Construction Concepts for Continuum Regression

Dan J. Spitzner

Revised: August 28, 2004

Abstract

Approaches for meaningful regressor construction in the linear prediction problem are investigated in a framework similar to partial least squares and continuum regression, but weighted to allow for intelligent specification of an evaluative scheme. A cross-validatory continuum regression procedure is proposed, and shown to compare well with ordinary continuum regression in empirical demonstrations. Similar procedures are formulated from model-based constructive criteria, but are shown to be severely limited in their potential to enhance predictive performance. By paying careful attention to the interpretability of the proposed methods, the paper addresses a long-standing criticism that the current methodology relies on arbitrary mechanisms.

KEY WORDS: linear prediction; principal components regression; partial least squares regression; continuum regression; weighted cross-validation

1 Introduction

Continuum regression (CR; Stone and Brooks, 1990) and its antecedents, partial least squares (PLS; Wold, 1975) and principal components regression (PCR; Massy, 1965), were developed

Dan Spitzner is Assistant Professor, Department of Statistics (0439), Virginia Tech, Blacksburg, VA 24061, USA. The author thanks Prof. Richard L. Smith for his comments and guidance, and an anonymous reviewer for his or her comments and an elementary proof of Lemma B.3

for linear prediction problems as regressor construction techniques aiming to enhance predictive performance. They have been used widely in scientific applications, especially in chemometrics, where the number of explanatory variables is often large, but also in general regression applications involving colinear data. Stone and Brooks (1990) argue for CR's particular relevance to "elastic" science, falling somewhere between "hardened" science, which involves a true, believable model, and "soft" science, which allows and sometimes relies on *ad hoc* data manipulation. On an empirical level, the methods have performed satisfactorily in many such scenarios, and they remain popular analysis tools. We will refer to them collectively as regularization by dimension reduction (RDR) methods.

The RDR setup was first proposed algorithmically, motivated largely on heuristic arguments loosely connected with standard statistical thinking. More recently, the statistical properties of established RDR procedures have been studied in Helland (1988), Frank and Friedman (1993), Sundberg (1993), and Björkström and Sundberg (1996, 1999), among others. Comparing CR to ridge regression (RR; Hoerl and Kennard, 1970), Sundberg (1993) suggests CR may serve as a means of stabilizing colinearity without subscribing to "shrinkage in principle," referring to the decision-theoretic objective of reducing mean squared error, which is often associated with RR.

Despite established theory for some aspects of CR and PLS, and empirical evidence of their efficacy, some important criticisms have fueled controversy over their use in practice. Questions remain over the statistical justification for the constructive mechanisms used in CR and PLS, leading many to view the methodology as a form of pseudo-statistics that works by "maximizing some arbitrary criterion" (Fearn, 1990; see also Brown, 1993). The initial impetus for this investigation was to explore alternative RDR formulations that are justified statistically either through: (*i.*) matching the criterion used to construct regressors with that used to evaluate predictive performance; or (*ii.*) constructing regressors *via* a formal model, such as the Bayesian model of Frank and Friedman (1993). In doing so, approach (*i.*) leads to a working RDR technique, which uses cross-validation in a novel way to form a "cross-validatory continuum." Its empirical performance compares well with existing RDR methods, and, in our

view, is more closely tied to the problem of prediction than are CR and PLS. The technique makes a substantial contribution to the methodology by admitting weighted predictive criteria, which permits analyses to carry forward according to the non-arbitrary specifications of the analyst. Taking approach (*ii.*), it is shown that the most common model-based mechanisms, when adapted to the RDR problem, are hampered by a type of singularity that severely limits their ability to enhance predictive performance. This phenomenon is understood by connecting RDR to RR through its constructive mechanisms, which is a different sort of connection than has been noted (and exploited) by other authors.

The paper is organized as follows. The linear prediction problem and framework for RDR is laid out in the remainder of this section. RDR by constructive cross-validation is described in Section 2, and demonstrated on existing data sets. Model-based RDR is described and investigated in Section 3. Further discussion and conclusions appear in Section 4.

1.1 Linear prediction

The linear prediction problem seeks to build a formula

$$\hat{Y}_0 \quad = \quad \bar{Y} + (x_0 - \bar{x})^T \hat{\beta}, \tag{1}$$

for predicting a univariate response Y_0 on the basis of a given p-dimensional vector of explanatory measurements x_0. The formula is to be derived from an observed vector of univariate responses, $Y = [Y_1, \ldots, Y_n]^T$ and associated vectors of explanatory measurements, which are collected into the $n \times p$ design matrix $X = [x_1, \ldots, x_n]^T$. The means $\bar{Y} = n^{-1} \sum Y_i$ and $\bar{x} = n^{-1} \sum x_i$ set baseline levels and $\hat{\beta}$, the driving object in (1), determines the manner in which Y_0 is explained by x_0. We will assume an initial preprocessing step of centering and scaling, by which $\bar{Y} = 0$, $\bar{x} = 0$, and the columns of X have a common sum of squares. As in many applications, $p > n$ is allowed, but we will assume the rank of X is $p_0 = \min(n - 1, p)$.

The vector $\hat{\beta}$ is to be calculated from the form

$$\hat{\beta} \quad = \quad R_m (R_m^T V R_m)^{-1} R_m^T X^T Y, \tag{2}$$

where $V = X^T X$ and R_m is some $p \times m$ matrix with linearly independent columns. One should recognize $\hat{\beta}$ as R_m times the least squares coefficients for Y regressed on a "constructed" regressor matrix XR_m. We will refer to R_m as a "regression components matrix," and the integer m as the "dimension" of the prediction formula. Helland (1988) deduces (2) explicitly from the original algorithmic definition of PLS.

It is helpful the think of R_m as the first m columns of a larger $p \times p_0$ regression components matrix $R = [R_m \; R_{-m}]$. The first step in RDR is to construct the full matrix R, after which cross-validation is applied to select R_m. This perspective implies the restriction $m \leq p_0$.

1.2 Regressor construction

In CR, R is constructed column-by-column, with the m'th column calculated recursively as the r_m which maximizes

$$CR_{m,\alpha} \quad = \quad (r_m^T X^T Y)^2 (r_m^T V r_m)^{\alpha/(1-\alpha)-1}, \tag{3}$$

subject to the normalization $\|r_m\| = 1$ and the orthogonality constraint $r_j^T V r_m = 0$ for $j = 1, ..., m-1$. In the full CR procedure, the index α is allowed to take any value in $0 \leq \alpha < 1$, and is selected data-dependently by cross-validation. At fixed values, $\alpha = 0$ defines ordinary least squares (OLS), $\alpha = 1/2$ defines PLS, and $\hat{\beta}$ tends to a PCR estimator as α tends to 1.

Frank and Friedman (1993), Sundberg (1993), and Björkström and Sundberg (1996, 1999) observe that at $m = 1$ the regression component maximizing (3) takes the form $r_1 \propto (V + \delta I)^{-1} X^T Y$, the core expression of a ridge regression estimator, where δ acts as a sort of ridge constant. Investigating this further, Björkström and Sundberg (1999) identify from among criteria similar to (3) that maximize a function of $r_m^T X^T Y$ and $r_m^T V r_m$ a wide subclass whose members each lead to solutions resembling RR and intersect with OLS, PCR, and PLS. This leads to "least squares ridge regression," which, as with CR, derives from a criterion in this subclass.

1.3 Evaluative cross-validation

A role of cross-validation in RDR is to fix the dimension m, and in CR to select the value α. For present purposes, such determinations will be made through the weighted cross-validation diagnostic

$$CV \quad = \quad n^{-1} \sum_{i,i'=1}^{n} \{Y_i - \hat{Y}_{-i}\}[G]_{ii'}\{Y_{i'} - \hat{Y}_{-i'}\}, \tag{4}$$

where $\hat{Y}_{-i}$ is the "deleted prediction" of Y_i calculated from (1) but with $\hat{\beta}$, $\bar{Y}$, and $\bar{x}$ derived without reference to Y_i and x_i. The values $[G]_{ii'}$ are entries of a (fixed) symmetric non-negative definite $n \times n$ weight matrix G. It is conventional to standardize CV as a "cross-validatory index," $\rho = 1 - CV/CV_0$ where CV_0 is (4) with deleted predictions calculated as $\hat{Y}_{-i} = (Y_1 + \cdots + Y_{i-1} + Y_{i+1} + \cdots + Y_n)/(n-1)$. The cross-validatory index cannot exceed 1, but can be negative, and larger values indicate better predictive performance.

In most applications, and throughout the literature on RDR, G is set to the $n \times n$ identity matrix, in which case (4) is said to be unweighted. Indeed, there is no compelling reason why the criterion (3) would better suit CR for one choice of G over another. From a strictly algorithmic point of view, any choice is arbitrary, and it suffices to set $G = I$ simply out of convenience. However, from a statistical point of view, the choice of G reflects a preference toward certain predictions over others, and the inability of (3) to adapt to weighted versions of cross validation highlights its arbitrariness and inflexibility. Our view is that some meaning is afforded to RDR procedures when the criterion used to construct R reflects an intelligent choice of G, which (3) does not.

