Published in final edited form as:
Electron J Stat. 2012 January 1; 6: 323–353. doi:10.1214/12-EJS676.

Structured penalties for functional linear models—partially empirical eigenvectors for regression

Timothy W. Randolph,
Fred Hutchinson Cancer Research Center, Biostatistics and Biomathematics Program, Seattle, WA 98109

Jaroslaw Harezlak, and
Indiana University School of Medicine, Department of Biostatistics, Indianapolis, IN 46202

Ziding Feng
Fred Hutchinson Cancer Research Center, Biostatistics and Biomathematics Program, Seattle, WA 98109

Timothy W. Randolph: trandolp@fhcrc.org; Jaroslaw Harezlak: harezlak@iupui.edu; Ziding Feng: zfeng@fhcrc.org

Abstract

One of the challenges with functional data is incorporating geometric structure, or local correlation, into the analysis. This structure is inherent in the output from an increasing number of biomedical technologies, and a functional linear model is often used to estimate the relationship between the predictor functions and scalar responses. Common approaches to the problem of estimating a coefficient function typically involve two stages: regularization and estimation. Regularization is usually done via dimension reduction, projecting onto a predefined span of basis functions or a reduced set of eigenvectors (principal components). In contrast, we present a unified approach that directly incorporates geometric structure into the estimation process by exploiting the *joint* eigenproperties of the predictors and a linear penalty operator. In this sense, the components in the regression are 'partially empirical' and the framework is provided by the generalized singular value decomposition (GSVD). The form of the penalized estimation is not new, but the GSVD clarifies the process and informs the choice of penalty by making explicit the joint influence of the penalty and predictors on the bias, variance and performance of the estimated coefficient function. Laboratory spectroscopy data and simulations are used to illustrate the concepts.

Keywords and phrases

Penalized regression; generalized singular value decomposition; regularization; functional data

1. Introduction

The coefficient function, β, in a functional linear model (fLM) represents the linear relationship between responses, y, and predictors, x, either of which may appear as a function. We consider the special case of scalar-on-function regression, formally written as $y = \int_I x(t)\beta(t)\,dt + \varepsilon$, where x is a random function, square integrable on a closed interval $I \subset \mathbb{R}$, and ε a vector of random i.i.d. mean-zero errors. In many instances, one has an approximate idea about the *informative structure* of the predictors, such as the extent to which they are smooth, oscillatory, peaked, etc. Here we focus on analytical framework for incorporating such information into the estimation of β.

The analysis of data in this context involves a set of n responses $\{y_i\}_{i=1}^{n}$ corresponding to a set of predictor curves $\{x_i\}_{i=1}^{n}$, each arising as a discretized sampling of an idealized function; i.e., $x_i \equiv (x_i(t_1), \ldots, x_i(t_p))$, for some, $t_1 < \cdot \cdot \cdot < t_p$, of I. In particular, the concept of geometric or spatial structure implies an order relation among the index parameter values. We assume the predictor functions have been sampled equally and densely enough to capture geometric structure of the type typically attributed to functions in (subspaces of) $L^2(I)$. For this, it will be assumed that $p > n$ although this condition is not necessary for our discussion.

Several methods for estimating β are based on the eigenfunctions associated with the auto-covariance operator defined by the predictors [16, 32]. These eigenfunctions provide an empirical basis for representing the estimate and are the basis for the usual ordinary least-squares and principal-component estimates in multivariate analysis. The book by Ramsay and Silverman [38] summarize a variety of estimation methods that involve some combination of the empirical eigenfunctions and smoothing, using B-splines or other technique, but none of these methods provide an analytically tractable way to incorporate presumed structure directly into the estimation process. The approach presented here achieves this by way of a penalty operator, $\mathcal{L}$, defined on the space of predictor functions.

The joint influence of the penalty and predictors on the estimated coefficient function is made explicit by way of the generalized singular value decomposition (GSVD) for a matrix pair. Just as the ordinary SVD provides the ingredients for an ordinary least squares estimate (in terms of the empirical basis), the GSVD provides a natural way to express a penalized least-squares estimate in terms of a basis derived from both the penalty and the predictors. We describe this in terms of the $n \times p$ matrix of sampled predictors, X, and an $m \times p$ discretized penalty operator, L. The general formulation is familiar as we consider estimates of β that arise from a squared-error loss with quadratic penalty:

$$\tilde{\beta}_{\alpha,L} = \arg\min_{\beta} \{ \|y - X\beta\|^2_{\mathbb{R}^n} + \alpha \|L\beta\|^2_{L^2} \}. \qquad (1)$$

What distinguishes our presentation from others using this formulation is an emphasis on the *joint* spectral properties of the pair (X, L), as arise from the GSVD. We investigate the analytical role played by L in imposing structure on the estimate and focus on how the structure of L's least-dominant singular vectors should be commensurate with the informative structure of β.

In a Bayesian view, one may think of L as implementing a prior that favors a coefficient function lying near a particular subspace; this subspace is determined jointly by X and L. We note, however, that informative priors must come from somewhere and while they may come from expectations regarding smoothness, other information often exists—including pilot data, scientific knowledge or laboratory and instrumental properties. Our presentation aims to elucidate the role of L in providing a flexible means of implementing informative priors, regardless of their origin.

The general concept of incorporating "structural information" into regularized estimation for functional and image data is well established [2, 12, 36]. Methods for penalized regression have adopted this by constraining high-dimensional problems in various "structured" ways (sometimes with use of an L^1 norm): locally-constant structure [49, 46], spatial smoothness [20], correlation-based constraints [52], and network-dependence structure described via a graph [26]. These general penalties have been motivated by a variety of heuristics: Huang et al. [24] refer to the second-difference penalty as an "intuitive choice"; Hastie et al. [20] refer to a “structured penalty matrix [which] imposes smoothness with regard to an underlying space, time or frequency domain”; Tibshirani and Taylor [50] note that the rows of L should “reflect some believed structure or geometry in the signal”; and the penalties of Slawski et al. [46] aim to capture “a priori association structure of the features in more generality than the fused lasso.”

