Significance Regression: Improved Estimation from Collinear Data for the Measurement Error Model

Tyler R. Holcomb

Manfred Morari *

Control and Dynamical Systems 210-41

California Institute of Technology

Pasadena CA 91125

Keywords: measurement error models, biased regression, partial least squares, multivariable regression, significance regression, collinearity

CIT-CDS Technical Memo 93-004

May 12, 1993

Abstract

This paper examines improved regression methods for the linear multivariable measurement error model (MEM) when the data suffers from "collinearity." The difficulty collinearity presents for reliable estimation is discussed and a systematic procedure, significance regression (SR-MEM), is developed to address collinearity. In addition to mitigating collinearity difficulties SR-MEM produces asymptotically unbiased estimates. The use of ordinary least squares (OLS) for the MEM is examined. For collinear data OLS can improve the mean squared error of estimation over the maximum likelihood (ML) unbiased estimator in a manner analogous to ridge regression (RR). The significance regression method developed for the classical model

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

(SR-classical) can also be used for data with measurement errors. SR-classical is similar SR-MEM and can yield better estimation than the ML estimator for collinear data. Numerical examples illustrate several points.

1 Introduction

The classical model assumes that the explanatory variables are known without error. However practitioners must often work with data where all the measurements are corrupted by measurement noise, not just the the dependent variables. For multivariable data sets there are also commonly correlations among the explanatory variables. These correlations increase the variance of the regressor and can cause unreliable prediction and estimation. This paper will address regression problems with strong correlations between the explanatory variables: that is, "collinear" problems. In particular, the data are assumed to be described by the measurement error model (MEM):

$$\mathbf{y} = \mathbf{Tr} + \mathbf{e}, \quad \text{and} \tag{1}$$

$$\mathbf{X} = \mathbf{T} + \mathbf{S}. \tag{2}$$

In this formulation, $\mathbf{T} \in \Re^{n_s \times n_i}$ represents the "true" but unobservable explanatory variables, while $\mathbf{X} \in \Re^{n_s \times n_i}$ and $\mathbf{y} \in \Re^{n_s}$ represent the n_s observations of the explanatory and dependent variables, respectively. Since this work focuses on collinear problems, it will be most applicable to problems where the condition number of $\mathbf{T}^T\mathbf{T}$ is "large." The unobservable errors affecting the explanatory and dependent variables are $\mathbf{S} \in \Re^{n_s \times n_i}$ and $\mathbf{e} \in \Re^{n_s}$, respectively. Problems with multiple dependent variables, where $\mathbf{Y} \in \Re^{n_s \times n_o}$, can also be treated using the methods presented in this paper [Holcomb et al., 1993, Holcomb and Morari, 1993]. The assumptions used in this paper are:

(A1) $\mathbf{S}$ and $\mathbf{T}$ are stochastically independent.

(A2) The elements of $\mathbf{T}$ are fixed (but unknown) variates.

(A3) $\mathcal{E}(\mathbf{S}) = 0$, $\mathcal{E}(\mathbf{e}) = 0$.

(A4) $\mathbf{S}$ and $\mathbf{e}$ are stochastically independent.

(A5) The fourth moments of the distributions of all the elements of $\mathbf{S}$ and $\mathbf{e}$ exist.

(A6) $\lim_{n_s\to\infty} \mathbf{T}^T\mathbf{T}/n_s = \mathbf{M}_T$ exists and is non-singular.

(A7) Each row of $\mathbf{S}$, $\mathbf{s}^T$, is stochastically independent and identically distributed, with $\mathcal{E}\left(\mathbf{s}\mathbf{s}^T\right) = \mathbf{\Sigma}$ and $\mathbf{\Sigma}$ non-singular. Likewise, the elements of $\mathbf{e}$ are also stochastically independent and homoscedastic with $\mathcal{E}\left(ee^T\right) = \sigma_e^2\mathbf{I}$.

(A8) In addition to (A6), as $n_s \to \infty$, $\mathbf{T}^T\mathbf{T} \to n_s\mathbf{M}_T$.

Further define $\mathbf{M}_X = \mathbf{M}_T + \mathbf{\Sigma}$. (A4) can be readily relaxed, but at the expense of more involved notation. (A2) can be relaxed by assuming that all limits are regular with probability one. For this more general assumption, all of the results of this paper hold "conditioned on $\mathbf{T}$." See Schneeweiß [Schneeweiß, 1976] for further discussion of these assumptions, their implications, and how to relax them.

For this model an asymptotically unbiased estimate of $\mathbf{r}$ is

$$\tilde{\mathbf{r}} = (\mathbf{X}^T\mathbf{X} - n_s\mathbf{\Sigma})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3}$$

which results directly from the minimization problem

$$\tilde{\mathbf{r}} = \arg\min_{\mathbf{b}\in\Re^{n_i}} (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) - n_s\mathbf{b}^T\mathbf{\Sigma}\mathbf{b}. \tag{4}$$

Under the additional assumption of normally distributed errors $\tilde{\mathbf{r}}$ is the maximum likelihood estimate of $\mathbf{r}$ [Lindley, 1947, Johnston, 1972]. Schneeweiß [Schneeweiß, 1976] has also determined the asymptotic properties of $\tilde{\mathbf{r}}$. Specifically, $\sqrt{n_s}(\tilde{\mathbf{r}} - \mathbf{r})$ has an asymptotically normal distribution with variance

$$\begin{aligned} \mathrm{Var}_\infty\left(\sqrt{n_s}(\tilde{\mathbf{r}} - \mathbf{r})\right) &= \mathbf{M}_T^{-1}(\mathbf{\Sigma}\mathbf{r}\mathbf{r}^T\mathbf{\Sigma} + \sigma_e^2\mathbf{M}_X)\mathbf{M}_T^{-1} && (5)\\ &= \lim_{n_s\to\infty} n_s(\tilde{\mathbf{r}} - \mathbf{r})(\tilde{\mathbf{r}} - \mathbf{r})^T. && (6)\end{aligned}$$

Equation 6 does not follow directly from Schneeweiß's development, but rather from Theorem 4.5.2 of [Chung, 1974]. If $\mathbf{M}_T\mathbf{M}_X^{-1}\mathbf{M}_T$ has one or more "small" eigenvalues then $\tilde{\mathbf{r}}$ will have "large" variance. If correlations exist among the explanatory variables then $\mathbf{M}_T$ will be "nearly" singular, $\mathbf{M}_T$ will have at least one "small" eigenvalue, $\mathbf{M}_T\mathbf{M}_X^{-1}\mathbf{M}_T$ will tend to have at least one "small" eigenvalue, and $\tilde{\mathbf{r}}$ will tend to have a "large" variance.. Thus, the well-known "collinearity problem" that bedevils ordinary least squares (OLS) estimators also poses difficulty for estimation in the MEM framework.

2 Significance Regression for the Classical Model

One can mollify the collinearity problem for classical regression by employing techniques such as stepwise regression [Draper and Smith, 1966], ridge regression [Hoerl and Kennard, 1970], principal components regression [Hill et al., 1977], and significance regression (SR) [Holcomb et al., 1993]. To build the foundation for addressing the collinearity problem in the context of the MEM, this section examines SR for problems of the form

$$\mathbf{y} = \mathbf{X}\mathbf{r} + \mathbf{e} \tag{7}$$

where $\mathcal{E}\left(\mathbf{e}\mathbf{e}^T\right) = \sigma_e^2\mathbf{I}$; equation 7 will be called the "classical model" in this paper. The SR approach encompasses the successful partial least squares algorithm [Wold et al., 1984], can have better prediction properties than ridge regression and principal components regression for a variety of problems [Fearn, 1983, Lorber et al., 1987, Stone and Brooks, 1990], and rests on a rigorous foundation that can be readily and clearly adapted to the MEM. A comprehensive motivation and derivation for significance regression for the classical model is presented in [Holcomb et al., 1993]; only the main points are considered here.

