalski, 1986] Sanchez, E. and Kowalski, B. (1986). Generalized rank annihilation factor analysis. *Analytical Chemistry*, 58(2):496–499.

[Stone and Brooks, 1990] 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. *Journal of the Royal Statistical Society*, B.52:237–269.

[Theobald, 1974] Theobald, C. M. (1974). Generalizations of mean squared error applied to ridge regression. *Journal of the Royal Statistical Society*, B.36:103–106.

[Wahlberg and Ljung, 1992] Wahlberg, B. and Ljung, L. (1992). Hard frequency-domain model error bounds from least-squares like identification techniques. *IEEE Transactions on Automatic Control*, 37(7).

[Wold, 1978] Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405.

[Wold et al., 1987] Wold, S., Geladi, P., Esbensen, K., and Öhman, J. (1987). Multi-way principal components- and pls-analysis. *Chemometrics and Intelligent Laboratory Systems*, 1:41–56.

[Wold et al., 1984] Wold, S., Ruhe, A., Wold, H., and Dunn, W. (1984). The collinearity problem in linear regression: The partial least squares approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, 5(3):753–743.

A Nomenclature

In general, capital letters represent matrices, lower case letters represent column vectors, and Greek letters represent scalars. Estimates are denoted by a tilde, "˜". The dimensions of matrices are denote by subscripted n's.

some scalars, vectors, and matrices

variable	dimension	description
$\tilde{b}$	$n_i \times 1$	is the biased estimate of r. See equation 4.
$\tilde{B}$	$n_i \times n_o$	is the biased estimate of R.
I	as appropriate	is the identity matrix.
q_1	$n_w \times 1$	is the component of r in the search space. $q_1 = W^T r$.
q_2	$n_i - n_w \times 1$	is the component of r orthogonal to the search space. $q_2 = W^{\perp T} r$.
r	$n_i \times 1$	is the "true" regression vector. See equation 1.
$\tilde{r}$	$n_i \times 1$	is the minimum variance unbiased estimate of r.
R	$n_i \times n_o$	is the "true" regression matrix. See equation 60.
S	$n_i \times n_o$	is the matrix representing the subspace under consideration. See equation 63
W	$n_i \times n_w$	is the matrix whose range defines the search space for $\tilde{b}$. See equation 4.
v	varies$\times$1	is a vector locally defined. Any given v may or may not relate to any other v.
x_j	$n_i \times 1$	is the jth input data sample.
X	$n_s \times n_i$	input data; each row corresponds to one input sample. Thus, $X^T = [x_1 \;\; x_2 \;\; \ldots \;\; x_{n_s}]$.
y_i	$n_o \times 1$	is the ith output data sample.
Y	$n_s \times n_o$	is the output data. Each row corresponds to one output sample. Thus, $Y^T = [y_1 \;\; y_2 \;\; \ldots \;\; y_{n_s}]$.
$\tilde{Y}$	$n_s \times n_o$	is the regression prediction of output data.
$\tau(S, Y)$	scalar	is the test statistic for S and a given Y.
$\tau_i^{opt}(Y)$	scalar	is the maximum of $\tau(S, Y)$ over all S in the allowable space.

dimensional descriptors

variable	description
n_d	is the number of "significant subspaces" to be generated.
n_i	is the number of inputs.
n_o	is the number of outputs.
n_s	is the number of samples.
n_p	is dimension of the allowable space in which to search for further w_i^{opt}. For scalar output problems, $n_p = n_i - i + 1$.
n_w	is the rank of W.

operators

operator	description
$\lvert \cdot \rvert$	is the absolute value.
$\lvert\lvert \cdot \rvert\rvert$	is the Euclidean norm. $\lvert\lvert a \rvert\rvert = \sqrt{< a, a >}$.
$\lvert\lvert \cdot \rvert\rvert_F$	is the Frobenius (matrix Euclidean) norm. $\lvert\lvert a \rvert\rvert_F = \sqrt{< A, A >} = \sqrt{\sum_{i,j} a_{i,j}^2}$, where $a_{i,j}$ are the components of A.
$[W \; \lvert V]$	is the matrix formed by placing W and V side-by-side.
$< \cdot, \cdot >$	is the inner product. For matrices A and B, $< A, B > = \mathrm{Tr}(AB^T)$.
$\mathcal{E}\,(\cdot)$	is the expectation.
$\mathrm{MSE}(\cdot)$	is the Mean Square Error as $n_s \to \infty$. See equation 5.
$\mathrm{PRESS}(\cdot)$	is the PRedicted Error Sum of Squares. See equations 6 and 7.
$\mathrm{Pr}\{event\}$	is the probability of *event* occurring.
$\mathrm{Range}(\cdot)$	is the range; the span of the column vectors of a matrix.
$\mathrm{Rank}(\cdot)$	is the dimension of the range of an operator.
$\mathrm{Span}(\cdot)$	is the space defined by all linear combinations of the elements in a set.
$\mathrm{Tr}(\cdot)$	is the trace, the sum of the diagonal elements of a matrix.
$\mathrm{Var}(\cdot)$	is the variance.