The most common penalty is a (discretized) derivative operator, motivated by the heuristic of penalizing roughness (see [21, 38]). Our perspective on this is more analytical: since the eigenfunctions of the second-derivative operator $\mathcal{L} = \mathcal{D}^2$ (with zero boundary conditions on [0, 1]) are of the form $\phi(t) = \sin(k\pi t)$, with eigenvalues $k^2\pi^2$ ($k = 1, 2, \ldots$), $\mathcal{L}$ implements the assumption that the coefficient function is well represented by low-frequency trigonometric functions. This is in contrast to ridge regression ($L = I$) which imposes no geometric structure. Although not typically viewed this way, the choice of $\mathcal{L} = \mathcal{D}^2$, or any differential operator, implies a favored basis for expansion of the estimate.

A purely empirical basis comprised of a few dominant right singular vectors of X is a common and theoretically tractable choice. This is the essence of principal component regression (PCR) and these vectors also form the basis for a ridge estimate. Although this empirical basis does not technically impose local spatial structure (no order relation among the index parameter values is used), it may be justified by arguing that a few principal component vectors capture the “greatest part” of a set of predictors [17]. Properties of this approach for signal regression is the focus of [7] and [16]. The functional data analysis framework of Ramsay and Silverman [38] provides two formulations of PCR. One in which the predictor curves are themselves smoothed prior to construction of principal components (chap. 8) and another that incorporates a roughness penalty into the construction of principal components (chap. 9), as originally proposed in [45]. In a related presentation on signal regression, Marx and Eilers [30] proposed a penalized B-spline approach in which predictors are transformed using a basis external to the problem (B-splines) and the estimated coefficient function is derived using the transform coefficients. Combining ideas from [30] and [21], the “smooth principal components” method of [8] projects predictors onto the dominant eigenfunctions to obtain an estimate then uses B-splines in a procedure that smooths the estimate. Reiss and Ogden [40] provide a thorough study on several of these methods and propose modifications that include two versions of PCR using B-splines and second-derivative penalties: FPCR$_C$ applies the penalty to the construction of the principal components (cf. [45]), while FPCR$_R$ incorporates the penalty into the regression (cf. [38]).

In the context of nonparametric regression ($X = I$) the formulation (1) plays a dominant role for smoothing [54]. Related to this, Heckman and Ramsay [22] proposed a differential equations model-based estimate of a function μ whose properties are determined by a linear differential operator chosen from a parameterized family of differential equations, $L\mu = 0$. In this context, however, the GSVD is irrelevant since X does not appear and the role of L is relatively transparent.

Algebraic details on the GSVD as it relates to penalized least-squares are given in section 3 with analytic expressions for various properties of the estimation process are described in section 3.2. Intuitively, smaller bias is obtained by an informed choice of L (the goal being small $L\beta$). The affect of such a choice on the variance is described analytically. Section 4 describes several classes of structured penalties including two previously-proposed special cases that were justified by numerical simulations. The targeted penalties of subsection 4.2 are studied in more detail in section 5 including an analysis of the mean squared error for a family of penalized estimates which encompasses the ridge, principal-component and James-Stein estimates.

The assumptions on L here are increasingly restrictive to the point where the estimates are only minor extensions of these well-studied estimates. The goal, however, is to analytically describe the substantial gains achievable by even mild extensions of these established methods.

In applications the selection of the tuning parameter, α in (1), is important and so Section 6 describes our application of REML-based estimation for this. Numerical illustrations are provided in section 7: the simulation in subsection 7.1 is motivated by Reiss and Ogden's study of fLMs [40]; 7.2 presents a simulation using experimentally-derived Raman spectroscopy curves in which the "true" β has naturally-occurring (laboratory) structure; and section 7.3 presents an application based on experimentally collected spectroscopy curves representing varied biochemical (nanoparticle) concentrations. An appendix looks at the simulation studied by Hall and Horowitz [16]. We begin in section 2 with a brief setup for notation and an introductory example. Note that for any $L \neq I$, the estimated β is not given in terms of the ordinary empirical singular vectors (of X), but rather in terms of a "partially empirical" basis arising from a simultaneous diagonalization of $X'X$ and $L'L$ via the GSVD. Hence, for brevity, we refer to $\tilde{\beta}_{\alpha,L}$ as a PEER (partially empirical eigenvector for regression) estimate whenever $L \neq I$.

2. Background and simple example

Let β represent a linear functional on $L^2(I)$ defining a linear relationship $y = \int_I x(t)\beta(t)\,dt + \varepsilon$ (observed with error, ε) between a response, y, and random predictor function, $x \in L^2(I)$. We assume a set of n scalar responses $\{y_i\}_{i=1}^n$ corresponding to the set of n predictors, $\{x_i\}_{i=1}^n$, each discretely sampled at common locations in I. Denote by X the $n \times p$ matrix whose ith row is a p-dimensional vector, x_i, of discretely sampled functions, and columns that are centered to have mean 0. The notation $\langle \cdot, \cdot \rangle$ will be used to denote the inner product on either $L^2(I)$ or $\mathbb{R}^p$, depending on the context.