We offer the following paradigm for specifying G in practice. First, note that the cross-validation diagnostic (4) is closely related to the quadratic loss functions used in decision-theoretic point estimation,

$$L^Q(\beta, \hat{\beta}) \quad = \quad \{\beta - \hat{\beta}\}^T Q^T Q \{\beta - \hat{\beta}\}, \tag{5}$$

where the $s \times p$ matrix $Q = [q_1, \ldots, q_s]^T$ identifies a predetermined set of response points at which it is especially important to make good predictions. The vector β is an unknown parameter of a

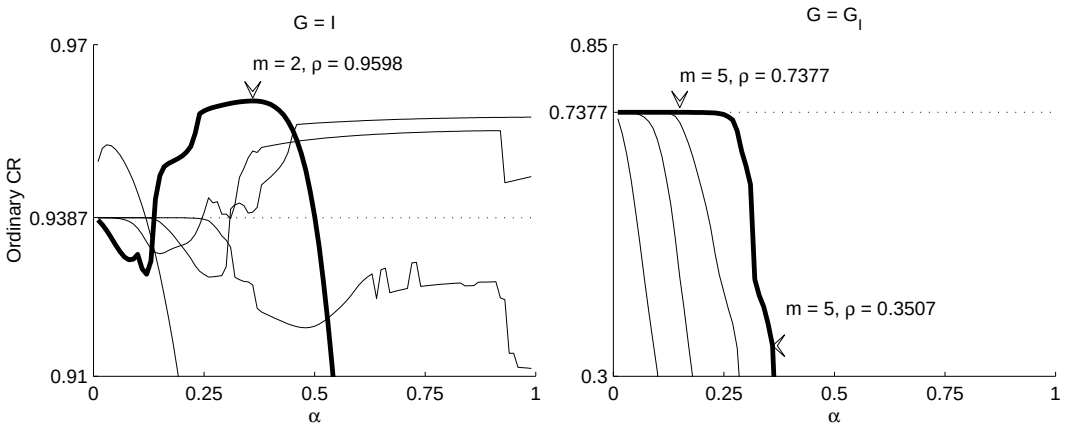

Figure 1: *Continuum regression for the infra-red calibration of protein data. Plots of ρ for each $m < p$ as α varies. The dotted line locates OLS, and the thick lines highlight the m which maximize ρ.*

hypothetical probability model $Y = X\beta + \epsilon$, in which ϵ represents a vector of mean-zero errors. Plugging the model into (4) and supposing each $\hat{Y}_{-i} \approx x_i^T \hat{\beta}$ and the entries of ϵ are negligible on average, it is seen that

$$nCV \quad \approx \quad \{\beta - \hat{\beta}\}^T X^T G X \{\beta - \hat{\beta}\} + const. \tag{6}$$

Comparing (5) and (6) it is seen that G may be treated through quadratic loss weighted by $Q^T Q = X^T G X$, and it is immediately obvious that $Q = X$ agrees with the default setting $G = I$, which matches with conventional intuition. For other Q, a suitable rule is to set

$$G_Q \quad = \quad I + X V^- \{Q^T Q - X^T X\} V^- X^T, \tag{7}$$

where V^- is the Moore-Penrose inverse of V. Whenever V is invertible or $Q = QV^-V$, expression (7) both solves $Q^T Q = X^T G_Q X$ and leads to $G_X = I$.

References to weighted cross-validation appears sparsely in the literature, but it has been used, for example, in spatial analyses involving unequally spaced measurements (Militino and Ugarte, 2001). To our knowledge, (7) and our proposal to use it in a general paradigm for targeting preferred response points is novel. Its derivation is straightforward after representing X by a singular value decomposition.

Continuum regression can be extremely sensitive to the analyst's choice of G. For illustration, let us revisit Stone and Brook's (1990) Example 3, which demonstrates CR on a data set having

$n = 12$ observations and $p = 6$ explanatory variables. Further details are given in Section 2.2.1. Stone and Brooks's Figure 3 is reproduced in the left panel of the current Figure 1, in which the unweighted ($G = I$) cross-validatory index is plotted at each $m = 1, \ldots, 5$ as α varies. The maximum cross-validatory index of $\rho = 0.9598$ is achieved at $m = 2$ and $\alpha = 0.36$. When cross-validation is weighted by $G = G_I$ the picture changes radically, as shown in the right panel of the figure. Here, at $\alpha = 0.36$ the maximum cross-validatory index for $m < p$ drops to $\rho = 0.3507$ for $m = 5$. The maximum cross-validatory index of $\rho = 0.7377$ is achieved at $m = 5$ and $\alpha = 0.15$, and is barely larger than that of ordinary least squares.

1.4 Construction and evaluation

We have drawn a distinction between two types of objective criteria used in RDR: *constructive* criteria, such as (3), which guide the construction of R, and *evaluative* criteria, such as (4), which evaluate a procedure's overall predictive performance. The controversy over RDR is directed more at its constructive aspect than its evaluative one. On the evaluative side, cross-validatory assessment is intrinsic to problems of prediction, and widely accepted in general. (See *e.g.,* Stone, 1974, for an in-depth discussion.) The constructive criterion (3), or the more general criterion of Björkström and Sundberg (1999), on the other hand, is sometimes defended on the basis that it balances covariance, $r_m^T X^T Y$, with variance, $r_m^T V r_m$, but it is unclear exactly why this is meaningful for prediction. This leaves vague the meaning and implications of established RDR techniques' approach to construction.

2 A cross-validatory continuum

As our first proposal to firm up the foundations of RDR, we seek to extend the interpretability of cross-validatory assessment to construction, but do so in a way that retains the spirit of techniques like CR. Our approach is to replace (3) with a continuum of cross-validation diagnostics.

2.1 Constructive cross-validation

To formulate a cross-validatory continuum, write $V_m^{-1} = R_m(R_m^T V R_m)^{-1} R_m^T$ for $m \geq 1$, $V_0^{-1} = 0$, and define the deleted prediction errors

$$d_{m,i} \quad = \quad Y_i - \hat{Y}_{-i} = \frac{Y_i - x_i^T V_m^{-1} X^T Y}{1 - x_i^T V_m^{-1} x_i}, \tag{8}$$

where now $\hat{Y}_{-i}$ is the deleted prediction of Y_i through (1) with R_m held fixed, and data-recentering omitted. The right-hand expression follows from Lemma B.2 in the appendices. By holding R_m fixed, this perspective treats the regression components as model-like parameters independent of the data actually measured, which distinguishes these constructive versions of $\hat{Y}_{-i}$ from those involved in the evaluative criterion (4). Our cross-validatory continuum consists of the weighted L_γ norms of the $d_{m,i}$ over the positive γ, weighted by the same G used in the evaluative criterion (4). For its definition, write G in its diagonalized form as $U_g G_0 U_g^T$ for some $n \times n$ orthonormal matrix $U_g = [u_{g,1}, \ldots, u_{g,n}]$ and diagonal matrix $G_0 = \mathrm{diag}(g_1, \ldots, g_n)$. The weighted L_γ norm of the deleted prediction errors is then

$$CV_{m,\gamma} \quad = \quad \left\{ \sum_{i=1}^{n} g_i^{\gamma/2} |u_{g,i}^T d_m|^\gamma \right\}^{1/\gamma}, \tag{9}$$

where $d_m = [d_{m,1}, \ldots, d_{m,n}]^T$ and $\gamma > 0$. (It is technically a pseudo-norm for $\gamma < 1$.) Notice that $CV_{m,\gamma}$ becomes $aCV_{m,\gamma}$ when G is rescaled by a^2.

Although (9) matches the principle underlying regressor construction with that of evaluation, defining it on a continuum does permit inexact matches with (4), and this may seem to taint our procedure with a certain arbitrariness. Such criticism is justified, and it is prudent to pay special attention at $\gamma = 2$, in which case (9) reduces to a weighted version of the PRESS diagnostic (Allen, 1974),

$$WPRESS_m \quad = \quad Y^T \{I - X V_m^{-1} X^T\} D_{CV} G D_{CV} \{I - X V_m^{-1} X^T\} Y, \tag{10}$$

where D_{CV} is a diagonal matrix with i'th diagonal entry $(1 - x_i^T V_m^{-1} x_i)^{-1}$.

Ideally, one would want to construct R so that each r_m, minimizes $CV_{m,\gamma}$ given R_{m-1}, but computational limitations (*e.g.*, numerous local minima) force us to settle for an approximate

optimization. We require $R_m^T V R_m = I$, which is equivalent to the restrictions adopted in CR, although the normalization constraint is different. The core subroutine which calculates r_m given R_{m-1} is carried out multiple steps, as follows:

STEP 1: Obtain candidate components $r_{m,1}, \ldots, r_{m,n}$, for which $r_{m,i}$ minimizes the i'th deleted prediction error in absolute value, $|d_{m,i}|$. These all have exact analytical solutions, described below.

STEP 2: Evaluate $CV_{m,\gamma}$ at each $r_{m,i}$ to produce the scores $v_{m,1}, \ldots, v_{m,n}$. From these, calculate an additional candidate by weighted averaging,

$$\bar{r}_m = \textstyle\sum_i v_{m,i}^{-1} r_{m,i} / \sum_i v_{m,i}^{-1}$$

STEP 3: Select r_m from among $r_{m,1}, \ldots, r_{m,n}$ and $\bar{r}_m$ as that candidate achieving the lowest value of $CV_{m,\gamma}$.