Typically, the specification of a regressor can be expressed as an unconstrained optimization problem. For example, the classical ordinary least squares regressor,

$$\tilde{\mathbf{p}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{8}$$

results directly from the minimization problem

$$\tilde{\mathbf{p}} = \arg\min_{\mathbf{b}\in\Re^{n_i}} (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}). \tag{9}$$

The variance of the regressor can be reduced if one constrains the allowable values for the final regressor. For example ridge regression uses a "soft" constraint derived from assuming a prior distribution for $\mathbf{r}$ [Gruber, 1990]. For ridge regression, the appropriate optimization problem is

$$\arg\min_{\mathbf{b}\in\Re^{n_i}} (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) + \mathbf{b}^T\mathbf{A}\mathbf{b}. \tag{10}$$

for some positive definite $\mathbf{A}$ that describes the prior distribution. Another approach is to constrain the regressor to a prespecified subspace, as in

$$\tilde{\mathbf{b}} = \arg\min_{\mathbf{b}\in\mathrm{Range}(\mathbf{W})} (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) \tag{11}$$

where $\mathbf{W} \in \Re^{n_i \times n_d}$ consists of orthonormal columns. For stepwise regression each column would consist of unit vectors describing coordinate axes. For example if one chooses to use the second and third of three variables, then $\mathbf{W} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$. For principal components regression $\mathbf{W}$ would be built from the principal components of $\mathbf{X}$. But how to choose the "best" $\mathbf{W}$? First, one clearly desires $\mathbf{r} \in \text{Range}(\mathbf{W})$ to assure that the regressor is an unbiased estimator. Moreover, if $< \mathbf{w}, \mathbf{r} > = \mathbf{w}^T \mathbf{r} = 0$, then $\mathbf{w}$ should not be used as a column for $\mathbf{W}$ since this will increase the variance without affecting the bias. One can quantify and exploit these observations by postulating the null hypothesis

$$\mathcal{H}_0^1 : \qquad < \mathbf{w}, \mathbf{r} > = 0 \tag{12}$$

and searching for directions ($\mathbf{w}$) that reject it. A natural test statistic for $\mathcal{H}_0^1$ is

$$\tau_{classical}\ (\mathbf{w}, \mathbf{y}) = \frac{< \mathbf{w}, \tilde{\mathbf{p}} >\ ^2}{\text{Var}(< \mathbf{w}, \tilde{\mathbf{p}} >\ )} \tag{13}$$

for which $\tilde{\mathbf{p}}$ is the OLS estimator and $\text{Var}(< \mathbf{w}, \tilde{\mathbf{p}} >\ ) = \sigma_e^2 \mathbf{w}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{w}$. For any given $\mathbf{w}$ and normally distributed errors, $\tau_{classical}\ (\mathbf{w}, \mathbf{y})$ has a non-central χ^2 distribution with one degree of freedom; when $\mathcal{H}_0^1$ holds, the non-centrality parameter is zero. Thus, one can build $\mathbf{W}$ by seeking mutually orthogonal directions that successively maximize $\tau_{classical}\ (\mathbf{w}, \mathbf{y})$; this method is precisely significance regression. For the classical model, the SR algorithm is:

Algorithm 1 (SR-classical)

$$\tilde{\mathbf{p}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \tag{14}$$

$$\mathbf{V}_{classical} = \sigma_e^2 (\mathbf{X}^T \mathbf{X})^{-1} \tag{15}$$

$$\mathbf{W}_0 = [0 \cdots 0]^T, \quad \mathbf{W}_0 \in \Re^{n_i} \tag{16}$$

DO $i = 1, n_d$

$$\mathbf{w}_i^{opt}(\mathbf{y}) = \frac{(\mathbf{I} - \mathbf{W}_{i-1} \mathbf{W}_{i-1}^T) \mathbf{V}_{classical}^i \tilde{\mathbf{p}}}{||(\mathbf{I} - \mathbf{W}_{i-1} \mathbf{W}_{i-1}^T) \mathbf{V}_{classical}^i \tilde{\mathbf{p}}||} \tag{17}$$

$$\mathbf{W}_i = [\mathbf{w}_1^{opt} | \mathbf{w}_2^{opt} | \cdots | \mathbf{w}_i^{opt}] \tag{18}$$

END DO.

$$\tilde{\mathbf{b}} = \mathbf{W}_{n_d} (\mathbf{W}_{n_d}^T \mathbf{X}^T \mathbf{X} \mathbf{W}_{n_d})^{-1} \mathbf{W}_{n_d}^T \mathbf{X}^T \mathbf{y} \tag{19}$$

To determine n_d, one can use use either cross-validation [Stone, 1974] on the prediction error or one can use hypothesis testing. In particular, if the null hypothesis

$$\mathcal{H}_0^{2,i}: \qquad < \mathbf{w}, \mathbf{r} > = 0 \quad \text{for all } w \in \text{Range}(I - \mathbf{W}_{i-1}\mathbf{W}_{i-1}^T) \tag{20}$$

is true then $n_d < i$. One can test $\mathcal{H}_0^{2,i}$ using the test statistic $\tau_{classical,i}^{opt}(\mathbf{y}) = \tau_{classical}(\mathbf{w}_i^{opt}(\mathbf{y}), \mathbf{y})$. For normal errors and the assumption that $\mathbf{w}_i^{opt}(\mathbf{y})$ is independent of $\mathbf{W}_{i-1}$, $\tau_{classical,i}^{opt}(\mathbf{y})$ has a non-central χ^2 distribution in $n_p = n_i - i + 1$ degrees of freedom. The non-centrality parameter is zero when $\mathcal{H}_0^{2,i}$ holds. Since $\mathbf{W}_{i-1}$ is a function of $\mathbf{y}$, the independence assumption is not strictly true. However the assumption is true as n_s becomes large. Moreover, the distributions computed using the independence assumption will be good approximations to the "true" distribution when the non-centrality parameter dominates the variance for all of the directions (columns) of $\mathbf{W}_{i-1}$. Importantly, these are the directions SR seeks. This independence assumption will be inappropriate when the value of the non-centrality parameter approaches n_p for any of the directions of $\mathbf{W}_{i-1}$. When the independence assumption is unwarranted, the χ^2-distribution will have too heavy of a tail relative to the "true" distribution. Therefore the n_d determined using the χ^2 test derived using the independence assumption will be less than or equal to the n_d one would determine using the "true" distribution.

3 Significance Regression for the Measurement Error Model

To construct an estimation method for collinear data in the MEM context, one proceeds in the same manner as before. First one considers the null hypothesis

$$\mathcal{H}_0^1: \qquad < \mathbf{w}, \mathbf{r} > = 0, \tag{21}$$

for which a natural test statistic is

$$\tau_{ideal}\,(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{< \mathbf{w}, \tilde{\mathbf{r}} >^{\;2}}{\text{Var}(< \mathbf{w}, \tilde{\mathbf{r}} >)}. \tag{22}$$

Computing $\text{Var}(< \mathbf{w}, \tilde{\mathbf{r}} >)$ is involved; however, one can use equation 5 to discern that as $n_s \to \infty$

$$\text{Var}(< \mathbf{w}, \tilde{\mathbf{r}} >) \to \frac{1}{n_s}\mathbf{w}^T\mathbf{M}_T^{-1}(\boldsymbol{\Sigma}\mathbf{r}\mathbf{r}^T\boldsymbol{\Sigma} + \sigma_e^2\mathbf{M}_X)\mathbf{M}_T^{-1}\mathbf{w} \tag{23}$$

For ease of notation let $\mathbf{V}_{ideal} = (1/n_s)\mathbf{M}_T^{-1}(\boldsymbol{\Sigma}\mathbf{r}\mathbf{r}^T\boldsymbol{\Sigma} + \sigma_e^2\mathbf{M}_X)\mathbf{M}_T^{-1}$. Unlike the classical case, $\mathbf{V}_{ideal}$ includes several terms that must themselves be estimated. First, one must determine $\mathbf{M}_T$. By (A2) the $\mathbf{T}$ are fixed variates. Lacking other information, one may conjecture that the values of $\mathbf{T}$ are repeated as $n_s \to \infty$, consider $\mathbf{M}_T$ to be $\mathbf{T}^T\mathbf{T}/n_s$, and approximate $\mathbf{M}_X$ as $\mathbf{X}^T\mathbf{X}/n_s$. A more difficult problem is the explicit appearance of the unknown vector $\mathbf{r}$ in equation 22. One can use the ML estimate $\tilde{\mathbf{r}}$ in place of $\mathbf{r}$, leading to the approximation

$$\mathbf{V}_{approx} = \frac{1}{n_s}\left(\frac{\mathbf{X}^T\mathbf{X}}{n_s} - \mathbf{\Sigma}\right)^{-1}\left(\mathbf{\Sigma}\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\mathbf{\Sigma} + \sigma_e^2\frac{\mathbf{X}^T\mathbf{X}}{n_s}\right)\left(\frac{\mathbf{X}^T\mathbf{X}}{n_s} - \mathbf{\Sigma}\right)^{-1}. \tag{24}$$

This approximation is asymptotically valid:

$$\lim_{n_s\to\infty} n_s\mathbf{V}_{approx} = n_s\mathbf{V}_{ideal} \tag{25}$$

As discussed below, this approximation does affect the distributional properties of $\tau_{ideal}(\mathbf{w}, \mathbf{X}, \mathbf{y})$. With these approximations, the approximate test statistic is

$$\tau_{approx}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{<\mathbf{w}, \tilde{\mathbf{r}}>^{\ 2}}{\mathbf{w}^T\mathbf{V}_{approx}\mathbf{w}}, \tag{26}$$

and the resulting significance regression algorithm for the MEM is:

Algorithm 2 (SR-MEM)

$$\tilde{\mathbf{r}} = (\mathbf{X}^T\mathbf{X} - n_s\mathbf{\Sigma})^{-1}\mathbf{X}^T\mathbf{y} \tag{27}$$

$$\mathbf{V}_{approx} = \frac{1}{n_s}\left(\frac{\mathbf{X}^T\mathbf{X}}{n_s} - \mathbf{\Sigma}\right)^{-1}\left(\mathbf{\Sigma}\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\mathbf{\Sigma} + \sigma_e^2\frac{\mathbf{X}^T\mathbf{X}}{n_s}\right)\left(\frac{\mathbf{X}^T\mathbf{X}}{n_s} - \mathbf{\Sigma}\right)^{-1} \tag{28}$$

$$\mathbf{W}_0 = [0\cdots0]^T, \quad \mathbf{W}_0 \in \Re^{n_i} \tag{29}$$

DO $i = 1, n_d$

$$\mathbf{w}_i^{opt}(\mathbf{X}, \mathbf{y}) = \frac{(\mathbf{I} - \mathbf{W}_{i-1}\mathbf{W}_{i-1}^T)\mathbf{V}_{approx}^{-i}\tilde{\mathbf{r}}}{||(\mathbf{I} - \mathbf{W}_{i-1}\mathbf{W}_{i-1}^T)\mathbf{V}_{approx}^{-i}\tilde{\mathbf{r}}||} \tag{30}$$

$$\mathbf{W}_i = [\mathbf{w}_1^{opt}|\mathbf{w}_2^{opt}|\cdots|\mathbf{w}_i^{opt}] \tag{31}$$

END DO.

$$\tilde{\mathbf{b}} = \mathbf{W}_{n_d}(\mathbf{W}_{n_d}^T(\mathbf{X}^T\mathbf{X} - n_s\mathbf{\Sigma})\mathbf{W}_{n_d})^{-1}\mathbf{W}_{n_d}^T\mathbf{X}^T\mathbf{y} \tag{32}$$

The $\mathbf{w}_i^{opt}$ and the associated $\tau_{approx,i}^{opt}(\mathbf{X}, \mathbf{y}) = \tau_{approx}(\mathbf{w}_i^{opt}, \mathbf{X}, \mathbf{y})$ have several useful properties. First, assume $\mathbf{w}_i^{opt}$ was computed using $\mathbf{V}_{ideal}$. Then

$$\frac{1}{n_s}\tau_{ideal,1}^{opt}(\mathbf{X}, \mathbf{y}) = \frac{(\tilde{\mathbf{r}}^T\mathbf{w}_1^{opt})^2}{n_s\mathbf{w}_1^{opt^T}\mathbf{V}_{ideal}\mathbf{w}_1^{opt}} \tag{33}$$

$$= \frac{(\tilde{\mathbf{r}}^T\mathbf{V}_{ideal}^{-1}\tilde{\mathbf{r}})^2}{n_s\tilde{\mathbf{r}}^T\mathbf{V}_{ideal}^{-1}\mathbf{V}_{ideal}\mathbf{V}_{ideal}^{-1}\tilde{\mathbf{r}}} \tag{34}$$

$$= \frac{1}{n_s}\tilde{\mathbf{r}}^T\mathbf{V}_{ideal}^{-1}\tilde{\mathbf{r}} \tag{35}$$

which asymptotically approaches a non-central χ^2-distribution with n_i degrees of freedom; when $\mathcal{H}_0^1$ holds for $\mathbf{w}_1^{opt}$, then $(1/n_s)\tau_{ideal,1}^{opt}(\mathbf{X},\mathbf{y})$ approaches a central χ^2-distribution [Billingsley, 1968, Theorem 5.1]. In practice one uses the asymptotically equivalent $\mathbf{V}_{approx}$ for which

$$\lim_{n_s\to\infty}\frac{1}{n_s}\tau_{approx}\left(\mathbf{w},\mathbf{X},\mathbf{y}\right)=\lim_{n_s\to\infty}\frac{1}{n_s}\tau_{ideal}\left(\mathbf{w},\mathbf{X},\mathbf{y}\right)\quad\forall\mathbf{w}. \tag{36}$$

For any given n_s, the distribution of $\tau_{approx,i}^{opt}(\mathbf{X},\mathbf{y})$ will differ from $\tau_{ideal,i}^{opt}(\mathbf{X},\mathbf{y})$. However, as shown in section 5 below, the distributions are typically similar in practice.

Lastly, $\tilde{\mathbf{b}}$ shares a crucial property with $\tilde{\mathbf{r}}$: $\tilde{\mathbf{b}}$ is asymptotically unbiased. For any $\mathbf{w}$,

$$\begin{aligned}
\lim_{n_s\to\infty}\tau_{approx}\left(\mathbf{w},\mathbf{X},\mathbf{y})\right) &= \lim_{n_s\to\infty}\frac{(\mathbf{w}^T\tilde{\mathbf{r}})^2}{\mathbf{w}^T\mathbf{V}_{approx}\mathbf{w}} && (37)\\
&= \lim_{n_s\to\infty}\frac{n_s(\mathbf{w}^T\tilde{\mathbf{r}})^2}{n_s\mathbf{w}^T\mathbf{V}_{approx}\mathbf{w}} && (38)\\
&= \begin{cases}\infty & \text{if } \mathbf{w}^T\mathbf{r}\neq 0\\ 0 & \text{if } \mathbf{w}^T\mathbf{r}=0\end{cases} && (39)
\end{aligned}$$

When n_s is large enough, $(\tau_{approx}\left(\mathbf{w},\mathbf{X},\mathbf{y})\right)$ will be large enough to overcome any given threshold for "significance" for all directions where $\mathbf{w}^T\mathbf{r}\neq 0$. This means that if the χ^2 test described above is used to determine n_d, then for n_s sufficiently large $\mathbf{r}\in\text{Range}(\mathbf{W}_{n_d})$ and $\tilde{\mathbf{b}}$ is an asymptotically unbiased estimator of $\mathbf{r}$.

4 Use of Classical Model Methods on the Measurement Error Model

The above section developed the SR algorithm for the MEM. This section examines the implications of using the OLS and SR methods with collinear data with measurement errors. Both methods will be shown to have certain advantages over $\tilde{\mathbf{r}}$. Using the classical ordinary least-squares regressor ($\tilde{\mathbf{p}}$, equation 8) for a measurement error model produces an asymptotically biased estimate of $\mathbf{r}$ since $\lim_{n_s\to\infty}(\tilde{\mathbf{p}}-\mathbf{r})=-\mathbf{M}_X^{-1}\mathbf{\Sigma}\mathbf{r}\neq 0$. However, $\tilde{\mathbf{p}}$ is an asymptotically unbiased estimate of the least-squares optimal predictor $\mathbf{p}$; see Fact 1 and Fact 2 of appendix B, [Berkson, 1950], and [Schneeweiß, 1976] for further discussion. Moreover, $\tilde{\mathbf{p}}$ acts as a "natural ridge regressor." For the MEM (equation 4), the natural generalization of classical ridge regressor (equation 10) is

$$\arg\min_{\mathbf{b}\in\Re^{n_i}}(\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b})-n_s\mathbf{b}^T\mathbf{\Sigma}\mathbf{b}+\mathbf{b}^T\mathbf{A}\mathbf{b}. \tag{40}$$