B PLS Review

This section restates known results for partial least squares in a manner more compatible with the results developed in this paper. The algorithm was primarily developed by Wold [Wold et al., 1984]. A tutorial was provided by Geladi and Kowalski, [Geladi and Kowalski, 1986] while Höskuldsson [Höskuldsson, 1988] analyzed the mathematical aspects of the algorithm. Helland [Helland, 1988] illustrated important properties of the loading vectors (w_i's). The view of PLS developed below draws strongly from Höskuldson's and Helland's work.

As described by Höskuldsson, the PLS vectors at a given step of the process are found by computing the eigenvector of maximum eigenvalue in the following equations:

$$X^T Y Y^T X w^{opt} = \lambda_w w^{opt} \tag{89}$$

$$Y^T X X^T Y c^{opt} = \lambda_c c^{opt}. \tag{90}$$

Another way to describe w^{opt} and c^{opt} is

$$\{w^{opt}, c^{opt}\} = \arg \max_{||w||=1,\ ||c||=1} < Xw, Yc > \tag{91}$$

which is equivalent to

$$\{w^{opt}, c^{opt}\} = \arg \max_{w,c} \frac{w^T X^T Y c}{\sqrt{w^T w c^T c}}. \tag{92}$$

Taking the gradient of equation 92 with respect to w and equating the result to zero yields

$$w^{opt} = \gamma X^T Y c^{opt} = \gamma X^T X \tilde{R} c^{opt}. \tag{93}$$

where γ is an irrelevant (but known) scalar. Thus, to compute the first PLS vectors, one solves the eigenvector equation 90 to determine c^{opt} and then uses 93 to produce w^{opt}. For the case of scalar output, c is a scalar so $w^{opt} = X^T y = X^T X \tilde{r}$.

Once the first PLS vector has been determined, w^{opt} is orthgononally removed from the input data, and the effect of w^{opt} and c^{opt} is subtracted from the output. The algorithm can thus be described:

$X_1 \quad = X \qquad (94)$

$Y_1 \quad = Y \qquad (95)$

Do $\quad i \quad = 1$ to n_d

Let c_i^{opt} be the eigenvector of maximum eigenvalue in

$$\lambda_{c_i} c_i = Y_i^T X_i^T X_i Y_i c_i \tag{96}$$

$$c_i^{opt} = \frac{c_i^{opt}}{||c_i^{opt}||} \tag{97}$$

$$w_i^{opt} = \frac{X_i^T Y_i c_i^{opt}}{||X_i^T Y_i c_i^{opt}||} \tag{98}$$

$$\beta_i = \frac{w_i^{opt\,T} X_i^T Y_i c_i^{opt}}{w_i^{opt\,T} X_i^T X_i w_i^{opt}} \tag{99}$$

$$\tilde{Y}_i = \beta_i X_i w_i^{opt} c_i^{opt\,T} \tag{100}$$

$$X_{i+1} = X_i \left(I - \frac{w_i^{opt} w_i^{opt\,T} X_i^T X_i}{w_i^{opt\,T} X_i^T X_i w_i^{opt}}\right) \tag{101}$$

$$Y_{i+1} = Y_i - \tilde{Y}_i \tag{102}$$

End Do

In PLS, one wishes to restrict the regressor to the input space and output space described by the PLS "loading vectors," w_i^{opt} and c_i^{opt}. Let $W = [w_1^{opt}|\ldots|w_{n_d}^{opt}]$ and let the orthonormal columns of C be such that $\text{Range}(C) = \text{Span}(\{c_1^{opt}, c_2^{opt}, \ldots, c_{n_d}^{opt}\})$. Moreover, define $C^\perp$ such that $[C|C^\perp]^T\ [C|C^\perp] = I$ and $n_c = \text{Rank}(C)$. The parameterization of all such allowable restriction regressors is $\tilde{B} = WVC^T$ where $V \in \Re^{n_d \times n_c}$. One solves for the least squared-error regressor in the classical manner.