The empirical covariance operator is $K=\frac{1}{n}X'X$, but for functional predictors, typically $p > n$ or else K is ill-conditioned or rank deficient. In this case, there are either infinitely many least-squares solutions, $\hat{\beta} \equiv \arg\min_\beta \|y - X\beta\|^2$, or else any such solution is highly unstable and of little use. The least-squares solution having minimum norm is unique, however, and it can be obtained directly by the singular value decomposition (SVD): $X = UDV'$ where the left and right singular vectors, u_k and v_k, are the columns of U and V, respectively, and D =[D$_1$ 0], where $D_1=\text{diag}\{\sigma_k\}_{k=1}^n$, typically ordered as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ ($r = \text{rank(X)}$). In terms of the SVD of X, the minimum-norm solution is $\widehat{\beta}_+=X^\dagger y=\sum_{\sigma_k\neq 0}(1/\sigma_k)u_k'y\,v_k$, where $X^\dagger$ denotes the Moore-Penrose inverse of X: $X^\dagger = VD^\dagger U'$, where $D^\dagger = \text{diag}\{1/\sigma_k \text{ if } \sigma_k \neq 0; 0 \text{ if } \sigma_k = 0\}$. The orthogonal vectors that form the columns of V are the eigenvectors of $X'X$ and sometimes referred to as a Karhunen-Lòeve (K-L) basis for the row space of X.

The solution $\hat{\beta}_+$ is Marquardt's generalized inverse estimator whose properties are discussed in [29]. For functional data, $\hat{\beta}_+$ is an unstable, meaningless solution. One obvious fix is to truncate the sum to $d < r$ terms so that $\{\sigma_k\}_{k=1}^d$ is bounded away from zero. This leads to the truncated singular value or principal component regression (PCR) estimate: $\tilde{\beta}_{\text{PCR}} \equiv \tilde{\beta}_{\text{PCR}}^d=V_dD_d^{-1}U_d'y$ where here, and subsequently, we use the notation $A_d \equiv \text{col}[a_1, \ldots, a_d]$ to denote the first d columns of a matrix A.

When $L = I$, the minimizer in (1) is the ridge penalty due to A. E. Hoerl [23]

$$\tilde{\beta}_{\alpha,I}=(X'X+\alpha I)^{-1}X'y=\sum_{k=1}^{r}\left(\frac{\sigma_k^2}{\sigma_k^2+\alpha}\right)\frac{1}{\sigma_k}u_k'y\,v_k, \qquad (2)$$

or, $\tilde{\beta}_{\alpha,I}=V_rF^{\alpha}D_r^{\dagger}U_r{}'y$, where $F^{\alpha}=\text{diag}\{\frac{\sigma_k^2}{\sigma_k^2+\alpha}\}$. The factor F^{α} acts to counterweight, rather than truncate, the terms $\frac{1}{\sigma_k}$ as they get large. This is one of many possible filter factors which address problems of ill-determined rank (for more, see [12, 19, 33]). Weighted (or generalized) ridge regression replaces $L = I$ with a diagonal matrix whose entries downweight those terms corresponding to the most variation [23]. Other “generalized ridge” estimates replace $L = I$ by a discretized second-derivative operator, $L = \mathcal{D}^2$. Indeed, the Tikhonov-Phillips form of regularization (1) has a long history in the context of differential equations [51, 36] and image analysis [15, 33] with emphasis on numerical stability; see also [28]. In a linear model context, the smoothing imposed by this penalty was mentioned by Hastie and Mallows [21], discussed in Ramsay and Silverman [38], and used (on the space of spline-transform coefficients) by Marx and Eilers [31], among others. The following simple example illustrates basic behavior for some of these penalties alongside an idealized PEER penalty.

2.1. A simple example

We consider a set of $n = 50$ bumpy predictor curves $\{x_i\}$ discretely sampled at $p = 250$ locations, as displayed in gray in the last panel of Figure 1. The true coefficient function, β, is displayed in black in this same panel. The responses are defined as $y_i = \langle x_i, \beta\rangle + \varepsilon_i$ (ε_i normal, uncorrelated mean-zero errors), and hence depend on the amplitudes of β’s three bumps centered at locations $t = 45, 110, 210$.

A detailed simulation with complete results are provided in section 7.1. Here we simply illustrate the estimation process for $L = I$, as in (2), in comparison with $L = \mathcal{D}^2$ and an idealized PEER penalty. The latter is constructed using a visual inspection of the predictors and lightly penalizes the subspace spanned by such structure, specifically, bumps centered at all visible locations (approximately $t = 15, 45, 80, 110, 160, 210, 240$).

The first five panels serve to emphasize the role played by the structure of basis vectors that comprise the series expansion in (2) (in terms of ordinary singular vectors) versus the analogous expansion (see (7)) in terms of generalized singular vectors. In particular, Figure 1 shows several partial sums of (7) for these three penalties. The ridge process (gray) is, naturally, dominated by the right singular vectors of X which become increasingly noisy in successive partial sums. The second-derivative penalized estimate (dashed) is dominated by low-frequency structure, while the targeted PEER estimate converges quickly to the informative features.

In this toy example, visual structure (spatial location) is used to define a regularization process that easily outperforms uninformed methods of penalization. Less visual examples where the penalty is defined by a set of laboratory-derived structure (in Raman spectroscopy curves) is given in sections 7.2 and 7.3; see Figure 2. In that setting, and in general, the role played by L is appropriately viewed in terms of a preferred subspace in $\mathbb{R}^p$ determined by its singular vectors. Algebraic details about how structure in the estimation process is determined jointly by X and $L \neq I$ are described next.

3. Penalized least squares and the GSVD

Of the many methods for estimating a coefficient function discussed in the Introduction, nearly are all aimed at imposing geometric or “functional” structure into the process via the use of basis functions in some manner. An alternative to choosing a basis outright is to exploit the structure imposed by an informed choice of penalty operator. The basis, determined by a pair (X, L), can be tailored toward structure of interest by the choice of L. When this is carried out in the least-squares setting of (1), the algebraic properties of the GSVD explicitly reveal how the structure of the estimate is inherited from the spectral properties of (X, L).

3.1. The GSVD

For a given linear penalty L and parameter $\alpha > 0$, the estimate in (1) takes the form

$$\tilde{\beta}_{\alpha,L} = (X'X + \alpha L'L)^{-1} X'y. \tag{3}$$

This cannot be expressed using the singular vectors of X alone, but the generalized singular value decomposition of the pair (X, L) provides a tractable and interpretable series expansion. The GSVD appears in the literature in a variety of forms and notational conventions. Here we provide the necessary notation and properties of the GSVD for our purposes (see, e.g., [19]) but refer to [4, 13, 35] for a complete discussion and details about its computation. See also the comments of Bingham and Larntz [3].