Letting $\mathbf{A} = n_s \boldsymbol{\Sigma}$ produces a "generalized MEM ridge regressor" $\tilde{\mathbf{p}} = (\mathbf{X}^T\mathbf{X} - n_s\boldsymbol{\Sigma} + n_s\boldsymbol{\Sigma})^{-1}\mathbf{X}^T\mathbf{y}$. Just as ridge regression has a mean-squared-error (MSE) advantage over OLS for the classical model for noisy and/or collinear data, so this MEM ridge regressor, $\tilde{\mathbf{p}}$, can have an MSE advantage over $\tilde{\mathbf{r}}$. Define

$$\mathrm{MSE}(\mathbf{v}) = \mathcal{E}\left((\mathbf{v} - \tilde{\mathbf{r}})(\mathbf{v} - \tilde{\mathbf{r}})^T\right) \tag{41}$$

for any vector $\mathbf{v}$. Note that

$$\begin{aligned} \mathrm{MSE}(\tilde{\mathbf{r}}) &\to \frac{1}{n_s}\mathbf{M}_T^{-1}(\boldsymbol{\Sigma}\mathbf{r}\mathbf{r}^T\boldsymbol{\Sigma} + \sigma_e^2\mathbf{M}_X)\mathbf{M}_T^{-1} \quad \text{and} & (42)\\ \mathrm{MSE}(\tilde{\mathbf{p}}) &\to \frac{1}{n_s}\mathbf{M}_X^{-1}(\boldsymbol{\Sigma}\mathbf{r}\mathbf{r}^T\boldsymbol{\Sigma} + \sigma_e^2\mathbf{M}_X)\mathbf{M}_X^{-1} + \mathbf{M}_X^{-1}\boldsymbol{\Sigma}\mathbf{r}\mathbf{r}^T\boldsymbol{\Sigma}\mathbf{M}_X^{-1} & (43) \end{aligned}$$

as $n_s \to \infty$; equation 43 is verified in Fact 4. As shown in Fact 7, either $\mathbf{r}^T\mathbf{M}_T\boldsymbol{\Sigma}^{-1}\mathbf{M}_T\mathbf{r}$ being "small" or σ_e^2 being "large" favor $\mathrm{MSE}(\tilde{\mathbf{p}})$ over $\mathrm{MSE}(\tilde{\mathbf{r}})$. These conditions can be loosely described as "poor signal-to-noise (SNR) ratio." Notice that even if all the individual explanatory variables have "good" SNR (the diagonal element of $\mathbf{M}_T$ is "large" relative the corresponding diagonal element of $\boldsymbol{\Sigma}$), collinearity may result in $\boldsymbol{\Sigma}^{-1/2}\mathbf{M}_T$ being "small" in the crucial $\mathbf{r}$ direction.

While $\tilde{\mathbf{p}}$ may be preferable to $\tilde{\mathbf{r}}$ for collinear data, $\tilde{\mathbf{p}}$ has its own well known difficulties with collinearity. As shown above, a more direct approach to collinearity for the classical model is SR-classical. SR-classical only differs from SR-MEM in two respects: the selection of the search space ($\mathbf{W}_{n_d}$) and the construction of the final estimate, $\tilde{\mathbf{b}}$. SR-classical proceeds by assuming

$$\mathrm{Var}_\infty\left(\sqrt{n_s}(\tilde{\mathbf{p}} - \mathbf{p})\right) = \sigma_e^2\mathbf{M}_X^{-1}. \tag{44}$$

However as shown in Fact 3,

$$\begin{aligned} \mathrm{Var}_\infty\left(\sqrt{n_s}(\tilde{\mathbf{p}} - \mathbf{p})\right) &= \lim_{n_s\to\infty} n_s(\tilde{\mathbf{p}} - \mathbf{p})(\tilde{\mathbf{p}} - \mathbf{p})^T & (45)\\ &= \mathbf{M}_X^{-1}(\boldsymbol{\Sigma}\mathbf{r}\mathbf{r}^T\boldsymbol{\Sigma} + \sigma_e^2\mathbf{M}_X)\mathbf{M}_X^{-1} & (46)\\ &= \mathbf{M}_X^{-1}\boldsymbol{\Sigma}\mathbf{r}\mathbf{r}^T\boldsymbol{\Sigma}\mathbf{M}_X^{-1} + \sigma_e^2\mathbf{M}_X^{-1}. & (47) \end{aligned}$$

While equation 47 is not identical equation 44, the only difference occurs in the $\mathbf{M}_X^{-1}\boldsymbol{\Sigma}\mathbf{r}$ direction; the other $n_i - 1$ directions are unaltered. Thus the two methods will not necessarily produce dramatically different $\mathbf{W}_{n_d}$. The other difference between SR-classical and SR-MEM is that SR-classical will produce a biased estimate even when $\mathbf{r} \in \mathrm{Range}(\mathbf{W}_{n_d})$.

However, for multivariable data tolerance of the collinearity tends to be more important than unbiasedness (hence the success of SR and other biased estimators for the classical model), so SR-classical will tend to exhibit much of the benefit of SR-MEM and will often be preferable to the ML estimator for collinear data.

5 Simulation Examples

The above development focused on theoretical understanding and derivation. This section presents several numerical examples that illustrate points made above. 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 on problems that illustrate particular difficulties. The "real world" successes of partial least squares (PLS) algorithm (*e.g.* [Martens and Næs, 1989, Mejdell, 1990, Ricker, 1988]) are suggested as evidence of the practical utility of SR since the PLS algorithm is closely related to SR.

The regression methods investigated were

- asymptotically unbiased estimation (ML, equation 3),
- ordinary least squares (OLS, equation 8),
- the MEM significance regression method (SR-MEM, algorithm 2), and
- the classical model significance regression method (SR-classical, algorithm 1).

In all cases, a 90% significance threshold was used to evaluate $\mathcal{H}_0^{2,i}$ and to determine n_d. All examples had ten explanatory variables ($n_i = 10$) and four dependent variables ($n_o = 4$). For each 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 C. The examples tended to be collinear in that the singular values of $\mathbf{T}$ (the square root of the eigenvalues of $\mathbf{T}^T\mathbf{T}$) and the values of the regression parameters varied over five orders of magnitude; moreover there were typically large variances in the explanatory data that had little effect on the dependent variables. The same one thousand examples were used in all studies. All errors were independent and homoscedastic; however the variances of the errors for each explanatory variable varied by one order of magnitude.

Since the examples were synthetic, $\mathbf{r}$ was known and a point estimate of the trace of the MSE could be computed for each example. Thus the measure used to evaluate the various regression methods was

$$RMS = \sqrt{\frac{\mathrm{Tr}(\ (\tilde{\mathbf{b}} - \mathbf{r})(\tilde{\mathbf{b}} - \mathbf{r})^T\ )}{\mathbf{r}^T\mathbf{r}}}. \tag{48}$$

The $\mathbf{r}^T\mathbf{r}$ term was included to produce a relative error and allow averaging over all one thousand examples. Also, for each example, the rank (relative performance) of each estimator was recorded and averaged over all examples computed. 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. If the performance of two regressors differed by less than 0.1%, both regressors were given the same rank. In all examples, thirty samples were used to compute the regressor ($n_s = 30$).

First studied was the impact of using $\mathbf{V}_{approx}$ in place of $\mathbf{V}_{ideal}$. Using the first synthetic example, the distribution of both $\tau^{opt}_{ideal,1}(\mathbf{X}, \mathbf{y})$ and $\tau^{opt}_{approx,1}(\mathbf{X}, \mathbf{y})$ were determined via Monte Carlo simulation. Since the *variation* of $\tilde{\mathbf{r}}$ causes the deviation from the "ideal" distribution, not the particular value of $\mathbf{r}$, $\mathbf{r}$ was set to zero. The distributions were sampled using 60 equal-width intervals between zero and thirty. One million samples were drawn from the (normal) distributions for $\mathbf{S}$ and $\mathbf{e}$, and the frequency for each interval was recorded. The results are plotted in figure one. The crosses ('+') are the frequencies of $\tau^{opt}_{ideal,1}(\mathbf{X}, \mathbf{y})$, while the circles ('o') are the frequencies of $\tau^{opt}_{approx,1}(\mathbf{X}, \mathbf{y})$. As one can see, the two test statistics had virtually identical distributions. In 99.5% of the samples, the values of $\tau^{opt}_{ideal,1}(\mathbf{X}, \mathbf{y})$ and $\tau^{opt}_{approx,1}(\mathbf{X}, \mathbf{y})$ lay in the same interval. The solid line is the probability density function for the χ^2 distribution with 10 degrees of freedom normalized for abscissa used in Figure 1. The distributions for both test statistics closely conform to each other and the "asymptotic" χ^2 distribution.