$$\begin{aligned}
V^{opt} &= \arg \min_{V \in \Re^{n_d \times n_c}} ||Y - XWVC^T||_F^2 && (103)\\
&= \arg \min_{V \in \Re^{n_w \times n_c}} ||\ Y[C \mid C^\perp] - XWVC^T[C \mid C^\perp]\ ||_F^2 && (104)\\
&= \arg \min_{V \in \Re^{n_w \times n_c}} ||\ [YC \mid YC^\perp] - [XWV \mid 0]\ ||_F^2 && (105)\\
&= (W^T X^T X W)^{-1} W^T X^T Y C, && (106)
\end{aligned}$$

from which the PLS restriction regressor is $\tilde{B} = W(W^T X^T X W)^{-1} W^T X^T Y C C^T$.

C Proofs

C.1 Reduced variance of "restriction regressors"

Section 2 claims $\text{Var}(WW^T\tilde{r}) \geq \text{Var}(\tilde{b})$. The proof below is based directly on lemma 3.1 of [Wahlberg and Ljung, 1992].

Theorem 1 *For the model in equation 1 and any given W such that $W^TW = I$, $\text{Var}(WW^T\tilde{r}) - \text{Var}(\tilde{b})$ is always a positive semi-definite symmetric matrix.*

Proof. First, recall

$$\text{Var}(WW^T\tilde{r}) = WW^T(X^TX)^{-1}WW^T\sigma_e^2 \ \text{ and} \tag{107}$$

$$\mathrm{Var}(\tilde{b}) \quad = \quad W(W^T X^T X W)^{-1} W^T \sigma_e^2. \tag{108}$$

Let Range(V) equal the null space of $W^T X^T X$. Then $[W \mid V]$ is full rank and $W^T X^T X V = 0$. Next,

$$\begin{aligned}
(X^T X)^{-1} &= [W \mid V] \left([W \mid V]^T X^T X [W \mid V] \right)^{-1} [W \mid V]^T && (109)\\
&= [W \mid V] \begin{bmatrix} (W^T X^T X W)^{-1} & 0 \\ 0 & (V^T X^T X V)^{-1} \end{bmatrix} [W \mid V]^T && (110)\\
&= W(W^T X^T X W)^{-1} W^T + V(V^T X^T X V)^{-1} V^T && (111)
\end{aligned}$$

Pre- and post-multiplying equation 111 by $WW^T \sigma_e$ yields

$$\begin{aligned}
WW^T (X^T X)^{-1} WW^T \sigma_e^2 &= W(W^T X^T X W)^{-1} W^T \sigma_e^2 && (112)\\
&+ WW^T V(V^T X^T X V)^{-1} V^T WW^T \sigma_e^2
\end{aligned}$$

which becomes

$$\mathrm{Var}(WW^T \tilde{r}) - \mathrm{Var}(\tilde{b}) = WW^T V(V^T X^T X V)^{-1} V^T WW^T \sigma_e^2. \tag{113}$$

Noting that $WW^T V(V^T X^T X V)^{-1} V^T WW^T$ is symmetric and positive semidefinite completes the proof. □

Since $\mathrm{Var}(WW^T \tilde{r})$ dominates $\mathrm{Var}(\tilde{b})$ by a positive semi-definite matrix, one may invoke [Horn and Johnson, 1985][page 471] and further state:

$$\begin{aligned}
\mathrm{Tr}\left(\mathrm{Var}(WW^T \tilde{r})\right) &\geq \mathrm{Tr}\left(\mathrm{Var}(\tilde{b})\right) && (114)\\
\|\mathrm{Var}(WW^T \tilde{r})\|_2 &\geq \|\mathrm{Var}(\tilde{b})\|_2, \text{ and} && (115)\\
\|\mathrm{Var}(WW^T \tilde{r})\|_F^2 &\geq \|\mathrm{Var}(\tilde{b})\|_F^2. && (116)
\end{aligned}$$

C.2 Equivalence of SR and PLS for scalar output case

This appendix links the significance regression method (SR) to PLS for scalar output problems. In particular, a proof is developed that shows Helland's formula for the PLS loading vectors satisfies the necessary condition for the significant subspaces for scalar output models. A "significant vector" is understood to be any of the n_A vectors w_i^{opt} that satisfies equation 34.