The candidate vectors $r_{m,1}, \ldots, r_{m,n}$ calculated in STEP 1 are not unique as minimizers of $|d_{m,i}|$. The preferred set of minimizers is that which leads the averaged candidate $\bar{r}_m$ to the greatest decrease in $CV_{m,\gamma}$. We translate this to a requirement where $r_{m,1}, \ldots, r_{m,n}$ are to be chosen as homogeneous as possible, postulating that they would then cluster around a true global minimum, which $\bar{r}_m$ would then have an improved chance of closely approximating. To this end, each individual $r_{m,i}$ is taken to be the minimizer of $|d_{m,i}|$ which forms the smallest angle with the fixed vector $X^T Y$.

The exact solutions for STEP 1 are now described, using a certain parameterization that greatly simplifies the problem. Let $\tilde{R}$ be any $p \times (p-m+1)$ matrix satisfying $\tilde{R}^T V \tilde{R} = I$ and $R_{m-1}^T V \tilde{R} = 0$. Any candidate can then be parameterized as $r_{m,i} = \tilde{R} z_i$ for some $(p-m+1) \times 1$ vector z_i satisfying $z_i^T z_i = 1$. Now write

$$a_i = Y_i - x_i^T V_{m-1}^{-1} X^T Y, \quad b_i = 1 - x_i^T V_{m-1}^{-1} x_i, \quad c_1 = \|\tilde{R} X^T Y\|,$$
$$u_1 = \tilde{R} X^T Y / c_1, \qquad v_{i,1} = u_1^T \tilde{R} x_i, \qquad v_{i,2} = \|\tilde{R} x_i - v_{i,1} u_1\|,$$

and if $v_{i,2} > 0$ set $u_{i,2} = (\tilde{R} x_i - v_{i,1} u_1)/v_{i,2}$. Note that $\|\tilde{R} x_i\|^2 = v_{i,1}^2 + v_{i,2}^2$. Now write $z_i = z_{i,1} u_1 + z_{i,2} u_{i,2}$, where $z_{i,1} = z_i^T u_1$ and $z_{i,2} = z_i^T u_{i,2}$. Using Lemma B.2 in the appendix,

this parameterization permits the deleted prediction error (8), calculated at any z_i, to be written

$$d_{m,i} = \frac{a_i - c_1(z_{i,1}^2 v_{i,1} + z_{i,2} v_{i,2})}{b_i - (z_{i,1} v_{i,1} + z_{i,2} v_{i,2})^2}. \tag{11}$$

Exact closed-form solutions for the preferred $z_{i,1}$ and $z_{i,2}$ minimizing (11) are described separately for the cases $v_{i,2} \neq 0$ and $v_{i,2} = 0$. Their technical derivation appears in in the appendix. To avoid pathological cases, which are also handled in the appendix, assume b_i is neither 0, nor $\|\tilde{R}x_i\|^2$.

Case 1, $v_{i,2} \neq 0$: Whenever

$$\left[av_{i,1}/c_1 + \tfrac{1}{2}v_{i,2}^2\right]^2 \geq (av_{i,1}/c_1)^2 + v_{i,2}^2, \tag{12}$$

a z_i exists for which $|d_{m,i}| = 0$. In this case, set

$$\theta^2 = \frac{1}{\|\tilde{R}x_i\|^2}\left\{av_{i,1}/c_1 + \tfrac{1}{2}v_{i,2}^2 + \sqrt{\left[av_{i,1}/c_1 + \tfrac{1}{2}v_{i,2}^2\right]^2 - (a/c_1)^2\|\tilde{R}x_i\|^2}\right\}, \tag{13}$$

and

$$z_{i,1} = \sqrt{\theta^2} \quad \text{and} \quad z_{i,2} = \frac{a/c1 - \theta^2 v_{i,1}}{z_{i,1} v_{i,2}}. \tag{14}$$

When (12) does not hold, set the quantity d to whichever of

$$\frac{a_i}{b_i} + \frac{1}{2b_i(b_i - \|\tilde{R}x_i\|^2)}\left\{a_i\|\tilde{R}x_i\|^2 - b_i c_1 v_{i,1} \pm \|\tilde{R}x_i\|\sqrt{(b_i c_1)^2 + a_i^2\|\tilde{R}x_i\|^2 - b_i c_1\left(2a_i v_{i,1} + c_1 v_{i,2}^2\right)}\right\} \tag{15}$$

is closest to zero. In this case, $|d|$ is the smallest possible value of $|d_{m,i}|$, and $db_i - a_i$ will equal one of

$$\frac{1}{2}\left[d\|\tilde{R}x_i\|^2 - c_1 v_{i,1} \pm \sqrt{\left[d\|\tilde{R}x_i\|^2 - c_1 v_{i,1}\right]^2 + (c_1 v_{i,2})^2}\right]. \tag{16}$$

Set

$$z_{i,1} = \frac{1}{\xi}\left\{\frac{d(v_{i,1}^2 - v_{i,2}^2) + c_1 v_{i,1} \pm \sqrt{\left[d\|\tilde{R}x_i\|^2 - c_1 v_{i,1}\right]^2 + (c_1 v_{i,2})^2}}{(2dv_{i,1} - c_1)v_{i,2}}\right\}, \tag{17}$$

using the "+" version when, and only when, $db_i - a_i$ matches the larger of (16); set $z_{i,2} = 1/\xi$, where ξ is such that $z_{i,1}^2 + z_{i,2}^2 = 1$.

Case 2, $v_{i,2} \neq 0$: Whenever

$$0 \leq av_{i,1}/c_1 \leq 1 \tag{18}$$

set $z_{i,1} = \sqrt{av_{i,1}/c_1}$ so that $|d_{m,i}| = 0$. When (18) does not hold, set d to whichever of

$$\frac{a_i}{b_i} \quad \text{or} \quad \frac{a_i - c_1 v_{i,1}}{b_i - v_{i,1}^2} \tag{19}$$

is closest to zero. If $d = a_i/b_i$, set $z_{i,1} = 0$; otherwise set $z_{i,1} = 1$.

Note in this case it is possible for $z_{i,1}^2 < 1$, which requires that $z_i = z_{i,1}u_1 + z_{i,3}u_{i,3}$ for some normalized vector $u_{i,3}$ orthogonal to u_1 and $z_{i,3} = \pm\sqrt{1 - z_{i,1}^2}$. The preferred solution is to set $u_{i,3}$ and $z_{i,3}$ in a manner that leaves the resulting $r_{m,1}, \ldots, r_{m,n}$ tightly clustered. Our subroutine handles this as follows: Among the preferred minimizers, z_{i^*}, of the $|d_{m,i^*}|$ having $v_{i^*,2} \neq 0$, find the index $i_0^* = i^*$ for which $|z_{i^*,1} - z_{i,1}|$ is smallest. Set $u_{i,3} = u_{i_0^*,2}$ and $z_{i,3} = \text{sign}(z_{i_0^*,2})\sqrt{1 - z_{i,1}^2}$.

2.2 Empirical demonstrations

We illustrate our cross-validatory procedure on three example data sets examined in Stone and Brooks (1990), beginning with example introduced at the end of Section 1.

2.2.1 Near infra-red calibration for protein

The data of Example 3 in Stone and Brooks are compiled from the infra-red spectrometer measurements of flour proteins in Table 1 of Fearn (1983). Only the first $n = 12$ rows are used. The $p = 6$ explanatory variables, $L_1, \ldots, L_6$, measure log(1/reflectance) at six wavelengths, and the associated response variable Y measures protein percentage. Following Stone and Brooks, we preprocess so that the (i, j) entries of X are $x_{i6} = \bar{L} = (L_{i1} + \cdots + L_{i6})/6)$ and $x_{ij} = L_{ij} - \bar{L}$ for $j = 1, ..., 5$, then center and scale.

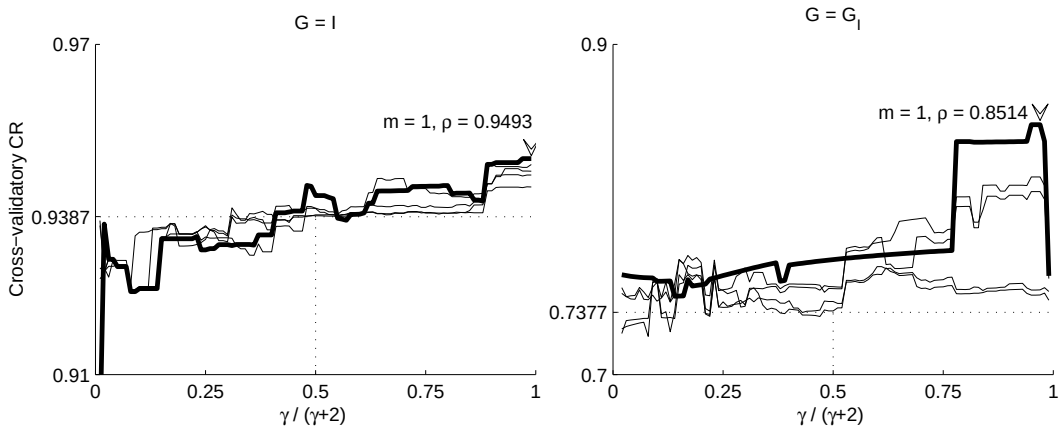

Figure 2: *Cross-validatory continuum regression for the infra-red calibration of protein data. Plots of ρ for each $m < p$ as α varies. The dotted lines locate OLS and $\gamma = 2$, and the thick lines highlight the m which maximize ρ.*

Recall Figure 1, which indicates the maximum cross-validatory index achieved by ordinary CR is $\rho = 0.9598$ when $G = I$ and $\rho = 0.7377$ when $G = G_I$. The ρ values for cross-validatory CR are shown in Figure 2, plotted for each $m < p$ as $\gamma/(\gamma + 2)$ varies between 0 and 1. Note that $\gamma/(\gamma + 2) = 0, 1/2$ and 1 correspond to $\gamma = 0, 2$ and ∞. For $G = I$, the maximum index of $\rho = 0.9493$ is achieved at $m = 1$ and $\gamma/(\gamma + 2) = 0.99$. For $G = G_I$, the maximum index of $\rho = 0.8514$ is achieved at $m = 1$ and $\gamma/(\gamma + 2) = 0.97$. For these data, it is seen that the the performance of our new procedure is comparable to that of ordinary CR for $G = I$, but leads to a vastly increased value of ρ for $G = G_I$.