Next studied was the effectiveness of SR-MEM; these results are in Table 1. SR-MEM had an $\overline{RMS}$ three orders of magnitude less than that of ML. Clearly SR-MEM mollified much of the difficulty caused by the correlations among the explanatory variables. The "ridging effect" of using $\tilde{\mathbf{p}}$ for estimating $\mathbf{r}$ was also examined; these results are shown in Table 2. In this example, the performance degradation due to bias was more than offset by the reduction in variance: OLS reduced the $\overline{RMS}$ by two orders of magnitude. Last studied was the effectiveness of using SR-classical. As shown in Table 3, SR was almost

Figure 1: Distribution of test statistic as determined by $1,000,000$ sample Monte Carlo simulation. Solid line, χ^2 distribution with ten degrees of freedom; "+," $\tau^{opt}_{ideal,1}(\mathbf{X}, \mathbf{y})$; "o," $\tau^{opt}_{approx,1}(\mathbf{X}, \mathbf{y})$.

method	$\overline{RMS}$	$\overline{\text{rank}}$
ML	$3,200$	2.0
SR − MEM	1.3	1.0

Table 1: Comparison of MEM-based methods over $1,000$ examples of synthetic data.

method	$\overline{RMS}$	$\overline{\text{rank}}$
ML	$3,200$	2.0
OLS	120	1.0

Table 2: Comparison of asymptotically unbiased estimator versus least-squares estimator over $1,000$ examples of synthetic data.

method	$\overline{RMS}$	$\overline{\text{rank}}$
ML	$3,200$	2.8
SR	5.6	1.5
$\text{SR} - \text{MEM}$	1.3	1.3

Table 3: Comparison SR-MEM and SR methods over $1,000$ examples of synthetic data.

method	$\overline{RMS_{\text{PRESS}}}$	$\overline{\text{rank}}$
null estimator	1.00	4.0
OLS	0.145	2.8
SR	0.103	1.6
$\text{SR} - \text{MEM}$	0.103	1.6

Table 4: Prediction performance over $1,000$ examples of synthetic data.

as good as SR-MEM. Thus, these simulations suggest that one can use SR-classical for collinear data with measurement errors without undue performance loss relative to SR-MEM and with considerable performance benefit relative to ML. Although not shown here, similar results were obtained if a 95% or 50% significance criterion was used; for these examples the performance of SR-MEM and SR did not strongly depend on the choice of significance level.

One might object that none of the estimators did better than the null estimator: using the estimator "always estimate zero" yields $\overline{RMS} = 1$. However, most of the synthetic examples had "large" components of $\mathbf{r}$ in directions where $\mathbf{T}$ had "small" variance. Thus the "success" of the null estimator was a reflection of the difficult nature of the examples used. With the null estimator one disavows using any variance information and instead relies solely on the mean of the training data. With the significance regression, the space where estimation is not attempted, and therefore bias may exist, is precisely $\mathbf{I} - \mathbf{W}_{n_d}\mathbf{W}_{n_d}^T$. Thus significance regression proceeds where the data are "significant" and suggests where to be wary. In the subspaces with "significant" data, the null estimator performed very poorly, as shown by studying predictive ability. To quantify predictive ability an additional one hundred samples ($\mathbf{X}_{new}$, $\mathbf{y}_{new}$) for each synthetic example were generated from the identical distribution as the training data, but the $\mathbf{y}_{new}$ were not corrupted by error ($e_{new} = 0$). Then

$$RMS_{\mathrm{PRESS}} = \sqrt{\frac{(\mathbf{X}_{new}\tilde{\mathbf{b}} - \mathbf{y}_{new})^T(\mathbf{X}_{new}\tilde{\mathbf{b}} - \mathbf{y}_{new})}{100}}. \tag{49}$$

Since the data were generated with the constraint

$$\sqrt{\frac{\mathbf{y}_{new}^T\mathbf{y}_{new}}{100}} = 1 \tag{50}$$

the RMS_{PRESS} was averaged over the examples without normalization. Since, as shown in Fact 2, OLS is more appropriate than MEM for prediction problems, the SR-MEM algorithm used

$$V = \frac{1}{n_s}(\frac{\mathbf{X}^T\mathbf{X}}{n_s})^{-1}(\mathbf{\Sigma}\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T\mathbf{\Sigma} + \sigma_e^2\frac{\mathbf{X}^T\mathbf{X}}{n_s})(\frac{\mathbf{X}^T\mathbf{X}}{n_s})^{-1} \tag{51}$$

and replaced equation 32 with

$$\tilde{\mathbf{b}} = \mathbf{W}_{n_d}(\mathbf{W}_{n_d}^T\mathbf{X}^T\mathbf{X}\mathbf{W}_{n_d})^{-1}\mathbf{W}_{n_d}^T\mathbf{X}^T\mathbf{y} \tag{52}$$

when computing $\tilde{\mathbf{b}}$ for prediction. The results are shown in Table 4. All of the methods did at least six times better than the null-estimator. Indeed, the null estimator was a worse predictor than all other investigated methods for all examples. Thus, the "superior" estimation performance of the null estimator occurred primarily in the subspace where the data were not "significant" — the space clearly delineated by SR-MEM. For prediction, SR-MEM and SR-classical construct $\tilde{\mathbf{b}}$ using 52; the only difference is the construction of $\mathbf{W}_{n_d}$. Consistent with the discussion at the close of section 4, SR and SR-MEM had similar predictive performances.

6 Conclusion

This work examined estimation and prediction from collinear data for the measurement error model (MEM) . A successful method for treating collinearity in the classical framework, significance regression (SR-classical), was generalized for the MEM. The resulting SR-MEM method improved estimation relative to the ML estimator and also provided asymptotically unbiased estimation. Also examined was the efficacy for the MEM of two methods derived for the classical model. The ordinary-least squares (OLS) regressor was seen to be the "optimal predictor" and a "natural ridge regressor" for MEM estimation problems. For noisy and/or collinear problems, OLS can have a smaller mean squared error (MSE) than the ML estimator. SR-classical was seen to compute search spaces ($\mathbf{W}$) similar to SR-MEM, but provide biased estimation. For multivariable data being tolerant of the collinearity tends to be more important than being unbiased, so SR-classical will tend to possess much of the benefit of SR-MEM and will often be preferable to the ML estimator for collinear data.

Acknowledgments: *The author's thank Håkan Hjalmarsson for his helpful comments concerning asymptotic distributions. This research was partially supported by the he Department of Energy, Office of Basic Energy Sciences, and 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 Manufacturing Company, and Shell Oil Company Foundation.*

References

[Berkson, 1950] Berkson, J. (1950). Are there two regressions? *American Statistical Association Journal*, 45:164–180.

[Billingsley, 1968] Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley.

[Chung, 1974] Chung, K. L. (1974). *A Course in Probability Theory.* Academic Press.

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

[Fearn, 1983] Fearn, T. (1983). A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *Applied Statistics*, 32(1):73–79.

[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.

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

[Hill et al., 1977] Hill, R. C., Fomby, T. B., and Johnson, S. (1977). Component selection norms for principal components regression. *Communications in Statistics A: Theory and Methods*, A6(4):309–334.

[Hoerl and Kennard, 1970] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

[Holcomb et al., 1993] Holcomb, T. R., Hjalmarsson, H., and Morari, M. (1993). Significance regression: A statistical approach to biased regression and partial least squares. CDS Technical Memo CIT-CDS 93-002, California Institute of Technology, Pasadena, CA 91125.

[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.

[Johnston, 1972] Johnston, J. (1972). *Econometric Methods, 2nd Edition.* McGraw-Hill.

[Lindley, 1947] Lindley, D. V. (1947). Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society*, B(Supplement) 52:218–244.

[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.

[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.

[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.

[Schneeweiß, 1976] Schneeweiß, H. (1976). Consistent estimation of a regression with errors in the variables. *Metrika*, 23:101–115.

[Stone, 1974] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, B.36:111–147.

[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.

[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, boldface upper case letters (both Roman and Greek) represent matrices, boldface lower case letters represent column vectors, and lower case Greek letters represent scalars. Estimates are denoted by a tilde, "˜". The dimensions of matrices are denote by subscripted n's.

operators

$[\mathbf{W} \mid \mathbf{V}]$	is the matrix formed by placing $\mathbf{W}$ and $\mathbf{V}$ side-by-side.
$< \cdot, \cdot >$	is the inner product. For vectors $\mathbf{a}$ and $\mathbf{b}$, $< \mathbf{A}, \mathbf{B} > = \mathbf{a}^T\mathbf{b}$.
$\mathbf{Diag}(\lambda_1, \ldots, \lambda_i)$	is the $i \times i$ diagonal matrix with corresponding λ_i's on the diagonal.
$\mathcal{E}(\cdot)$	is the expectation.
$\mathrm{MSE}(\cdot)$	is the Mean Square Error. See equation 41.
$\mathrm{Range}(\cdot)$	is the range; the span of the column vectors of a matrix.
$\mathrm{Tr}(\cdot)$	is the trace, the sum of the diagonal elements of a matrix.
$\mathrm{Var}(\cdot)$	is the variance.

some scalars, vectors, and matrices

$\tilde{\mathbf{b}}$	$n_i \times 1$	is the biased estimate of $\mathbf{r}$. See equation 32.
$\mathbf{I}$	as appropriate	is the identity matrix.
$\mathbf{p}$	$n_i \times 1$	is the least-squares optimal predictor. See equation 57.
$\tilde{\mathbf{p}}$	$n_i \times 1$	least-squares estimate of the MSE optimal predictor, also known as the Ordinary Least Squares (OLS) regressor. See equation 8.
$\mathbf{r}$	$n_i \times 1$	is the "true" regression vector. See equation 1.
$\tilde{\mathbf{r}}$	$n_i \times 1$	is the unbiased estimate of $\mathbf{r}$. See equation 3
$\mathbf{T}$	$n_s \times n_i$	is the "true" explanatory variable data. See equation 2.
$\mathbf{S}$	$n_s \times n_i$	is the measurement noise corrupting the explanatory data. $\mathbf{S}^T = [\mathbf{s}_1 \ \ \mathbf{s}_2 \ \ \ldots \ \ \mathbf{s}_{n_s}]$ See equation 2.
$\mathbf{s}_j$	$n_i \times 1$	is the measurement noise corrupting the explanatory data in the jth data sample.