Assume X is an $n{\times}p$ matrix $(n \le p)$ of rank n, L is an $m{\times}p$ matrix $(m \le p)$ of rank m. We also assume that $n \le m \le p \le m + n$, and the rank of the $(n + m) \times p$ matrix $Z: = [X' \; L']'$ is p. A unique solution is guaranteed if the null spaces of X and L intersect trivially: Null$(L) \cap$ Null$(X) = \{0\}$. This is not necessary for implementation, but it is natural in our applications and simplifies the notation. In addition, the condition $p > n$ is not required, but rather than develop notation for multiple cases, this will be assumed.

Given X and L, the following matrices exist and form the decomposition below: an $n \times n$ matrix U and an $m \times m$ matrix V, each with orthonormal columns, $U'U = I$, $V'V = I$; diagonal matrices S $(n \times n)$ and M $(m \times m)$; and a nonsingular $p \times p$ matrix W such that

$$\begin{array}{lll} X = U\underline{S}W^{-1}, & \underline{S} = \begin{bmatrix} 0 & S \end{bmatrix}, & S = \mathrm{diag}\{\sigma_k\} \\ L = V\underline{M}W^{-1} & \underline{M} = \begin{bmatrix} M & 0 \end{bmatrix}, & M = \mathrm{diag}\{\mu_k\}. \end{array} \tag{4}$$

Here, S and M are of the form $S = \begin{bmatrix} S_1 & 0 \\ 0 & I_{p-m} \end{bmatrix}$ and $M = \begin{bmatrix} I_{p-n} & 0 \\ 0 & M_1 \end{bmatrix}$, whose submatrices S_1 and M_1 have $\ell := n + m - p$ diagonal entries ordered as

$$\begin{array}{c} 0 \le \sigma_1 \le \sigma_2 \le \cdots \le \sigma_\ell \le 1 \\ 1 \ge \mu_1 \ge \mu_2 \ge \cdots \ge \mu_\ell \ge 0 \end{array} \quad \text{where} \quad \sigma_k^2 + \mu_k^2 = 1, \quad k = 1, \ldots, \ell, \tag{5}$$

These matrices satisfy

$$W'X'XW = \begin{bmatrix} 0 & 0 & 0 \\ 0 & S_1^2 & 0 \\ 0 & 0 & I \end{bmatrix} = \underline{S}'\underline{S}, \quad W'L'LW = \begin{bmatrix} I & 0 & 0 \\ 0 & M_1^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \underline{M}'\underline{M}, \tag{6}$$

with $\underline{S}'\underline{S} + \underline{M}'\underline{M} = I$.

Denote the columns of U, V and W by u_k, v_k and w_k, respectively. When $L = I$, the the generalized singular vectors u_k and v_k are those in the ordinary SVD of X (as denoted in Section 2) but their ordering is reversed. In this case, the corresponding ordinary singular values are equal to $\gamma_k := \sigma_k/\mu_k$ for $\mu_k > 0$. When $L \neq I$, however, the GSVD vectors and values of the pair (X, L) differ from those in the SVD of X. By the convention for ordering the GS values and vectors, the last few columns of W span the subspace Null(L) (or, if Null(L) is empty, they correspond to the smallest GS values, μ_k). We set d = dim(Null(L)) and note that $\mu_k = 0$ for $k > n - d$. In the special case that $m = p$ and L is a $p \times p$ nonsingular matrix, we have $L = VMW^{-1}$ and $XL^{-1} = U(\underline{S}\,\underline{M}^{\dagger})V'$, which connects the SVD of XL^{-1} to the GSVD of (X, L). In general, we define the $n \times m$ matrix $\underline{\Gamma}:=\underline{S}\,\underline{M}^{\dagger}=[\begin{matrix} 0 & S_1 M_1^{\dagger} \end{matrix}]$ and the $l \times l$ diagonal matrix $\Gamma:=S_1 M_1^{\dagger}$ with entries $\gamma_k = \sigma_k/\mu_k$ for $\mu_k > 0$, and $\gamma_k = 0$ for $\mu_k = 0$.

Now, equation (6) and some algebra gives $(X'X + aL'L)^{-1} = W(\underline{S}'\underline{S} + a\underline{M}'\underline{M})^{-1}W'$, and so $\tilde{\beta}_{a,L} = W(\underline{S}'\underline{S} + a\underline{M}'\underline{M})^{-1}\underline{S}'U'y$. A consequence of the ordering adopted for the GS values and vectors is that the first $p{-}n$ columns of W don't enter into the expression for $\tilde{\beta}_{a,L}$; see equation (4). Therefore, we will re-index the columns of W to reflect this and also so that the indexing coincides with that established for the GS values and vectors in (5). That is, denote the columns of W as follows: $W := \text{col}[w_{(1)}, \ldots, w_{(p-n)} \mid w_1, \ldots, w_{n-d} \mid w_{n-d+1}, \ldots, w_n]$. Therefore, the L-penalized estimate can be expressed as a series in terms of the GSVD as

$$\tilde{\beta}_{\alpha,L}=\sum_{k=1}^{n-d}\left(\frac{\sigma_k^2}{\sigma_k^2+\alpha\mu_k^2}\right)\frac{1}{\sigma_k}u_k'y\,w_k+\sum_{k=n-d+1}^{n}u_k'y\,w_k. \tag{7}$$

This GSV expansion corresponds to a new basis for the estimation process: the estimate is expressed in terms of GS vectors $\{w_k\}$ determined jointly by X and L; cf. the ridge estimate in (2).