With $\gamma = 2$ fixed, for $G = I$, the maximum index of $\rho = 0.9425$ is achieved at $m = 1$. For $G = G_I$, the maximum index of $\rho = 0.7697$ is achieved at $m = 1$. In both configurations, the ρ achieved by cross-validatory CR exceeds that of OLS, and for $G = G_I$ also that of ordinary CR.

2.2.2 Cement heat evolution data

Our second example is Example 1 of Stone and Brooks, analyzing data from Table 20.2 of Hald (1952), which measures the heat evolved in $n = 13$ cement samples in the 180 days after water was added. The response, Y, measures heat evolved in calories per gram, and we have $p = 4$ explanatory variables, each measuring the estimated percentage by weight of a specific

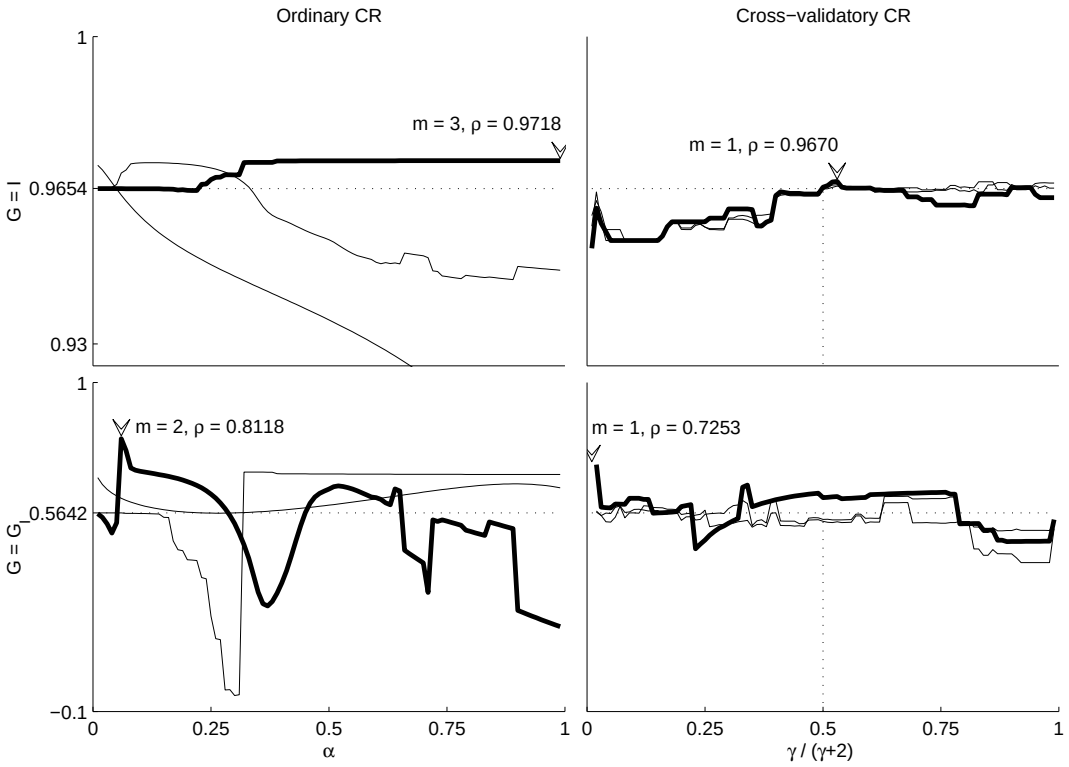

Figure 3: *Ordinary CR and cross-validatory CR for the heat evolution data. Plots of ρ for each $m < p$ as α varies. The dotted lines locate OLS and $\gamma = 2$, and the thick lines highlight the m which maximize ρ.*

compound in the cement.

Figure 3 plots ρ for ordinary CR and cross-validatory CR. For ordinary CR with $G = I$ (upper-left panel), the level $\rho = 0.9718$ is attained for $m = 3$ at $\alpha \approx 0.41$ and remains there as α increases, indicating PCR leads to the optimal predictive performance on this continuum. Setting $G = G_I$ (lower-left panel), the maximum index of $\rho = 0.8118$ is achieved at $m = 2$ and $\alpha = 0.06$. For cross-validatory CR with $G = I$ (upper-right panel), the maximum index of $\rho = 0.9670$ is achieved at $m = 1$ and $\gamma/(\gamma + 2) = 0.53$. With $G = G_I$ (lower-right panel), the maximum index of $\rho = 0.7253$ is achieved at $m = 1$ and $\gamma/(\gamma + 2) = 0.01$. For these data, we see that the maximum ρ of our new procedure is comparable, but less than that of ordinary CR. It is nevertheless able to enhance predictive performance over OLS.

With $\gamma = 2$ fixed, for $G = I$, the maximum index of $\rho = 0.9657$ is achieved at $m = 1$. For $G = G_I$, the maximum index of $\rho = 0.7253$ is achieved at $m = 1$. Both of these values exceed those of OLS, but not ordinary CR.

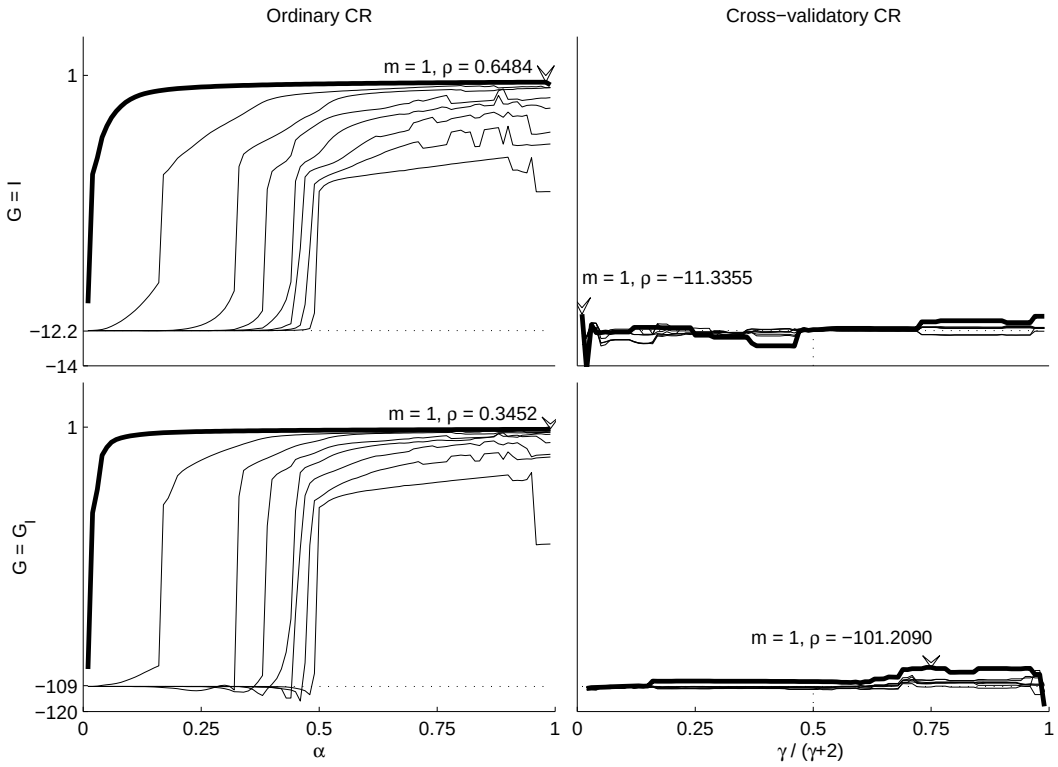

Figure 4: *Ordinary CR and cross-validatory CR for the road accident data. Plots of ρ for each $m < p$ as α varies. The dotted lines locate OLS and $\gamma = 2$, and the thick lines highlight the m which maximize ρ.*

For situations where inexact matches are allowed between the weights of evaluation and construction, let us note that when $G = G_I$ in (4) but $G = I$ in (9), cross-validatory CR achieves a maximum $\rho = 0.8527$ at $m = 1$ and $\gamma/(\gamma + 2) = 0.35$, which exceeds that of ordinary CR.

2.2.3 Road accident data

We now apply cross-validatory CR to Stone and Brooks's "thoroughly soft" Example 2, reanalyzing the data from Table 8.1 of Weisberg (1980). The response, Y, measures 1973 accident rates along $n = 13$ stretches of "minor arterial highways" in Minnesota, and just the first $p = 9$ explanatory variables are used.