$\mathbf{\Sigma}$	$n_i \times n_i$	is the measurement error covariance matrix for the explanatory data. See (A7).
$\mathbf{W}$	$n_i \times n_w$	is the matrix whose range defines the search space for $\tilde{\mathbf{b}}$. See equation 32.
$\mathbf{V}_{approx}$	$n_i \times n_i$	is the estimate of the variance matrix used in algorithm 2. See equation 28.
$\mathbf{V}_{ideal}$	$n_i \times n_i$	is the estimate of the variance matrix using the unknown quantity $\mathbf{r}$. See equation 23.
$\mathbf{v}$	varies $\times 1$	is a vector locally defined. Any given $\mathbf{v}$ may or may not relate to any other $\mathbf{v}$.
$\mathbf{x}_j$	$n_i \times 1$	is the measurement of the jth data sample.
$\mathbf{X}$	$n_s \times n_i$	is the measured explanatory data; each row corresponds to one sample of explanatory data. Thus, $\mathbf{X}^T = [\mathbf{x}_1 \;\; \mathbf{x}_2 \;\; \ldots \;\; \mathbf{x}_{n_s}]$. See equation 2.
$\mathbf{y}$	$n_s \times 1$	is the measured dependent variable. See equation 1.
$\tau(\mathbf{w}, \mathbf{X}, \mathbf{y})$	scalar	is the appropriate test statistic for $\mathbf{w}$ and a given $\mathbf{X}$ and $\mathbf{y}$. See equations 13, 22 and 26.
$\tau_i^{opt}(\mathbf{y})$	scalar	is the maximum of $\tau(\mathbf{w}, \mathbf{X}, \mathbf{y})$ over all allowable $\mathbf{w}$.

dimensional descriptors

n_d	is the number of "significant subspaces" to be generated.
n_i	is the number of explanatory variables.
n_s	is the number of samples.
n_p	is dimension of the allowable space in which to search for further $\mathbf{w}_i^{opt}$. For problems with a single dependent variable, $n_p = n_i - i + 1$.
n_w	is the rank of $\mathbf{W}$.

B Proofs for Facts

This section presents the arguments that verify several facts stated in section 4. Most of these facts are straightforward; they have been relegated to the appendix so that the requisite algebra would not distract from the main points. Any mention of Schneeweiß refers to [Schneeweiß, 1976].

Fact 1 *The least-squares optimal predictor is* $\mathbf{p} = (\mathbf{T}^T\mathbf{T}/n_s + \mathbf{\Sigma})^{-1}\mathbf{T}^T\mathbf{T}/n_s\mathbf{r}$.

We desire to compute

$$\mathbf{p} = \min_{v \in \Re^{n_i}} \mathcal{E}\left((\mathbf{y} - \mathbf{X}\mathbf{v})^T(\mathbf{y} - \mathbf{X}\mathbf{v})\right). \tag{53}$$

First compute the expectation:

$$\begin{aligned}
\mathcal{E}\left((\mathbf{y} - \mathbf{X}\mathbf{v})^T(\mathbf{y} - \mathbf{X}\mathbf{v})\right) &= \mathcal{E}\left((\mathbf{Tr} + \mathbf{e} - (\mathbf{T} + \mathbf{S})\mathbf{v})^T(\mathbf{Tr} + \mathbf{e} - (\mathbf{T} + \mathbf{S})\mathbf{v})\right) && (54)\\
&= \mathbf{r}^T\mathbf{T}^T\mathbf{Tr} + n_s\sigma_e^2 - 2\mathbf{r}^T\mathbf{T}^T\mathbf{Tv} + \mathbf{v}^T\mathbf{T}^T\mathbf{Tv} + n_s\mathbf{v}^T\mathbf{\Sigma v}. && (55)
\end{aligned}$$

Computing the gradient and equating it to zero,

$$\nabla_{\mathbf{v}}\mathcal{E}\left((\mathbf{y} - \mathbf{X}\mathbf{v})^T(\mathbf{y} - \mathbf{X}\mathbf{v})\right) = -2\mathbf{T}^T\mathbf{Tr} + 2\mathbf{T}^T\mathbf{Tv} + 2n_s\mathbf{\Sigma v} = 0, \tag{56}$$

implies

$$\mathbf{p} = (\frac{\mathbf{T}^T\mathbf{T}}{n_s} + \mathbf{\Sigma})^{-1}\frac{\mathbf{T}^T\mathbf{T}}{n_s}\mathbf{r}. \tag{57}$$

□

Fact 2 *The OLS regressor asymptotically equals the least-squares optimal predictor. That is* $lim_{n_s\to\infty}\tilde{\mathbf{p}} = lim_{n_s\to\infty}\mathbf{p}$.

$$\begin{aligned}
\lim_{n_s\to\infty}\tilde{\mathbf{p}} &= \lim_{n_s\to\infty}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} && (58)\\
&= \lim_{n_s\to\infty}\left(\frac{1}{n_s}(\mathbf{T}^T\mathbf{T} + \mathbf{T}^T\mathbf{S} + \mathbf{S}^T\mathbf{T} + \mathbf{S}^T\mathbf{S})\right)^{-1} && (59)\\
&\qquad \left(\frac{1}{n_s}(\mathbf{T} + \mathbf{S})^T(\mathbf{Tr} + \mathbf{v})\right) &&\\
&= \mathbf{M}_X^{-1}\mathbf{M}_T\mathbf{r} && (60)\\
&= \lim_{n_s\to\infty}\mathbf{p} && (61)
\end{aligned}$$

□

Fact 3 *($\sqrt{n_s}(\tilde{\mathbf{p}} - \mathbf{p})$) has an asymptotically normal distribution with zero mean and variance* $\mathbf{M}_X^{-1}(\mathbf{\Sigma}\mathbf{r}\mathbf{r}^T\mathbf{\Sigma} + \sigma_e^2\mathbf{M}_X)\mathbf{M}_X^{-1}$.

This fact is an algebraic variation of a result due to Schneeweiß. We make use of the algebraic relation $\mathbf{y} = \mathbf{X}\mathbf{r} - \mathbf{S}\mathbf{r} + \mathbf{e}$. Also define $\mathbf{S}_e = [\mathbf{e}|\ \mathbf{S}]$ and $\mathbf{g} = [1\ |\ -\tilde{\mathbf{r}}]^T$.

$$
\begin{aligned}
\tilde{\mathbf{p}} - \mathbf{p} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - \mathbf{p} && (62)\\
&= \left(\frac{\mathbf{X}^T\mathbf{X}}{n_s}\right)^{-1}\left(\frac{\mathbf{X}^T\mathbf{X}}{n_s}\mathbf{r} - \frac{\mathbf{X}^T\mathbf{S}}{n_s}\mathbf{r} + \frac{\mathbf{X}^T\mathbf{e}}{n_s} + \mathbf{\Sigma}\mathbf{r} - \mathbf{\Sigma}\mathbf{r}\right) - \mathbf{p} && (63)\\
&= \left(\frac{\mathbf{X}^T\mathbf{X}}{n_s}\right)^{-1}\left(\frac{\mathbf{T}^T\mathbf{T}}{n_s}\mathbf{r} + \frac{\mathbf{S}^T\mathbf{T}}{n_s}\mathbf{r} + \frac{\mathbf{T}^T\mathbf{S}}{n_s}\mathbf{r} + \frac{\mathbf{S}^T\mathbf{S}}{n_s}\mathbf{r} - \frac{\mathbf{T}^T\mathbf{S}}{n_s}\mathbf{r} - \frac{\mathbf{S}^T\mathbf{S}}{n_s}\mathbf{r}\right. && (64)\\
&\qquad \left. + \frac{\mathbf{T}^T\mathbf{e}}{n_s} + \frac{\mathbf{S}^T\mathbf{e}}{n_s} + \mathbf{\Sigma}\mathbf{r} - \mathbf{\Sigma}\mathbf{r}\right) - \mathbf{p}\\
&= \left(\frac{\mathbf{X}^T\mathbf{X}}{n_s}\right)^{-1}\left(\frac{\mathbf{T}^T\mathbf{S}_e}{n_s}\mathbf{g} + \frac{\mathbf{S}^T\mathbf{S}_e}{n_s}\mathbf{g} + \mathbf{\Sigma}\mathbf{r}\right) + && (65)\\
&\qquad \left(\frac{\mathbf{X}^T\mathbf{X}}{n_s}\right)^{-1}\left(\frac{\mathbf{T}^T\mathbf{T}}{n_s} + \frac{\mathbf{S}^T\mathbf{T}}{n_s} + \frac{\mathbf{T}^T\mathbf{S}}{n_s} + \frac{\mathbf{S}^T\mathbf{S}}{n_s} - \mathbf{\Sigma}\right)\mathbf{r} - \mathbf{p}.
\end{aligned}
$$