For brevity, set o:= $n - d$ and recall that A_{o} denotes the first $n - d$ columns of a matrix A. Now denote by A_{φ} the *last* d columns of A. In particular, the range of W_{φ} is Null(L). Using this notation,

$$W=\text{col}[w_{(1)},\ldots,w_{(p-n)}|w_1,\ldots,w_{n-d}|w_{n-d+1},\ldots,w_n] \equiv [W_{(*)}|W_{\text{o}}|W_{\varphi}] \tag{8}$$

and equation (7) may be written concisely as

$$\tilde{\beta}_{\alpha,L}=W_oF^{\alpha}\Gamma_o^{\dagger}U_o'y+W_{\varphi}U_{\varphi}'y, \tag{9}$$

where $F^{\alpha}=\text{diag}\left\{\frac{\sigma_k^2}{\sigma_k^2+\alpha\mu_k^2}\right\}_{k=1}^{n-d}$.

In summary, the utility of a penalty L depends on whether the true coefficient function shares structural properties with this GSVD basis, $\{w_k\}_{k=1}^{n}$. With regard to this, the importance of the parameter a may be reduced by a judicious choice of L since the terms in (7) corresponding to the vectors $\{w_k: \mu_k = 0\}$ are independent of the parameter a [53].

As we'll see, bias enters the estimate to the extent that the vectors $\{w_k: \mu_k \neq 0\}$ appear in the expansion (7). The portion of $\tilde{\beta}_{a,L}$ that extends beyond the subspace Null(L) is constrained by a sphere (of radius determined by a); this portion corresponds to bias. Hence,

L may be chosen in such a way that the bias and variance of $\tilde{\beta}_{\alpha,L}$ arises from a specific type of structure, potentially decreasing bias without increasing complexity of the model. As a common example, $L = \mathcal{D}^2$ introduces a bias toward smoothness with structure imposed by the low-frequency trigonometric functions that correspond to its smallest eigenfunctions.

3.2. Bias and variance and the choice of penalty operator

Begin by observing that the penalized estimate $\tilde{\beta}_{\alpha,L}$ in (3) is a linear transformation of any solution to the normal equations. Indeed, define $X^{\#}=X^{\#}_{\alpha,L}=(X'X+\alpha L'L)^{-1}X'$ and note that if $\hat{\beta}$ denotes any solution to $X'X\beta = X'y$, then $\tilde{\beta}_{\alpha,L} = X^{\#}X\hat{\beta}+X^{\#}\varepsilon$. The *resolution* operator $X^{\#}X$ reflects the extent to which the estimate in (7) is linearly transformed relative to an exact solution. In particular, $E(\tilde{\beta}_{\alpha,L}) = X^{\#}X\beta$. Additionally, we have $\text{bias}(\tilde{\beta}_{\alpha,L}) = (I - X^{\#}X)\beta = \alpha(X'X + \alpha L'L)^{-1}L'L\beta$, and so $\|\text{bias}(\tilde{\beta}_{\alpha,L})\| \le \|\alpha(X'X + \alpha L'L)^{-1}L'\|\|L\beta\|$. Hence bias can be controlled by the choice of L, with an estimate being unbiased whenever $L\beta = 0$. There is a tradeoff, of course, and equation (11) below quantifies the effect on the variance as determined by W_{φ} (i.e., $\{w_k\}_{k=n-d+1}^{n}$) if Null(L) is chosen to be too large.

More generally, the decompositions in (4) lead to an expression for the resolution matrix as $X^{\#}X = W(\underline{S}'\underline{S} + \alpha\underline{M}'\underline{M})^{-1}\underline{S}'\underline{S}W^{-1}$, and

$$\begin{aligned} I - X^{\#}X &= \alpha W(\underline{S}'\underline{S}+\alpha\underline{M}'\underline{M})^{-1}\underline{M}'\underline{M}W^{-1} \\ &= \alpha W\left(\begin{bmatrix} 0 & 0 & 0 \\ 0 & S_1^2 & 0 \\ 0 & 0 & I \end{bmatrix} + \alpha\begin{bmatrix} I & 0 & 0 \\ 0 & M_1^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right)^{-1}\begin{bmatrix} I & 0 & 0 \\ 0 & M_1^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}W^{-1} \\ &= W\begin{bmatrix} I & 0 & 0 \\ 0 & \alpha(S_1^2+\alpha M_1^2)^{-1}M_1^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}W^{-1}. \end{aligned}$$

For notational convenience, define $\tilde{W} := W'^{-1}$ (note, $\tilde{W}$ plays a role analogous to $V \equiv V'^{-1}$ in the SVD). The bias of $\tilde{\beta}_{\alpha,L}$ can be expressed as

$$\text{bias}(\tilde{\beta}_{\alpha,L})=(I - X^{\#}X)\beta=\sum_{k=1}^{n-d}\frac{\alpha\mu_k^2}{\sigma_k^2+\alpha\mu_k^2}w_k\tilde{w}_k'\beta+\sum_{j=1}^{p-n}w_{(j)}\tilde{w}_{(j)}'\beta \qquad (10)$$

where $\tilde{w}_k$ is the kth column of $\tilde{W}$, and the $w_{(j)}$'s come from the first $p - n$ columns of W; see (8).