Figure 4 plots ρ for ordinary CR and cross-validatory CR. In this example, the cross-validatory indices for ordinary least squares is extremely low, at $\rho = -12.1775$ for $G = I$ and $\rho = -109.3379$ for $G = G_I$. Applying ordinary CR (left panels), ρ increases roughly in α

at each $m < p$, tending at $m = 1$ to maximum values of $\rho = 0.6484$ for $G = I$ and $\rho = 0.3452$ for $G = G_I$ as α approaches 1. As in the previous example, PCR appears best along this continuum. The ρ values for cross-validatory CR (right panels) remain far below that of PCR, hovering around the values attained by ordinary least squares. With $G = I$ the maximum index of $\rho = -11.3353$ is achieved at $m = 1$ and $\gamma/(\gamma + 2) = 0.01$. With $G = G_I$, the maximum index of $\rho = -101.2090$ is achieved at $m = 1$ and $\gamma/(\gamma + 2) = 0.75$. The new procedure does not increase predictive performance over PCR, but does provide an enhancement over OLS.

With $\gamma = 2$ fixed, for $G = I$, the maximum index of $\rho = -12.1112$ is achieved at $m = 4$. For $G = G_I$, the maximum index of $\rho = -107.2434$ is achieved at $m = 1$. Both of these values exceed those of OLS, but not ordinary CR.

3 Model-based regularization

In this section, we consider model-based criteria for constructing R in a meaningful way, building upon a model similar to the Bayesian formulation described in Frank and Friedman (1993).

3.1 Models for dimension reduction

Estimators of the form (2) arise as solutions to maximum penalized likelihood estimation under the the linear regression model $Y = X\beta + \epsilon$ with the error vector ϵ following a mean-zero normal distribution with covariance matrix $\sigma^2 I$. Specifically, the log-likelihood for β is $LogL(\beta; Y) = -\|Y - X\beta\|^2/(2\sigma^2) - (n/2)\log 2\pi\sigma^2$, and if S_m is any $p \times m$ matrix satisfying $S_m^T R_m = I$, an appropriate penalized log-likelihood is

$$PLogL_m(\beta; Y) = LogL(R_m S_m^T \beta; Y) - \frac{\tau}{2\sigma^2}\beta^T S_m R_m^T A R_m S_m^T \beta, \tag{20}$$

where τ is a nonnegative scalar and A is symmetric nonnegative-definite $p \times p$ matrix. (Technically, to form a true log-likelihood, $LogL$ should be constrained to reflect that the data are centered, but this will not effect our results.) The β maximizing $PLogL_m(\beta; Y)$ is

$$\tilde{\beta}_\tau \quad = \quad R_m(R_m^T\{V + \tau A\}R_m)^{-1}R_m^T X^T Y, \tag{21}$$

which tends to $\hat{\beta}$ in (2) as τ tends to zero. Alternately, a Bayesian perspective motivates (21) as the posterior mean under a mean-zero multivariate normal prior for β with covariance matrix $\tau^{-1}\sigma^2 A_m^{-1}$, where $A_m^{-1} = R_m(R_m A R_m)^{-1} R_m^T$ for $m \geq 1$, and $A_0^{-1} = 0$.

When $\tau > 0$, the matrix A may be seen as a weighting parameter that may be specified to match the analyst's choice of G in (4). For instance, the results of Strawderman (1978) and Stein (1981), working entirely in an RR setting, suggest that the setting $A = V(Q^T Q)^{-1} V$ induces shrinkage properties favorable to the specific Q in the quadratic loss function (5). This suggests that for $G = I$, which matches $Q = X$ through (7), an appropriate setting is $A = V$.

When $n < p$, V is not of full rank and it is necessary to focus on a lower-dimensional core model. Here, let us write X according to its singular value decomposition $X = U_0 \Lambda U_1^T$, where U_0 and U_1 are respectively $n \times n$ and $n \times p$ orthonormal matrices, and Λ is an $n \times n$ diagonal matrix with non-zero diagonal entries. An equivalent rotated model is now $U_0^T Y = \Lambda\alpha + \epsilon^*$, with $\alpha = U_1^T \beta$, and ϵ^* a mean-zero normal random error vector with covariance matrix $\sigma^2 U_0^T U_0 = \sigma^2 I$. An equivalent rotated Bayesian prior, now placed on α, has covariance matrix $\tau^{-1}\sigma^2 U_1^T A_m^{-1} U_1$. Under the restriction $A = U_1 U_1^T A U_1 U_1^T$, we may write $\tilde{A}_m^{-1} = U_1^T A_m^{-1} U_1 = \tilde{R}_m(\tilde{R}_m \tilde{A} \tilde{R}_m)^{-1} \tilde{R}_m^T$, where $\tilde{A} = U_1^T A U_1$ and $\tilde{R}_m = U_1^T R_m$. An analog to (21) is now

$$\begin{aligned} \tilde{\alpha}_\tau &= \tilde{R}_m(\tilde{R}_m\{\Lambda^T\Lambda + \tau\tilde{A}\}\tilde{R}_m)^{-1}\tilde{R}_m^T X^T Y \qquad (22) \\ &= U_1^T R_m(R_m^T\{V + \tau A\}R_m)^{-1} R_m^T X^T Y. \end{aligned}$$

Observing that $\tilde{\alpha}_\tau$ is simply a rotation of $\tilde{\beta}_\tau$, it is clear that we may proceed from either model $Y = X\beta + \epsilon$ or $U_0^T Y = \Lambda\alpha + \epsilon^*$, provided we impose the restriction $A = U_1 U_1^T A U_1 U_1^T$. Thus, to simplify our exposition, we shall always assume the former, with V, possibly acting in place of $\Lambda^T\Lambda$, always invertible.

The discussion above suggests, for any n and p, a preference for constructive criteria that are invariant to orthogonal rotations of the data. In other words, if Y and X were replaced by $U_0^T Y$ and $U_0^T X$, for some $n \times n$ orthonormal matrix, it is preferred that the manner in which R_m is assessed would remain unaffected. Basic cross-validation diagnostics like (9) do not have

this property, but all of the criteria we consider in this section do.

3.2 Model-based constructive criteria

One advantage of the model-based perspective is that it establishes a viewpoint where R_m is a "parameter" to be "estimated," and in doing so opens up an array of meaningful alternatives to the constructive criterion (3). These are model-based "fit-diagnostics," each assessing R_m according to a particular estimation principle. To retain generality, we describe them below leaving τ arbitrary, and for convenience write $C = V + \tau A$ and $C_m^{-1} = R_m(R_m^T C R_m)^{-1} R_m^T$ for $m \geq 1$, $C_0^{-1} = 0$, so that $\tilde{\beta}_\tau = C_m^{-1} X^T Y$. Our orthogonality constraint on the regression components is also modified as $R^T C R = 1$.

- *Generalized cross-validation:* A rotation-invariant version of (10), with V_m^{-1} replaced by C_m^{-1}, is the generalized cross-validation diagnostic

$$WGCV_m \quad = \quad \frac{Y^T\{I - XC_m^{-1}X^T\}G\{I - XC_m^{-1}X^T\}Y}{[\frac{1}{n}Trace\ \{I - XC_m^{-1}X^T\}]^2}, \tag{23}$$

  originally proposed in Golub *et al.* (1979) in an unweighted form. In the weighted form above, if Y and X were replaced by $U_0^T Y$ and $U_0^T X$, one would also need to replace G by $U_0^T G U_0$. In a constructive algorithm, r_m given R_{m-1} would be chosen to minimize $WGCV_m$.

- *Mean squared error diagnostics:* For a decision-theoretic perspective, R_m may be assessed through an unbiased estimate of $E[L^Q(\beta, \tilde{\beta})]$,

$$\begin{aligned} WMSE_m \quad &= \quad Y^T X V^{-1}\{I - VC_m^{-1}\}Q^T Q\{I - C_m^{-1}V\}V^{-1}X^T Y \\ &\qquad +\sigma^2 Trace\{2Q^T Q C_m^{-1} - Q^T Q V^{-1}\}. \end{aligned} \tag{24}$$

  In a constructive algorithm, r_m given R_{m-1} would be chosen to minimize $WMSE_m$.

- *Significance diagnostics:* Analogous to an F statistic for testing the significance of a model component, the logarithm of a penalized likelihood ratio is

$$PLogL_m - PLogL_{m-1} = \frac{1}{2\sigma^2} Y^T X\{C_m^{-1} - C_{m-1}^{-1}\}X^T Y,$$

which similarly assesses the contribution of r_m, given R_{m-1}. By Lemma B.2 in the appendices, this is equivalent to

$$\hat{F}_m = \frac{(r_m^T\{I - CC_{m-1}^{-1}\}X^TY)^2}{r_m^T\{I - CC_{m-1}^{-1}\}Cr_m}. \tag{25}$$

In a constructive algorithm, r_m given R_{m-1} would be chosen to maximize $\hat{F}_m$.

- *Bayes factors:* A Bayesian analogue of the F statistic is a Bayes factor (*cf.* Smith and Spiegelhalter, 1980), which is derived from the marginal distribution of $\tilde{\beta}_\tau$. With

$$\begin{aligned} LogL_m^{II} = & -\frac{1}{2\sigma^2}\left\{Y^TXV^{-1}\{V^{-1} + \tau^{-1}A_m^{-1}\}^{-1}V^{-1}X^TY\right. \\ & \left. +\sigma^2\log[\det\{V^{-1} + \tau^{-1}A_m^{-1}\}]\right\} - \frac{p_0}{2}\log 2\pi\sigma^2, \end{aligned} \tag{26}$$

a Bayes factor is derived as

$$BF_m = Log_m^{II} - Log_{m-1}^{II}. \tag{27}$$

In a constructive algorithm, r_m given R_{m-1} would be chosen to maximize BF_m.