Applying the necessary condition of equation 34 to the gradient of $\tau(w, y)$, described in equation 31, yields the condition

$$(I - W_{i-1}^{sig} {W_{i-1}^{sig}}^T)(\tilde{r} - \frac{\sigma_e^2 \tau(w_i^{opt}, y)}{\tilde{r}^T w_i^{opt}} (X^T X)^{-1} w_i^{opt}) = 0. \tag{117}$$

that must be satisfied in turn by each significant vector $w_i^{opt}(y)$.

Next consider Helland's method [Helland, 1988] for computing PLS loading vectors for model 2.

Algorithm 3 (Generation of PLS loading vectors)

$$\tilde{r} = (X^T X)^{-1} X^T y \tag{118}$$

$$W_0^{pls} = [0|\cdots|0]^T, \quad W_0^{pls} \in \Re^{n_i} \tag{119}$$

DO $i = 1, n_A$

$$v = (I - W_{i-1}^{pls} {W_{i-1}^{pls}}^T)(X^T X)^i \tilde{r} \tag{120}$$

$$w_i^{pls} = \frac{v}{||v||} \tag{121}$$

$$W_i^{pls} = [w_1^{pls}|w_2^{pls}|\cdots|w_i^{pls}] \tag{122}$$

END DO.

Now we can state the theorems linking PLS and SR. Theorem 2 shows that the PLS vectors are also "significant vectors" when they exist. Additional discussion shows that there are almost surely as many "significant vectors" as there are inputs.

Theorem 2 *Any vector satisfying the necessary condition for the ith significant vector is a non-zero scalar multiple of* w_i^{pls}.

Proof.

This theorem is proven inductively for the $n_A (n_d \leq n_A \leq n_i)$ significant vectors that exist.

For $i = 1$: Substituting the first PLS loading vector $\frac{(X^T X)\tilde{r}}{||(X^T X)\tilde{r}||}$ into expression 31 yields

$$\nabla_w \tau(w, y) = \frac{2\left(\tilde{r}\widetilde{(r^T(X^T X)\tilde{r})r}^T (X^T X)(X^T X)^{-1} - (\tilde{r}^T (X^T X)\tilde{r})^2 (X^T X)^{-1}\right)(X^T X)\tilde{r}\sigma_e^2}{(\tilde{r}^T (X^T X)(X^T X)^{-1}(X^T X)\tilde{r}\sigma_e^2)^2} \tag{123}$$

$$= \frac{2\tilde{r}(\tilde{r}^T (X^T X)\tilde{r})^2 \sigma_e^2 - 2\tilde{r}(\tilde{r}^T (X^T X)\tilde{r})^2 \sigma_e^2}{(\tilde{r}^T (X^T X)\tilde{r}\sigma_e^2)^2} \tag{124}$$

$$= 0. \tag{125}$$

Thus, any vector satisfying the necessary condition for the first significant vector is a non-zero scalar multiple of the first PLS loading vector.

Assume true for $i-1$**:** ¿From algorithm 3, the first column of W_{i-1}^{pls} is known to be $\frac{X^TX\tilde{r}}{||X^TX\tilde{r}||}$. Moreover, the jth column is $\sum_{k=1}^{j}\alpha_{k,j}(X^TX)^k\tilde{r}$. Since $(X^TX)^k\tilde{r}$ is linearly independent of $(X^TX)^j\tilde{r}\ \forall j\neq k,\quad j,k<i$ (due to the existence assumption) and $<(X^TX)^k\tilde{r},(X^TX)^j\tilde{r}>\neq 0 \quad \forall j$ and k, $\alpha_{k,j}\neq 0\ \forall\ k\leq j\ \ j,k<i$. The jth column of W_{i-1}^{sig} is also $\sum_{k=1}^{j}\alpha_{k,j}(X^TX)^k\tilde{r}$ because $W_{i-1}^{sig}=W_{i-1}^{pls}$ by assumption.