A counterpart is an expression for the variance in terms of the GSVD. Let Σ denote the covariance for ε. Then $\text{var}(\tilde{\beta}_{\alpha,L}) = \text{var}(X^{\#}X\beta + X^{\#}\varepsilon) = X^{\#}\Sigma(X^{\#})'$. Assuming $\Sigma = \sigma_\varepsilon I$, this simplifies to

$$\text{var}(\tilde{\beta}_{\alpha,L})=\sigma_\varepsilon^2 X^{\#}(X^{\#})'=\sigma_\varepsilon^2\left(\sum_{k=1}^{n-d}\frac{\sigma_k^2}{(\sigma_k^2+\alpha\mu_k^2)}w_k w_k'+\sum_{k=n-d+1}^{n}w_k w_k'\right). \qquad (11)$$

An interesting perspective of the bias-variance tradeoff is provided by the relationship between the GS-values in (5) and their role in equations (10) and (11). Moreover, these lead to an explicit expression for the mean squared error (MSE) of a PEER estimate. Since $E(\tilde{\beta}_{\alpha,L}) = X^{\#}X\beta$,

$$\begin{aligned}\mathrm{MSE}(\tilde{\beta}_{\alpha,L}) &= E(\|\beta-\tilde{\beta}_{\alpha,L}\|^2) = E(\|\beta\|^2+\|\tilde{\beta}_{\alpha,L}\|^2-2\langle\beta,\tilde{\beta}_{\alpha,L}\rangle)\\ &= \|\beta-X^{\#}X\beta\|^2+\sigma_\varepsilon^2\,\mathrm{trace}(X^{\#}X^{\#'})\\ &= \|\mathrm{bias}(\tilde{\beta}_{\alpha,L})\|^2+\sigma_\varepsilon^2\sum_{k=1}^{n-d}\frac{\sigma_k^2}{(\sigma_k^2+\alpha\mu_k^2)^2}\|w_k\|^2.\end{aligned}$$

As a final remark, recall that one perspective on ridge estimation defines fictitious data from an orthogonal "experiment," represented by an L, and expresses I as $I = L'L$ [29]. Regardless of orthogonality this applies to any penalized estimate and L may similarly be viewed as augmenting the data, influencing the estimation process through its eigenstructure; the response, y, is set to zero for these supplementary "data". In this view, equation (3) can be written as $Z\beta = \underline{y}$ where $Z=\begin{bmatrix}X\\ \sqrt{\alpha}L\end{bmatrix}$ and $\underline{y}=\begin{bmatrix}y\\ 0\end{bmatrix}$. This formulation proves useful in section 5.3 when assuring that the estimation process is stable with respect to perturbations in X and the choice of L.

4. Structured penalties

A *structured penalty* refers to a second term in (1) that involves an operator chosen to encourage certain functional properties in the estimate. A prototypical example is a derivative operator which imposes smoothness via its eigenstructure. Here we describe several examples of structured penalties, including two that were motivated heuristically and implemented without regard to the spectral properties that define their performance. Sections 3.2 and 5.3 provide a complete formulation of their properties as revealed by the GSVD.

4.1. The penalty of C. Goutis

The concept of using a penalty operator whose eigenstructure is targeted toward specific properties in the predictors appears implicitly in the work of C. Goutis [14]. This method aimed to account for the "functional nature of the predictors" without oversmoothing and, in essence, considered the inverse of a smoothing penalty. Specifically, if Δ denotes a discretized second-derivative operator (with some specified boundary conditions), the minimization in (1) was replaced by $\min_\beta\{\|Y-X\Delta'\Delta\beta\|_{\mathbb{R}^n}^2+\alpha\|\Delta\beta\|_{L^2}^2\}$. Here, the term $X\Delta'\Delta\beta$ can be viewed as the product of $X\Delta'$ (derivatives of the predictor curves) and $\Delta\beta$ (derivative of β). Defining $\gamma:=\Delta'\Delta\beta$ and seeking a penalized estimate of γ leads to

$$\begin{aligned}\tilde{\gamma} &= (X'X+\alpha(\Delta'\Delta)^{\dagger})^{-1}X'y\\ &= \arg\min_{\gamma}\{\|y-X\gamma\|^2+\alpha\langle\gamma,(\Delta'\Delta)^{\dagger}\gamma\rangle\}.\end{aligned} \qquad (12)$$

In [14], the properties of $\tilde{\gamma}$ were conjectured to result from the eigenproperties of $(\Delta'\Delta)^{\dagger}$. This was explored by ignoring X and plotting some eigenvectors of $(\Delta'\Delta)^{\dagger}$. The properties of this method become transparent, however, when formulated in terms of the GSVD. That is, let $L:=((\Delta'\Delta)^{\dagger})^{1/2}$ and note the functions that define $\hat{\gamma}$ are influenced most by the highly oscillatory eigenvectors of L which correspond to its smallest eigenvalues; see equations (5) and (7).

This approach was applied in [14] only for prediction and has drawbacks in producing an interpretable estimate, especially for non-smooth predictor curves. The general insight is valid, however, and modifications of this penalty can be used to produce more stable results. The operator $(\Delta'\Delta)^{\dagger}$ essentially reverses the frequency properties of the eigenvectors of Δ and is an extreme alternative to this smoothing penalty. An eigenanalysis of the pair (X, L), however, suggests penalties that may be more suited to the problem. This is illustrated in Section 7.

4.2. Targeted penalties

Given some knowledge about the relevant structure, a penalty can be defined in terms of a subspace containing this structure. For example, suppose $\beta \in \mathcal{Q}:=\text{span}\{q_j\}_{j=1}^{d}$ in $L^2(\mathcal{I})$. Set $Q=\sum_{j=1}^{d} q_j \otimes q_j$ and consider the orthogonal projection $P_{\mathcal{Q}} = QQ^{\dagger}$. (Here, $q \otimes q$ denotes the rank-one operator $f \mapsto \langle f, q\rangle q$, for $f \in L^2(\mathcal{I})$.) For $L = I - P_{\mathcal{Q}}$, then $\beta \in \text{Null}(L)$ and $\tilde{\beta}_{\alpha,L}$ is unbiased. The problem may still be underdetermined so, more pragmatically, define a decomposition-based penalty

$$L \equiv L_{Q}=a(I-P_{Q})+bP_{Q} \tag{13}$$

for some $a, b \geq 0$. Heuristically, when $a > b > 0$ the effect is to move the estimate towards Q by preferentially penalizing components orthogonal to $\mathcal{Q}$; i.e., assign a prior favoring structure contained in the subspace $\mathcal{Q}$. To implement the tradeoff between the two subspaces, we view a and b as inversely related, ab = const. The analytical properties of estimates that arise from this are developed in the next section and illustrated numerically in Section 7. For example, bias is substantially reduced when $\beta \subset \mathcal{Q}$, and equation (18) quantifies the tradeoff with respect to variance when the prior $\mathcal{Q}$ is chosen poorly.