- *Information criteria:* The last set of diagnostics we will consider are the information criteria,

$$IC_m^\xi = n\log\left(\frac{Y^T\{I - XC_m^{-1}X^T\}Y}{n}\right) + \xi(R_m), \tag{28}$$

where $\xi(R_m)$ is penalty function. Among its most common variations, "Akaike's Information Criterion" (AIC) (Akaike, 1974) uses the penalty function $\xi(R_m) = 2m$, and the "Bayesian information criterion" (BIC) (Schwartz, 1978) uses $\xi(R_m) = m\log n$. In a constructive algorithm, r_m given R_{m-1} would be chosen to minimize IC_m^ξ.

3.3 Diagnostic breakdown

Examining (21), note that $\beta = \tilde{\beta}_\tau$ solves the equations $C\beta = X^TY$ whenever R_m is such that $X^TY = CR_ma$ for a suitable $m \times 1$ vector a. In other words, whenever $C^{-1}X^TY$ is in the span of the columns of R_m, the RDR estimator $\tilde{\beta}_\tau$ reduces to the ridge regresion estimator $C^{-1}X^TY$.

When this phenomenon occurs, the machinery of RDR has "broken down" and the possibility of any enhancement in predictive performance disappears.

This is different type of connection between RDR and RR than has been noticed by other authors. In Section 1.2 it was noted that the CR criterion (with $\tau = 0$) leads to $r_1 \propto (V + \delta I)^{-1} X^T Y$ for some δ. Consequently, $\tilde{\beta}_\tau$ is proportional to a ridge regression estimator when $m = 1$. The present observation is that, for any known or unknown value of τ, including $\tau = 0$, and for any m, whenever the constructive criterion leads $C^{-1} X^T Y$ to fall in the span of R_m, $\tilde{\beta}_\tau$ becomes identically the ridge regression estimator associated with τ and A.

The following discussion will demonstrate that such breakdown is pervasive across the array of standard model-based diagnostics outlined above. It is associated specifically with weight settings consistent with $G = I$, in which $Q^T Q = V$ and $A = V$. Because $G = I$ is usually the default setting in analyses concerned with predictive performance, our results seriously limit the extent to which model-based diagnostics are suitable for the constructive aspect of RDR.

The mathematical development begins with the simple, but fundamental result:

Lemma 3.1 *For arbitrary $\tau \geq 0$, $\hat{F}_1$ is maximized when r_1 solves $C r_1 \propto X^T Y$. Thereafter, for $m > 1$, $\hat{F}_m = 0$. RDR based on maximizing $\hat{F}_m$ would therefore lead to breakdown.*

PROOF: At $m = 1$, $C_{m-1}^{-1} = 0$, and the Cauchy-Schwartz inequality implies that (25) constrained by $r_1 C r_1 = 1$ is maximized when $C r_1 \propto X^T Y$. Thereafter, for $m > 1$, $C C_{m-1}^{-1} X^T Y = X^T Y$, implying $\hat{F}_m = 0$. *Q.E.D*

Extension of Lemma 3.1 to the remaining diagnostics follows by connecting each one to $\hat{F}_m$. To this end, define the quadratic forms

$$Q_m^{GCV}(G) = Y^T \{I - X C_m^{-1} X^T\} G \{I - X C_m^{-1} X^T\} Y, \tag{29}$$

$$Q_m^{MSE}(Q) = \hat{\beta}^T \{I - V C_m^{-1}\} Q^T Q \{I - C_m^{-1} V\} \hat{\beta} \tag{30}$$

$$Q_m^{IC} = Y^T \{I - X C_m^{-1} X^T\} Y, \tag{31}$$

$$Q_m^{II} = Y^T X \{V^{-1} - C_m^{-1}\} X^T Y. \tag{32}$$

along with the quantities

$$\hat{D}_m(G) = \frac{r_m^T\{I - CC_{m-1}^{-1}\}X^TG\{I - XC_{m-1}^{-1}X^T\}Y}{r_m^T\{I - CC_{m-1}^{-1}\}X^TY},$$

$$E_m(G) = \frac{r_m^T\{I - CC_{m-1}^{-1}\}X^TGX\{I - C_{m-1}^{-1}C\}r_m}{r_m^T\{I - CC_{m-1}^{-1}\}C\{I - C_{m-1}^{-1}C\}r_m},$$

and let H_Q be defined through Q as $H_Q = XV^{-1}Q^TQV^{-1}X^T$. Direct application of Lemma B.2 in the appendix implies the following recursive formulas:

$$Q_m^{GCV}(G) = Q_{m-1}^{GCV}(G) - \hat{F}_{m-1}\{2\hat{D}_m(G) - E_m(G)\}, \tag{33}$$

$$Q_m^{MSE}(Q) = Q_{m-1}^{MSE}(Q) - \hat{F}_{m-1}\{2\hat{D}_m(H_Q) - E_m(H_Q)\} \tag{34}$$

$$Q_m^{IC} = Q_{m-1}^{IC} - \hat{F}_{m-1} \tag{35}$$

$$\hat{Q}_m^{II} = \hat{Q}_{m-1}^{II} - \hat{F}_{m-1}. \tag{36}$$

Several corollaries of Lemma 3.1 are now easily deduced:

Corollary 3.1.1 *Suppose the penalty function ξ is a function of m only, $\xi(R_m) = \xi(m)$. For arbitrary $\tau \geq 0$, IC_1^ξ is minimized when r_1 solves $Cr_1 \propto X^TY$. Thereafter, for $m > 1$, IC_m^ξ is constant. RDR based on maximizing IC_m^ξ would therefore lead to breakdown.*

PROOF: The information criteria may be written $IC_m^\xi = n\log(Q_m^{IC}/n) + \xi(R_m)$, a monotone function of Q_m^{IC} plus a function of m only. Hence, by (35) and Lemma 3.1, IC_1^ξ minimized when r_1 solves $Cr_1 \propto X^TY$. Thereafter $\hat{F}_m$ will be zero and, also by (35), $IC_m^\xi = IC_1^\xi$. *Q.E.D*

Corollary 3.1.2 *Set $G = I$, $Q = X$, and $A = V$. For arbitrary $\tau \geq 0$, both $WGCV_1$ and $WMSE_1$ are minimized when r_1 solves $Cr_1 \propto X^TY$. Thereafter, for $m > 1$, $WGCV_m$ and $WMSE_m$ are constant. RDR based on maximizing either $WGCV_m$ or $WMSE_m$ would therefore lead to breakdown.*

PROOF: Note that

$$\hat{D}_m(G) = \hat{D}_m(XV^{-1}X^TG)$$

$$E_m(G) = E_m(XV^{-1}X^TG) = E_m(GXV^{-1}X^T),$$

so $\hat{D}_m(I) = \hat{D}_m(H_X)$ and $E_m(I) = E_m(H_X)$. These relationships along with the identity $V_m^{-1}VV_m^{-1} = V_m^{-1}$ imply

$$\hat{D}_m(I) = \hat{D}_m(H_X) = 1 \quad \text{and} \quad E_m(I) = E_m(H_X) = \frac{1}{1+\tau}. \tag{37}$$

Under the stated conditions, the diagnostics may be written

$$\begin{aligned} \log WGCV_m &= \log Q_m^{GCV}(I) - 2\log\left(n - \frac{m}{1+\tau}\right), \\ WMSE_m &= Q_m^{MSE}(V) + 2\frac{m}{1+\tau} - \min(n,p). \end{aligned}$$

Hence each is an increasing function of Q_m^{GCV} or Q_m^{MSE}, respectively, plus a function of m only. By (37), (33) and (34) the quadratic forms simplify as

$$\begin{aligned} Q_m^{GCV}(I) &= Q_{m-1}^{GCV}(I) - b\hat{F}_{m-1}, \\ Q_m^{MSE}(V) &= Q_{m-1}^{MSE}(V) - b\hat{F}_{m-1}, \end{aligned}$$

where $b = 2 - (1+\tau)^{-1}$. The result then follows from a similar deduction as that of Corollary 3.1.1. *Q.E.D*

Finally, Lemma 3.1 extends to the Bayes factors through

Corollary 3.1.3 *Set $A = V$. For arbitrary $\tau > 0$, BF_1 is maximized when r_1 solves $Cr_1 \propto X^TY$. Thereafter, for $m > 1$, BF_m is constant. RDR based on maximizing BF_m would therefore lead to breakdown.*

PROOF: By Lemma B.1 in the appendix, one has the identity

$$(V^{-1} + \tau^{-1}A_m^{-1})^{-1} = \{I - VC_m^{-1}\}V, \tag{38}$$

and by this $-2LogL_m^{II}$ may be written

$$\frac{1}{\sigma^2}\hat{Q}_m^{II} + \xi_{II}(R_m) + p_0 \log 2\pi\sigma^2,$$

where $\xi_{II}(R_m) = \log\left[\det(V^{-1} + \tau^{-1}A_m^{-1})\right]$. Corollary B.3.1 in the appendix shows that when $A = V$, $\xi_{II}(R_m)$ is in fact a function of m only. Therefore, by (36) and Lemma 3.1, the result follows as in Corollary 3.1.1. *Q.E.D*

4 Conclusions and discussion

Cross-validatory continuum regression has been favorably demonstrated in several respects: (*i.*) Its flexible construction mechanism can be adjusted to match the weights of a non-arbitrary evaluative criterion. (*ii.*) It inherits the interpretability of cross-validatory assessment and direct relevance to problems of prediction. (*iii.*) It often leads to predictive performance comparable to that of ordinary CR, and can, in some cases, lead to substantially improved performance. For fixed $\gamma = 2$, it was seen in all of our empirical demonstrations to provide an enhancement over OLS. (*iv.*) Its core constructive subroutine relies on direct calculation rather than iterative searches, permitting quick execution.