Now

$$
\begin{aligned}
&\lim_{n_s\to\infty}\left(\frac{\mathbf{X}^T\mathbf{X}}{n_s}\right)^{-1}\left(\frac{\mathbf{T}^T\mathbf{T}}{n_s} + \frac{\mathbf{S}^T\mathbf{T}}{n_s} + \frac{\mathbf{T}^T\mathbf{S}}{n_s} + \frac{\mathbf{S}^T\mathbf{S}}{n_s} - \mathbf{\Sigma}\right)\mathbf{r} && (66)\\
&\qquad = \mathbf{M}_X^{-1}(\mathbf{M}_T + \mathbf{\Sigma})\mathbf{r} - \mathbf{M}_X^{-1}\mathbf{\Sigma}\mathbf{r}\\
&\qquad = \mathbf{r} - \mathbf{M}_X^{-1}\mathbf{\Sigma}\mathbf{r} && (67)\\
&\qquad = \lim_{n_s\to\infty}\tilde{\mathbf{p}} && (68)\\
&\qquad = \lim_{n_s\to\infty}\mathbf{p}. && (69)
\end{aligned}
$$

When equation 65 is pre-multiplied by $\sqrt{n_s}$, the the second line of equation 65 still asymptotically vanishes due to (A8). Moreover, the $\left(\mathbf{T}^T\mathbf{S}_e\mathbf{g}/n_s + \mathbf{S}^T\mathbf{S}_e\mathbf{g}/n_s + \mathbf{\Sigma}\mathbf{r}\right)$ portion of equation 65 is identical to the $(I_n \otimes \gamma')\mathrm{col}\,[\mathcal{W}'\Xi/T + \mathcal{W}'V/T - \mathbf{\Sigma}_{\mathcal{W}\mathcal{V}}]$ term of equation 4.1 of Schneeweiß. From here one follows Schneeweiß's development to verify the fact. □

Fact 4 *As* $n_s \to \infty$, $\mathrm{MSE}(\tilde{\mathbf{p}}) \to (1/n_s)\mathbf{M}_X^{-1}(\mathbf{\Sigma}\mathbf{r}\mathbf{r}^T\mathbf{\Sigma} + \sigma_e^2\mathbf{M}_X)\mathbf{M}_X^{-1} + \mathbf{M}_X^{-1}\mathbf{\Sigma}\mathbf{r}\mathbf{r}^T\mathbf{\Sigma}\mathbf{M}_X^{-1}$.

This fact is a direct extension of Fact 3.

$$
\begin{aligned}
\mathrm{MSE}(\tilde{\mathbf{p}}) &= \mathcal{E}\left((\mathbf{p} - \tilde{\mathbf{r}})(\tilde{\mathbf{p}} - \mathbf{r})^T\right) && (70)\\
&= \mathcal{E}\left((\tilde{\mathbf{p}} - \mathbf{r} - \mathbf{M}_X^{-1}\mathbf{\Sigma}\mathbf{r} + \mathbf{M}_X^{-1}\mathbf{\Sigma}\mathbf{r})(\tilde{\mathbf{p}} - \mathbf{r} - \mathbf{M}_X^{-1}\mathbf{\Sigma}\mathbf{r} + \mathbf{M}_X^{-1}\mathbf{\Sigma}\mathbf{r})^T\right) && (71)
\end{aligned}
$$

$$= \frac{1}{n_s}\mathcal{E}\left(n_s(\tilde{\mathbf{p}} - \mathbf{r} + \mathbf{M}_X^{-1}\mathbf{r})(\tilde{\mathbf{p}} - \mathbf{r} + \mathbf{M}_X^{-1}\mathbf{r})^T\right) - \mathbf{M}_X^{-1}\boldsymbol{\Sigma}\mathbf{r}\frac{1}{\sqrt{n_s}}\mathcal{E}\left(\sqrt{n_s}(\tilde{\mathbf{p}} + \mathbf{r} - \mathbf{M}_X^{-1}\mathbf{r})^T\right)$$
$$-\frac{1}{\sqrt{n_s}}\mathcal{E}\left(\sqrt{n_s}(\tilde{\mathbf{p}} - \mathbf{r} + \mathbf{M}_X^{-1}\boldsymbol{\Sigma}\mathbf{r})\right)\mathbf{r}^T\boldsymbol{\Sigma}\mathbf{M}_X^{-1T} + \mathbf{M}_X^{-1}\boldsymbol{\Sigma}\mathbf{r}\mathbf{r}^T\boldsymbol{\Sigma}\mathbf{M}_X^{-1T}. \tag{72}$$

As $n_s \to \infty$,

$$\mathcal{E}\left(n_s(\tilde{\mathbf{p}} - \mathbf{r} + \mathbf{M}_X^{-1}\boldsymbol{\Sigma}\mathbf{r})(\tilde{\mathbf{p}} - \mathbf{r} + \mathbf{M}_X^{-1}\boldsymbol{\Sigma}\mathbf{r})^T\right) \to \mathbf{M}_X^{-1}(\boldsymbol{\Sigma}\mathbf{r}\mathbf{r}^T\boldsymbol{\Sigma} + \sigma_e^2\mathbf{M}_X)\mathbf{M}_X^{-1} \tag{73}$$

and

$$\mathcal{E}\left(\sqrt{n_s}(\tilde{\mathbf{p}} - \mathbf{r} + \mathbf{M}_X^{-1}\mathbf{r})\right) \to 0 \tag{74}$$

by the same argument used in Fact 3, so the fact holds. □

Fact 5 *Both* $\tilde{\mathbf{p}}$ *and* $\tilde{\mathbf{r}}$ *are scale invariant.*

Define $\mathbf{X}_D = \mathbf{X}\mathbf{D}$ for any non-singular $\mathbf{D}$; $\mathbf{D}$ is the "scaling." Then

$$\tilde{\mathbf{p}}(\mathbf{X}_D, \mathbf{y}) = (\mathbf{D}^T\mathbf{X}^T\mathbf{X}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{X}^T\mathbf{y} = \mathbf{D}^{-1}\tilde{\mathbf{p}}(\mathbf{X}, \mathbf{y}) \text{ and} \tag{75}$$
$$\tilde{\mathbf{r}}(\mathbf{X}_D, \mathbf{y}) = (\mathbf{D}^T\mathbf{X}^T\mathbf{X}\mathbf{D} - n_s\mathbf{D}^T\boldsymbol{\Sigma}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{X}^T\mathbf{y} = \mathbf{D}^{-1}\tilde{\mathbf{r}}(\mathbf{X}, \mathbf{y}). \tag{76}$$

Thus both $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{r}}$ are scale invariant. □

Recall that for any two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimensions

$$\mathbf{A} \geq \mathbf{B} \Leftrightarrow \mathbf{D}\mathbf{A}\mathbf{D} \geq \mathbf{D}\mathbf{B}\mathbf{D} \tag{77}$$

for any non-singular symmetric $\mathbf{D}$ where "$\mathbf{A} \geq \mathbf{B}$" means $\mathbf{A} - \mathbf{B}$ is a positive semi-definite matrix. Choose $\mathbf{D} = \boldsymbol{\Sigma}^{-\frac{1}{2}}$. Then the rescaled problem has $\boldsymbol{\Sigma}_D = \mathbf{I}$. Under this scaling, as $n_s \to \infty$

$$\mathrm{MSE}(\tilde{\mathbf{r}}) \to \frac{1}{n_s}\mathbf{M}_T^{-1}(\mathbf{r}\mathbf{r}^T + \sigma_e^2\mathbf{M}_X^{-1})\mathbf{M}_T^{-1}, \text{ and} \tag{78}$$
$$\mathrm{MSE}(\tilde{\mathbf{p}}) \to \frac{1}{n_s}\mathbf{M}_X^{-1}(\mathbf{r}\mathbf{r}^T + \sigma_e^2\mathbf{M}_X)\mathbf{M}_X^{-1} + \mathbf{M}_X^{-1}\mathbf{r}\mathbf{r}^T\mathbf{M}_X^{-1}. \tag{79}$$