More generally, one may penalize each subspace differently by defining $L = a_1(I - P_{\mathcal{Q}})\ \mathcal{L}_1(I - P_{\mathcal{Q}}) + a_2\ P_{\mathcal{Q}}\ \mathcal{L}_2\ P_{\mathcal{Q}}$, for some operators $\mathcal{L}_1$ and $\mathcal{L}_2$. This idea could be carried further: for any orthogonal decomposition of $L^2(\mathcal{I})$ by subspaces $\mathcal{Q}_1, \ldots, \mathcal{Q}_J$, let P_j be the projection onto $\mathcal{Q}_j$. Then the multi-space penalty $L=\sum_{j=1}^{J} \alpha_j P_j$ leads to the estimate

$$\tilde{\beta}=\arg\min_{\beta}\{\|y-X\beta\|^2+\sum_{j=1}^{J}\alpha_j\|P_j\beta\|^2\}.$$

This concept was applied in the context of image recovery (where X represents a linear distortion model for a degraded image y) by Belge et al. [1].

The examples here illustrate ways in which assumptions about the structure of a coefficient function can be incorporated directly into the estimation process. In general, any estimation of β imposes assumptions about its structure (either implicitly or explicitly) and section 3.2 shows that the bias-variance tradeoff involves a choice on the *type* of bias (spatial structure) as well as the *extent* of bias (regularization parameter(s)).

5. Some analytical properties

Any direct comparison between estimates using different penalty operators is confounded by the fact there is no simple connection between the generalized singular values/vectors and the ordinary singular values/vectors. Therefore, we first consider the case of targeted or projection-based penalties (13). Within this class, we introduce a parameterized family of estimates that are comprised of *ordinary* singular values/vectors. Since the ridge and PCR estimates are contained in (or a limit of) this family, a comparison with some targeted PEER estimates is possible. For more general penalized estimates, properties of perturbations provide some less precise relationships; see proposition 5.6.

5.1. Transformation to standard form

We have reason to consider decomposition-based penalties (13) in which L is invertible. In this case, an expression for the estimate does not involve the second term in (9), and decomposing the first term into two parts will be useful. For this, we find it convenient to use the standard-form transformation due to Elden [11] in which the penalty L is absorbed into X. This transformation also provides a computational alternative to the GSVD which, for projection-based penalties in particular, can be less computationally expensive; see, e.g., [25]. By this transformation of X, a general PEER estimate ($L \neq I$) can be expressed via a ridge-regression process.

Define the *X-weighted generalized inverse of L* and the corresponding transformed X as

$$L_X^{\dagger}:=(I-[X(I-L^{\dagger}L)]^{\dagger}X)L^{\dagger} \quad \text{and} \quad \mathbb{X}:=X\,L_X^{\dagger};$$

see [11, 19]. In terms of the GSVD components (4), the transformed X is $\mathbb{X} = U\underline{\Gamma}V'$. In particular, the diagonal elements of $\Gamma=S_1M_1^{\dagger}$ are the ordinary singular values of $\mathbb{X}$, but in reversed order.

Now define $\tilde{\beta}_{\varphi}:= [X(I-L^{\dagger}L)]^{\dagger}y$, the component of the regularized solution $\tilde{\beta}_{a,L}$ that is in Null(L). The PEER estimate can be obtained from a ridge-like penalization process with respect to $\mathbb{X}$ as follows. Defining a ridge estimate in the transformed space as

$$\tilde{\mathbb{\beta}}_{\alpha} = \arg\min_{\mathbb{\beta}}\{||\mathbb{y}-\mathbb{X}\mathbb{\beta}||^2+\alpha||\mathbb{\beta}||^2\} \quad \text{where} \quad \mathbb{y}=y-X\tilde{\beta}_{\emptyset}, \tag{14}$$

then the PEER estimate is recovered as

$$\tilde{\beta}_{\alpha,L} = L_X^{\dagger}\tilde{\mathbb{\beta}}_{\alpha}+\tilde{\beta}_{\emptyset}.$$

Note that the transformed estimate as given in terms of the GSVD factors is: $\tilde{\mathbb{\beta}}_{\alpha} = VF\underline{\Gamma}^{\dagger}U'y$, where F=diag$\{\gamma_k^2/(\gamma_k^2+\alpha)\}$.

In what follows we consider invertible L in which case $L_X^{\dagger}=L^{-1}$, $[X(I-L^{\dagger}L)]^{\dagger}=0$, and $\mathbb{y} = y$. In particular, $\tilde{\beta}_{a,L} = L^{-1}\tilde{\mathbb{\beta}}_{\alpha}$. For the penalty (13) of the form $L = a(I-P_Q) + bP_Q$, then $L^{-1}=\frac{1}{a}(I-P_Q)+\frac{1}{b}P_Q$, and so $\mathbb{X}=\frac{1}{a}X(I-P_Q)+\frac{1}{b}XP_Q$. The regularization parameter, previously denoted by α, can be absorbed into the values a and b, so we will denote this PEER estimate $\tilde{\beta}_{a,L}$ simply as $\tilde{\beta}_{a,b}$.

Remark 5.1—When $a=b=\sqrt{\alpha}$, this is simply a ridge estimate: $\tilde{\beta}_{a,b} = \tilde{\beta}_{\alpha,I}$. Therefore, the best performance among this family of estimates is as least as good as the performance of ridge, regardless of the choice of $\mathcal{Q}$.