Regarding future research, we speculate the candidate regression components $r_{m,1}, \ldots, r_{m,n}$ calculated in our constructive algorithm might be used to identify "predictive outliers," *i.e.*, responses that are especially difficult to cross-validate. We also anticipate some further computational efficiency is possible by integrating the core subroutine with the rest of the algorithm.

Our investigation of model-based construction revealed a new connection between RDR and RR through the "breakdown" phenomenon. Although the model based-approaches benefit in terms of interpretability from the presence of a model, the pervasiveness of breakdown in configurations associated with $G = I$ suggests a serious limitation to their potential for practical benefit. Our investigation demonstrates that breakdown is excluded from neither frequentist nor Bayesian lines of thought, but it does not close this line of inquiry. For instance, it is conceivable the model-based diagnostics would lead to enhanced performance in configurations associated with $G = G_I$ or others. Moreover, the possible Bayesian formulations of RDR are far from exhausted. Let us note in particular that a normal mixture prior placed on β would set up a Bayesian model averaging model, which admits a suitable construction mechanism in the form of a hierarchical prior placed on τ and R, and computational method based standard Markov Chain Monte Carlo routines. Such a setup is mentioned in George and McCulloch (1993).

REFERENCES

Allen, D. M. (1974), The relationship between variable selection and data augmentation and a method for prediction, *Technometrics*, 16:125–127.

Björkström, A., and Sundberg, R. (1996), Continuum regression is not always continuous, *Journal of the Royal Statistical Society B,* 58:703–710

Björkström, A., and Sundberg, R. (1999), A generalized view on continuum regression, *Scandinavian Journal of Statistics,* 26:17–30.

Brown, P. J. (1993), *Measurements, Regression, and Calibration,* Oxford University Press.

Frank, I. E., and Friedman, J. H. (1993), A statistical view of some chemometrics regression tools (with discussion), *Technometrics,* 35(2):109–148.

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

Fearn, T. (1990). Discussion of the paper by Stone and Brooks, *Journal of the Royal Statistical Society B,* 52(2):260–261.

George, E. I., and McCulloch, R. E. (1993), Variable selection via Gibbs sampling, *Journal of the American Statistical Association,* 88(423):881–889.

Golub, G. H., Heath, M., and Wahba, G. (1979), Generalized cross-validaion as a method for choosing a good ridge parameter, *Technometrics,* 21(2):215–223.

Hald, A. (1952), *Statistical Theory with Engineering Applications,* Wiley.

Helland, I. S. (1988), On the structure of partial least squares regression, *Communications in Statistics: Simulation and Computation*, 17:581–607.

Hoerl, A. E., and Kennard, R. W. (1970), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12(1):55–67.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis,* Academic Press.

Massy, W. F. (1965), Principal components regression in exploratory statistical research, *Journal of the American Statistical Association,* 60:234–246.

Militino, A. F., and Ugarte, M. D. (2001), Assessing the covariance function in geostatistics, *Statistics and Probability Letters,* 52:199–206.

Schwartz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, 6:461–464.

Smith, A. F. M., and Spiegelhalter (1980), Bayes factors and choice criteria for linear models, *Journal of the Royal Statistical Society B,* 42:213–220

Stein, C. (1981), Estimation of the mean of a multivariate normal distribution, *Annals of Statistics* 9:1135–1151.

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

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(2):237–269; corrigendum, 54(3):906–907.

Strawderman, W. E. (1978), Minimax adaptive generalized ridge regression estimators, *Journal of the American Statistical Association,* 73(363):623–627.

Sundberg, R. (1993), Continuum regression and ridge regression, *Journal of the Royal Statistical Society B,* 55(3):653–659.

Weisberg, S. (1980), *Applied Linear Regression,* Wiley.

Wold, H. (1975), Soft modeling by latent variables, In: *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett* (J. Gani, ed.), Academic Press.

A Technical results for Section 2

Lemma A.1 *Under the conditions specified in Section 2 and* $v_{i,2} \neq 0$*, solutions (14,17) minimize (11) over all* $z_{i,1}, z_{i,2}$ *subject to* $z_{i,1}^2 + z_{i,2}^2 = \eta^2$ *with* $0 \leq \eta^2 \leq 1$*, and among the possible minimizers of (11) form the smallest angle with* $\tilde{R}X^TY$*.*

PROOF: Note the z_i forming the smallest angle with $\tilde{R}X^TY$ are those having $z_{i,1}$ as large as possible.

The condition $d_{m,i} = 0$ holds if, and only if, $z_{i,2} = (a_i/c_1 - z_{i,1}^2 v_{i,1})/(z_{i,1} v_{i,2})$, and with $z_{i,1}^2 + z_{i,2}^2 = \eta^2$ this is equivalent to

$$\begin{aligned} z_{i,1}^2 &= \frac{1}{\|\tilde{R}x_i\|^2} \left\{ \left[a v_{i,1}/c_1 + \tfrac{1}{2} v_{i,2}^2 \eta^2 \right] \right. \\ &\qquad \left. \pm \sqrt{\left[a v_{i,1}/c_1 + \tfrac{1}{2} v_{i,2}^2 \eta^2 \right]^2 - (a/c_1)^2 \|\tilde{R}x_i\|^2} \right\}. \end{aligned} \tag{39}$$

The "+" solution always leads to $z_{i,1}^2 \geq 0$ since $-1 \leq t/\sqrt{t^2 - b_0} \leq 0$ whenever $t \leq 0$ and $t^2 \geq b_0$ for any b_0. In (14), $z_{i,1}$ is set to the positive square root of the "+" solution to make it as large as possible. Moreover, the derivative of the "+" solution with respect to η^2 is

$$\frac{d}{d\eta^2} z_{i,1}^2 = \frac{v_{i,2}^2}{2\|\tilde{R}x_i\|^2} \left\{ 1 + \frac{a v_{i,1}/c_1 + \tfrac{1}{2} v_{i,2}^2 \eta^2}{\sqrt{\left[a v_{i,1}/c_1 + \tfrac{1}{2} v_{i,2}^2 \eta^2 \right]^2 - (a/c_1)^2 \|\tilde{R}x_i\|^2}} \right\},$$

which is positive since $-1 \leq t/\sqrt{t^2 - b_0} \leq 1$ for all t with $t^2 \geq b_0$. The preferred solution therefore has $\eta^2 = 1$, which leads to (14).

When (12) does not hold, rewrite (11) as $\zeta^T \Omega \zeta = d_{m,i} b_i - a_i$, where $\zeta = [z_{i,1}, z_{i,2}]^T$ and

$$\Omega = \begin{bmatrix} v_{i,1}(d_{m,i} v_{i,1} - c_1) & v_{i,2}(d_{m,i} v_{i,1} - c_1/2) \\ v_{i,2}(d_{m,i} v_{i,1} - c_1/2) & d_{m,i} v_{i,2}^2 \end{bmatrix} \tag{40}$$

The quantities in (17) are the eigenvalues of Ω, which we shall label ω_1 for the smaller and ω_2 for the larger. One can therefore find a rotation $\tilde{\zeta} = [\tilde{z}_{i,1}, \tilde{z}_{i,2}]^T$ of ζ for which $\omega_1 \tilde{z}_{i,1}^2 + \omega_2 \tilde{z}_{i,2}^2$ and $\tilde{z}_{i,1}^2 + \tilde{z}_{i,2}^2 = \eta^2$. The solutions have

$$\tilde{z}_{i,1}^2 = \frac{d_{m,i} b_i - a_i - \omega_2 \eta^2}{\omega_1 - \omega_2} \quad \text{and} \quad \tilde{z}_{i,2}^2 = -\frac{d_{m,i} b_i - a_i - \omega_1 \eta^2}{\omega_1 - \omega_2}, \tag{41}$$

and (17) are the inverse rotations for suitable $d_{m,i}$ and $\eta^2 = 1$.

To calculate (41) it is necessary to substitute the $d_{m,i} = d$ that minimizes $|d_{m,i}|$. For $0 \leq \tilde{z}_{i,1}^2 \leq \eta^2$, one must have $\omega_1 \eta^2 \leq d_{m,i} b_i - a_i \leq \omega_2 \eta^2$, or equivalently

$$\begin{aligned} &\frac{b_i (b_i - \|\tilde{R}x_i\|^2 \eta^2)}{\eta^4} \left(d_{m,i} - \frac{a_i (2b_i - \|\tilde{R}x_i\|^2 \eta^2) - b_i c_1 v_{i,1} \eta^2}{2 b_i (b_i - \|\tilde{R}x_i\|^2 \eta^2)} \right)^2 \\ &\qquad \leq \|\tilde{R}x_i\|^2 \frac{(b_i c_1)^2 + a_i^2 \|\tilde{R}x_i\|^2 - b_i c_1 (2 a_i v_{i,1} + c_1 v_{i,2}^2 \eta^2)}{4 b_i (b_i - \|\tilde{R}x_i\|^2 \eta^2)} \end{aligned} \tag{42}$$

The quantities (15) are the roots of this equation, which lead to $\tilde{z}_{i,1}^2 = \eta^2$ or $\tilde{z}_{i,1}^2 = 0$.