Thus for n_s large,

$$\mathrm{MSE}(\tilde{\mathbf{r}}) - \mathrm{MSE}(\tilde{\mathbf{p}}) \to \frac{1}{n_s}\left(\mathbf{M}_T^{-1}(\mathbf{r}\mathbf{r}^T + \sigma_e^2\mathbf{M}_X^{-1})\mathbf{M}_T^{-1} - \right. \tag{80}$$
$$\left.\mathbf{M}_X^{-1}(\mathbf{r}\mathbf{r}^T + \sigma_e^2\mathbf{M}_X^{-1})\mathbf{M}_X^{-1}\right) - \mathbf{M}_X^{-1}\mathbf{r}\mathbf{r}^T\mathbf{M}_X^{-1}$$
$$= \frac{1}{n_s}\mathbf{M}_T^{-1}\mathbf{r}\mathbf{r}^T\mathbf{M}_T^{-1} - (1 + \frac{1}{n_s})\mathbf{M}_X^{-1}\mathbf{r}\mathbf{r}^T\mathbf{M}_X^{-1} \tag{81}$$
$$+\frac{\sigma_e^2}{n_s}\left(\mathbf{M}_T^{-1}\mathbf{M}_X\mathbf{M}_T^{-1} - \mathbf{M}_X^{-1}\right).$$

Fact 6 *For large but finite n_s* $\mathrm{MSE}(\tilde{\mathbf{r}}) \geq \mathrm{MSE}(\tilde{\mathbf{p}})$ *if and only if the right-hand side equation 81 is positive semi-definite.*

This fact is a restatement of the algebra developed immediately above. □

One should remember that equation 81 was derived assuming that the data had been rescaled such that $\mathbf{\Sigma} = \mathbf{I}$. The third term (**w**ith the σ_e^2/n_s co-efficient) of equation 81 is always positive definite. However the sum of first two terms is an indefinite matrix: these terms may sum to the null matrix, to a rank one matrix with either a positive or a negative eigenvalue, or to a rank two matrix with one positive and one negative eigenvalue. The positive definite third term may or may not overcome the negative eigenvalue, depending on the values of the parameters; therefore the right hand side (RHS) of equation 81 may or may not be positive semi-definite.

Fact 7 $\mathrm{MSE}(\tilde{\mathbf{r}}) \geq \mathrm{MSE}(\tilde{\mathbf{p}})$ *for any large but given n_s if* $\sigma_e^2 \geq (n_s+1)\sum_{i=1}^{n_i}(\rho_i^2\lambda_i^2) \,/\, ((\lambda_i+1)(2\lambda_i+1))$, *where $\mathbf{v}_i$ are the eigenvalues of* $\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{M}_T\mathbf{\Sigma}^{-\frac{1}{2}}$ *with corresponding eigenvalues $\lambda_i > 0$ and $\rho_i = \mathbf{v}_i^T\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{r}$.*

A sufficient condition for the RHS of equation 81 is to positive semi-definite is

$$\frac{\sigma_e^2}{n_s}\left(\mathbf{M}_T^{-1}\mathbf{M}_X\mathbf{M}_T^{-1} - \mathbf{M}_X^{-1}\right) \geq (1+\frac{1}{n_s})\mathbf{M}_X^{-1}\mathbf{r}\mathbf{r}^T\mathbf{M}_X^{-1} \tag{82}$$

when the data have been rescaled such that $\mathbf{\Sigma} = \mathbf{I}$. Since the LHS of equation 82 is positive definite, Farebrother's 1976 result (Theorem 2.5.2 of [Gruber, 1990]) reveals that equation 82 holds iff

$$\frac{\sigma_e^2}{n_s+1} \geq \mathbf{r}^T\mathbf{M}_X^{-1}\left(\mathbf{M}_T^{-1}\mathbf{M}_X\mathbf{M}_T^{-1} - \mathbf{M}_X^{-1}\right)^{-1}\mathbf{M}_X^{-1}\mathbf{r}. \tag{83}$$

Since the $\mathbf{v}_i$ are also the eigenvalues of $\mathbf{M}_X$ with corresponding eigenvalues $\lambda_i + 1$, we may diagonalize the matrices and substitute $\mathbf{M}_T = \mathbf{\Lambda}$ and $\mathbf{M}_X = \mathbf{\Lambda} + \mathbf{I}$ where $\mathbf{\Lambda} = \mathbf{Diag}(\lambda_1, \ldots, \lambda_{n_i})$. Then

$$\frac{\sigma_e^2}{n_s+1} \geq \begin{bmatrix} \rho_1 \\ \vdots \\ \rho_{n_i} \end{bmatrix}^T (\mathbf{\Lambda}+\mathbf{I})^{-1}\left(\mathbf{\Lambda}^{-1}(\mathbf{\Lambda}+\mathbf{I})\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda}+\mathbf{I})^{-1}\right)^{-1}(\mathbf{\Lambda}+\mathbf{I})^{-1}\begin{bmatrix} \rho_1 \\ \vdots \\ \rho_{n_i} \end{bmatrix} \tag{84}$$

$$\frac{\sigma_e^2}{n_s+1} \geq \begin{bmatrix} \rho_1 \\ \vdots \\ \rho_{n_i} \end{bmatrix}^T (\mathbf{\Lambda}+\mathbf{I})^{-1}\left(\mathbf{Diag}\left(\frac{2\lambda_i+1}{(\lambda_i+1)\lambda_i^2}\right)\right)^{-1}(\mathbf{\Lambda}+\mathbf{I})^{-1}\begin{bmatrix} \rho_1 \\ \vdots \\ \rho_{n_i} \end{bmatrix} \tag{85}$$

$$\frac{\sigma_e^2}{n_s+1} \geq \sum_{i=1}^{n_i} \frac{\rho_i^2 \lambda_i^2}{(\lambda_i+1)(2\lambda_i+1)}. \tag{86}$$

Notice that inequality 86 will hold when λ_i is "small" for any direction for which ρ_i "large," or , equivalently, when $\mathbf{r}^T \mathbf{M}_T \mathbf{M}_T \mathbf{r}$ is "small" for properly scaled data ($D = \mathbf{\Sigma}^{-\frac{1}{2}}$). For unscaled data this condition is that $\mathbf{r}^T \mathbf{M}_T \mathbf{\Sigma}^{-1} \mathbf{M}_T \mathbf{r}$ be "small." □

C Generation of Data for Simulation Examples

The simulations were conducted using Matlab [Moler et al., 1990]. The Matlab functions used to generate the data are described below. The parameters used with these routines were: `n_train = 30`, `n_test = 100`, `d = 10`, `d_ind = 3`, `max_exp = 5`, `min_exp = 0`, `o_noise = 0.2`, `x_noise_max = 1`, and `x_noise_min = 0.1`. The distribution for the $\mathbf{T}$ was specifically constructed to favor $\mathbf{r}^T \mathbf{M}_T \mathbf{\Sigma}^{-1} \mathbf{M}_T \mathbf{r}$ being "small." Also, the distribution for $\mathbf{T}$ is symmetric about the origin. The data are generated calling the data generation routine,

```
% use data such that null pred has rms 1
    [X,y,Xt,yt,b] = gen_dat2(n_train,n_test,d,1,d_ind,max_exp,min_exp,o_noise);

% corrupt the explanatory data with   noise
    rand('normal');

    W = diag(scaled_rand(x_noise_max,x_noise_min,d));

    rand('normal');

    X       = X +   rand(n_train,d)*W;
    Xt      = Xt+   rand(n_test,d)*W;
    W       = W*W;
```

The "true" regression vectors ($\mathbf{r}$) 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 $\mathbf{X}$ are chosen independently of the $\mathbf{r}$ and the singular values (the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$) also vary over 5 orders of magnitude. Thus, there will be large variances in the $\mathbf{X}$ data which do not lie in any of the directions of the columns of $\mathbf{R}$ and therefore have little effect on the dependent variables. 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 explanatory variables vary independently of all other explanatory 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.

C.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 explanatory variables
%  o        is the number of dependent variables
%  d_ind    is the number of explanatory variables 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 data and
%            generating the regression vector
%  noise    std deviation of the normal additive noise
```

```
%
%
%  X          is the explanatory training data
%  Xt         is the explanatory testing data
%  y          is the dependent (noise corrupted) training data
%  yt         is the dependent (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;
```

C.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';
```