For $i,\ i\leq n_A$**:** From the i-1 step, $W_{i-1}=W_{i-1}^{sig}=W_{i-1}^{pls}$. The necessary condition for the ith significant vector, Equation 117, becomes

$$(I-W_{i-1}W_{i-1}^T)\tilde{r}=(I-W_{i-1}W_{i-1}^T)\frac{\sigma_e^2\tau_{w_i^{opt}}}{\tilde{r}^Tw_i^{opt}}(X^TX)^{-1}w_i^{opt}. \tag{126}$$

Notice that the left-hand side (LHS) of equation 126 can not be zero. If it were, then the ith significant vector could not exist.

Now describe w_i^{opt} as the sum of w_i^{pls} and some non-zero vector $v\in$ Range$(I-W_{i-1}W_{i-1}^T)$. This v is distinct from the v in equation 120. Then $w_i^{opt}=w_i^{pls}+v$, and equation 126 becomes

$$(I-W_{i-1}W_{i-1}^T)\tilde{r}=(I-W_{i-1}W_{i-1}^T)\frac{\sigma_e^2\tau_{w_i^{opt}}}{\tilde{r}^Tw_i^{opt}}(X^TX)^{-1}\left(w_i^{pls}+v\right). \tag{127}$$

By extension of the argument from the $i-1$ step, $w_i^{pls}=\sum_{k=1}^{i}\alpha_{k,i}(X^TX)^k\tilde{r}$, where $\alpha_{k,i}$ has the same definition as above. Multiplying w_i^{pls} by $(X^TX)^{-1}$ produces $\sum_{k=1}^{i}\alpha_{k,i}(X^TX)^{k-1}\tilde{r}$ which, after multiplying through the projection matrix $(I-W_{i-1}W_{i-1}^T)$ yields $\alpha_{1,i}(I-W_{i-1}W_{i-1}^T)\tilde{r}$. Thus equation 127 becomes

$$(I-W_{i-1}W_{i-1}^T)\tilde{r}=\frac{\sigma_e^2\tau_{w_i^{opt}}}{\tilde{r}^Tw_i^{opt}}\left(\alpha_{1,i}(I-W_{i-1}W_{i-1}^T)\tilde{r}\ +(I-W_{i-1}W_{i-1}^T)(X^TX)^{-1}v\right). \tag{128}$$

Consider now the second term of the RHS. If $\{X^TX\tilde{r},\ldots,(X^TX)^{n_i}\tilde{r}\}$ spans $\Re^{n_i}$, then one can quickly see that $(I-W_{i-1}W_{i-1}^T)(X^TX)^{-1}v$ is always non-zero. Additionally, one can show that if $\{X^TX\tilde{r},\ldots,(X^TX)^{n_i}\tilde{r}\}$ does not span $\Re^{n_i}$, then the additional basis vectors needed to span $\Re^{n_i}$ are eigenvectors of X^TX orthogonal to the columns of W_{i-1}. Thus, $(I-W_{i-1}W_{i-1}^T)(X^TX)^{-1}v\neq 0\ \forall v\in$ Range$(I-W_{i-1}W_{i-1}^T)$.

The LHS of equation 128 and the first term of the RHS are both vectors pointing in the same direction, namely $(I - W_{i-1}W_{i-1}^T)\tilde{r}$, so a vector $v \in \text{Range}(I - W_{i-1}W_{i-1}^T)$ satisfying $v \neq \eta w_i^{pls}$ for any scalar $\eta \neq 0$ would make equation 126 insoluble. However w_i^{opt} exists by assumption so equation 126 must have a well-defined solution and v therefore equals ηw_i^{pls} for some scalar η. Since

- a w_i^{opt} exists that satisfies equation 126,
- w_i^{opt} must be a vector pointing in the same direction as w_i^{pls}, and
- equation 126 is invariant to the length of w_i^{opt},

$w_i^{opt} = \eta w_i^{pls}$ for any $\eta \neq 0$ must satisfy equation 126. $\square$

The PLS loading vectors satisfy the necessary condition for the significant vectors for any $n_d \leq n_A$. Since both PLS and SR compute $\tilde{b}$ using equation 4 and the same search space, $\text{Range}(W_{n_d})$, the two methods yield the same $\tilde{b}$. Thus PLS is a useful algorithm for computing the SR search space. However, the above proof raises the question: what is the value of n_A? Drawing directly from Helland's results we know that n_A (which Helland calls M) is equal to the minimum number of right singular vectors of X (principal component vectors of X) required to form a basis for $\tilde{r}$. See theorem 1 and theorem 2 of [Helland, 1990]. Since $n_A < n_i$ only if $\tilde{r}$ is orthogonal to one of the right singular vectors of X, $n_A = n_i$ almost surely in practice.