We leave the following for the reader to verify: (*i*) Whenever the right side of (42) is negative, so is $b_i(b_i - \|\tilde{R}x_i\|^2\eta^2)$, and therefore the parabola in $d_{m,i}$ on the left side is concave; (*ii.*) As is implied by their derivatives with respect to η^2, both roots in (15) have a pole at $\eta^2 = b_i/\|\tilde{R}x_i\|^2$ and are monotonic in the ranges $\eta^2 < b_i/\|\tilde{R}x_i\|^2$ and $\eta^2 > b_i/\|\tilde{R}x_i\|^2$; (*iii.*) The roots in (15) straddle the ratio a_i/b_i whenever $b_i(b_i - \|\tilde{R}x_i\|^2\eta^2) > 0$. (The difference of the squares of the two terms in the braces is $c_1 v_{i,2} b_i(b_i - \|\tilde{R}x_i\|^2\eta^2) > 0$.)

Since (12) does not hold, $d_{m,i} = 0$ is impossible, and there are only two possible configurations for (15): either $b_i(b_i - \|\tilde{R}x_i\|^2\eta^2) < 0$ and the roots (15) straddle zero, or $b_i(b_i - \|\tilde{R}x_i\|^2\eta^2) > 0$ and both roots lie entirely to the left or right of zero.

It remains to show that the smallest $|d_{m,i}|$ is achieved at $\eta^2 = 1$. Set $\eta^2 = 1$ and calculate (15). If $b_i(b_i - \|\tilde{R}x_i\|^2) < 0$, the pole at $\eta^2 = b_i/\|\tilde{R}x_i\|^2$ implies any smaller possible $|d_{m,i}|$ leads to the existence of a solution with $d_{m,i} = 0$, which contradicts that (12) does not hold. If $b_i(b_i - \|\tilde{R}x_i\|^2) > 0$, the monotonicity of the derivative implies any smaller possible $|d_{m,i}|$ leads to the preferred solution at $\eta^2 = 0$, for which $d_{m,i} = a_i/b_i$. But this is a contradiction since the roots (15) straddling a_i/b_i imply one of them would be closer to zero at $\eta^2 > 0$. *Q.E.D.*

Lemma A.2 *Under the conditions specified in Section 2 and $v_{i,2} = 0$, the respective solutions $z_{i,1} = \sqrt{av_{i,1}/c_1}$, $z_{i,1} = 0$, or $z_{i,1} = 1$ minimize (11) over all $z_{i,1}$ with $0 \le z_{i,1}^2 \le 1$, and among the possible minimizers of (11) form the smallest angle with $\tilde{R}X^TY$.*

PROOF: Clearly, $d_{m,i} = 0$ is possible whenever $0 \le av_{i,1}/c_1 \le 1$, and $z_{i,1} = \sqrt{av_{i,1}/c_1}$ is the largest of the solutions. Otherwise,

$$z_{i,1} \quad = \quad \sqrt{\frac{d_{m,i}b_i - a_i}{v_{i,1}^2 d_{m,i} - c_1 v_{i,1}}},$$

where $d_{m,i} = d$ is set to the possible value minimizing $|d_{m,i}|$. Working case-by-case for $b_i < 0$, $0 < b_i < v_{i,1}^2$, and $b_i > v_{i,1}^2$, it is not difficult to show that one of the quantities in (19) yields the correct $d_{m,i}$. *Q.E.D.*

The pathological cases of $b_i = 0$ and $b_i = \|\tilde{R}x_i\|^2$ are covered by the next two lemmas. The

proofs are omitted, but are straightforward, following similar arguments as those of the previous two lemmas.

Lemma A.3 *For the case of $v_{i,2} \neq 0$ the solutions (14, 17) minimize (11) over all $z_{i,1}, z_{i,2}$ subject to $z_{i,1}^2 + z_{i,2}^2 = \eta^2$ with $0 \leq \eta^2 \leq 1$, and among the possible minimizers of (11) form the smallest angle with $\tilde{R}X^TY$, provided*

$$d \;=\; \frac{1}{\|\tilde{R}x_i\|^2}\left[\tfrac{1}{4}c_1^2 v_{i,2}^2 \eta^2/a_i - (a_i - c_1 v_{i,1}\eta^2)/\eta^2\right],$$

when $b_i = 0$ or

$$d \;=\; \frac{\frac{1}{4}c_1^2 v_{i,1}^2/\|\tilde{R}x_i\|^2 + a_i c_1 v_{i,1}/b_i - (a_i/b_i)^2\|\tilde{R}x_i\|^2}{(a_i/b_i)\|\tilde{R}x_i\|^2 - c_1 v_{i,1}}$$

when $b_i = \|\tilde{R}x_i\|^2$.

Lemma A.4 *For the case of $v_{i,2} = 0$ the solutions $z_{i,1} = 1$ when $b_i = 0$, or $z_{i,1} = 0$ when $b_i = \|\tilde{R}x_i\|^2$ minimize (11) over all $z_{i,1}$ with $0 \leq z_{i,1}^2 \leq 1$, and among the possible minimizers of (11) form the smallest angle with $\tilde{R}X^TY$.*

B Technical results for Section 3

Lemmas B.1 and B.2 below are standard results from matrix analysis; see *e.g.,* Appendix A of Mardia *et al.* (1979).

Lemma B.1 *For matrices A, U, and V,*

$$(A - UV^T)^{-1} \;=\; A^{-1} + A^{-1}U(I - V^TA^{-1}U)^{-1}V^TA^{-1}$$

whenever the expressions make sense.

Lemma B.2 *For a non-singular matrix A, the partitioning*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \textit{implies} \quad A^{-1} = \begin{bmatrix} B_{11}^{-1} & -A_{11}^{-1}A_{12}B_{22}^{-1} \\ -B_{22}^{-1}A_{21}A_{11}^{-1} & B_{22}^{-1} \end{bmatrix},$$

where $B_{11} = A_{11} - A_{12}A_{22}^{-1}A_{21}$ and $B_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12}$. Moreover,

$$\det A \;=\; (\det A_{11})(\det B_{22}) = (\det A_{22})(\det B_{11}).$$

Lemma B.3 *For B be an invertible matrix, c a constant, and P the projection matrix into an m dimensional subspace, $\mathcal{U}$, of $\mathbb{R}^p$. If $PBP = P$, then*

$$\log\left[\det\left(B^{-1} + c^{-1}\frac{Prr^TP}{r^TPr}\right)^{-1}\right]$$

is constant for any $p \times 1$ vector r.

PROOF: Set

$$H = \begin{bmatrix} B^{-1} & -q \\ q^T & a \end{bmatrix},$$

where $a = cr^TPr$ and $q = Pr$, and note by Lemma B.2 that $\det H$ may be written equivalently as

$$a\det(B^{-1} + a^{-1}qq^T) = \det(B^{-1})(a + q^TBq).$$

Substituting a and q, this is

$$(cr^TPr)\det[B^{-1} + (cr^TPr)^{-1}Prr^TP] =$$

$$\det(B^{-1})(cr^TPr + r^TPBPr) = \det(B^{-1})(cr^TPr + r^TPr),$$

since the projection matrix P is symmetric and $PBP = P$. Dividing each side by cr^TPr gives the desired conclusion. *Q.E.D*

Corollary B.3.1 *When $A = V$, the quantity*

$$\log\left[\det\left(V^{-1} + c^{-1}A_m^{-1}\right)^{-1}\right], \tag{43}$$

depends on R_m only through m.

PROOF: Define $\tilde{P} = I - A^{1/2}A_{m-1}^{-1}A^{1/2}$, $\tilde{V} = A^{-1/2}VA^{-1/2}$, $\tilde{r} = A^{1/2}r_m$, and $\tilde{B} = (\tilde{V}^{-1} + \tau^{-1}(I - \tilde{P}))^{-1}$. Using Lemma B.2, (43) may be written

$$\log[\det A] + \log\left[\det\left(\tilde{B}^{-1} + \tau^{-1}\frac{\tilde{P}\tilde{r}\tilde{r}^T\tilde{P}}{\tilde{r}^T\tilde{P}\tilde{r}}\right)^{-1}\right].$$

Following also from this notation, one has

$$\tilde{P}\tilde{B}\tilde{P} \quad = \quad A^{-1/2}\{I - AA_{m-1}^{-1}\}\left(V^{-1} + \tau^{-1}A_{m-1}^{-1}\right)^{-1}\{I - A_{m-1}^{-1}A\}A^{-1/2}. \tag{44}$$

When $A = V$, the identity (38) reduces this expression to

$$V^{-1/2}\{I - VV_{m-1}^{-1}\}V\{I - V_{m-1}^{-1}V\}V^{-1/2} \quad = \quad I - V^{1/2}V_{m-1}^{-1}V^{1/2} = \tilde{P},$$

demonstrating that $\tilde{P}\tilde{B}\tilde{P} = \tilde{P}$. Lemma B.3 implies the result. *Q.E.D.*