D Generation of Data for Simulation Examples

The simulation exploration was conducted using Matlab [Moler et al., 1990]. The two Matlab M-files used to generate the data are described below. The parameters used with these routines were: `n_train = 30`, `n_test = 100`, `d = 10`, `o = 4`, `d_ind = 3`, `max_exp = 5`, `min_exp = 0`, and `noise = 0.5`.

The generation routine is specifically designed to produce difficult examples. The "true" regression vectors (columns of R) are drawn from a spherically symmetric distribution about the origin (all directions are equally probable). However, the length of these vectors varies over 5 orders of magnitude. Thus, from a Bayesian viewpoint, the prior distribution for the regression vector is not particularly informative. The X are chosen independently of the R and the singular values (the square roots of the eigenvalues of X^TX) also vary over 5 orders of magnitude. Thus, there will be large variances in the X data which do not lie in any of the directions of the columns of R and therefore have little effect on the output. This will trouble principal component regression methods that proceed by examining directions in the order of the value of their singular values (principal components). Lastly, three of the input variables vary independently of all other input variables, but the remaining seven are correlated. This covariance structure can cause difficulties for both variable subset selection methods such as step-wise regression [Frank and Friedman, 1992] and for scaling methods such as auto-scaling (using "standardized variables") that weight the explanatory data solely on the variance of each individual explanatory variable.

D.1 Routine to generate random regression problems

```
function [X,y,Xt,yt,b] =gen_dat2(n_train,n_test,d,o,d_ind
                                   ,max_exp,min_exp,noise)

% this function generates data for linear regression problems
%
%
%  n_train  is the number of samples to be the training set
%  n_test   is the number of samples to be the testing set
%  d        is the number of inputs
%  o        is the number of outputs
%  d_ind    is the number of inputs NOT rotated
%            and thus "independent"
%  max_exp  the largest order of magnitude contemplated
%  min_exp  the smallest order of magnitude contemplated
%            used for scaling the input data and
%            generating the regression vector
%  noise    std deviation of the normal additive noise
%
```

```
%
%  X          is the input training data
%  Xt         is the input testing data
%  y          is the output (noise corrupted) training data
%  yt         is the output (not noise corrupted) testing data


scale = diag(abs(scaled_rand(max_exp,min_exp,d)));
% these b's are for the same direction as singular vectors
for i=1:o
    b(:,i) = scaled_rand(max_exp,min_exp,d);
end

% need to build random orthogonal matrix
% only rotate d - d_ind columns; let the rest be
% 'approx' independent

d_rot =  d - d_ind;
if  d_ind == d
  v = eye(d);
else
  rand('uniform')
  v = rand (d_rot);
  [u,s,v] = svd(v);
  if d_rot == d
      v = u*v;
  else
      v = [ eye(d_ind), zeros(d_ind,d_rot); zeros(d_rot,d_ind), u*v];
  end

end
```

```
% use v as an additional rotation on the data and regression vector

rand('normal')
X =  rand(n_train,d) * scale * v;
Xt =  rand(n_test,d) * scale * v;
b = v'*b;
yt = Xt*b;
%desire RMS of null predictor to be 1
rms = sqrt(trace(yt'*yt)/ (n_test * o) ) ;
b=b/rms;
yt = Xt*b;
y = X*b + rand(n_train,o)*noise;
```

D.2 Routine to generate "exponential" random numbers

```
function vect =  scaled_rand(u,l,d)

% this function generates a vector of random numbers that are
% 'exponentially' distributed;  that is, the probability of
%  a number having any given  order of magnitude within
%  the valid range is roughly equal
%
%   u      lowest order of magnitude allowed
%   l      highest order of magnitude allowed
%   d      is the dimension of the vector generated
%
%  10^l < number < 10^u
%

rand('uniform');

for i = 1:d
```

```
  vect(i) = 10^ ( (u - 1) * rand(1,1)  + 1);
end

vect = vect';
```