5.2. SVD targeted penalties

Consider the special case in which $\mathcal{Q}$ is the span of the d largest right singular vectors of an $n \times p$ matrix X of rank n. Let $X=U\left[\begin{array}{cc} 0 & D \end{array}\right]V'$ be an ordinary singular value decomposition where D is a diagonal matrix of singular values. For consistency with the GSVD notation, these will be ordered as $0 \le \sigma_1 \le \cdots \le \sigma_n$. Hence the *last d* columns of V correspond to the d largest singular values of X. For the rest of this section, we adopt the convention for indexing the columns of V as use for W in (8). In particular, $Q = V_\varphi$.

We are interested interested in the penalty $L = a(I - P_{\mathcal{Q}}) + bP_{\mathcal{Q}}$, where $d = \dim(\mathrm{Null}(I - P_{\mathcal{Q}}))$. Similar to before, set o = $n - d$ and define o × o and $d \times d$ submatrices, D_{o} and D_φ, of D as

$$D=\left[\begin{array}{cc} D_{\mathrm{o}} & 0 \\ 0 & D_\varphi \end{array}\right]; \quad \text{also set} \quad \Lambda=\left[\begin{array}{cc} aI_{\mathrm{o}} & 0 \\ 0 & bI_d \end{array}\right]. \tag{15}$$

Here, $P_{\mathcal{Q}} = V_\varphi V_\varphi{}'$ and $(I - P_{\mathcal{Q}}) = V_{\mathrm{o}} V_{\mathrm{o}}{}'$ and so,

$$\begin{aligned}\mathbb{X}&=\tfrac{1}{a}UDV'(V_{\mathrm{o}}V_{\mathrm{o}}{}')+\tfrac{1}{b}UDV'(V_\varphi V_\varphi{}')=\tfrac{1}{a}UD\left[\begin{array}{c} V_{\mathrm{o}}{}' \\ 0 \end{array}\right]+\tfrac{1}{b}UD\left[\begin{array}{c} 0 \\ V_\varphi{}' \end{array}\right]\\ &=U\left[\begin{array}{c} \tfrac{1}{a}D_{\mathrm{o}}V_{\mathrm{o}}{}' \\ 0 \end{array}\right]+U\left[\begin{array}{c} 0 \\ \tfrac{1}{b}D_\varphi V_\varphi{}' \end{array}\right]=U(D\Lambda^{-1})V'.\end{aligned}$$

This decomposition implies that the ridge estimate in (14) in the transformed space is of the following form: setting $G = D\Lambda^{-1}$, denoting its diagonal entries by $\{\gamma_k\}$, and defining $F=\mathrm{diag}\{\gamma_k^2/(\gamma_k^2+1)\}$ gives $\hat{\beta} = VFG^\dagger U'y$. Now,

$$L^{-1}V=\frac{1}{a}V_{\mathrm{o}}V_{\mathrm{o}}{}'V+\frac{1}{b}V_\varphi V_\varphi{}'V=\left[\begin{array}{c} V_{\mathrm{o}} \\ V_\varphi \end{array}\right]\left[\begin{array}{cc} \tfrac{1}{a}I_{\mathrm{o}} & 0 \\ 0 & \tfrac{1}{b}I_d \end{array}\right]=V\Lambda^{-1}$$

and so $\tilde{\beta}_{a,b} = L^{-1}\hat{\beta} = L^{-1}(VFG^\dagger U'y) = V\Lambda^{-1}F\Lambda D^{-1}U'y$. By the decomposition (15),

$$\tilde{\beta}_{a,b}=V_{\mathrm{o}}F_{\mathrm{o}}D_{\mathrm{o}}^{-1}U_{\mathrm{o}}{}'y+V_\varphi F_\varphi D_\varphi^{-1}U_\varphi{}'y.$$

This shows that the estimate decomposes as follows.

Theorem 5.2—Let $\mathcal{Q}$ be the span of the largest d right singular vectors of X. Set $L = a(I - P_{\mathcal{Q}}) + bP_{\mathcal{Q}}$. Then, in terms of the notation above, the estimate $\tilde{\beta}_{a,b}$ decomposes as

$$\tilde{\beta}_{a,b}=\sum_{k=1}^{n-d}\left(\frac{\sigma_k^2}{\sigma_k^2+a^2}\right)\frac{1}{\sigma_k}u_k'y\,v_k+\sum_{k=n-d+1}^{n}\left(\frac{\sigma_k^2}{\sigma_k^2+b^2}\right)\frac{1}{\sigma_k}u_k'y\,v_k, \tag{16}$$

where the left and right terms are independent of b and a, respectively.

Similar arguments can be used to decompose an estimate for arbitrary $\mathcal{Q}$:

$$\tilde{\beta}_{a,b}=\sum_{k=1}^{n-d}\left(\frac{\sigma_k^2}{\sigma_k^2+a^2\mu_k^2}\right)\frac{1}{\sigma_k}u_k'y\,w_k+\sum_{k=n-d+1}^{n}\left(\frac{\sigma_k^2}{\sigma_k^2+b^2\mu_k^2}\right)\frac{1}{\sigma_k}u_k'y\,w_k. \tag{17}$$

In this case, however, all terms are dependent on both a and b. Indeed, using notation as in (9) one can decompose $\mathbb{X}=\frac{1}{a}U_{\mathrm{o}}\Gamma_{\mathrm{o}}V_{\mathrm{o}}'+\frac{1}{b}U_{\varphi}\Gamma_{\varphi}V_{\varphi}'$ and obtain $\tilde{\beta}= VF\underline{\Gamma}^{\dagger}U'y$. However, $L^{-1}V= WM^{\dagger}$, and the non-orthogonal terms provided by W do not decompose the estimate into terms from the orthogonal sum $\mathcal{Q}\oplus\mathcal{Q}^{\perp}$.

The following corollary, along with Remark 5.1, records the manner in which (16) is a family of penalized estimates, parameterized by a, $b \geq 0$ and $d \in \{1, \ldots, n\}$, that extends some standard estimates.

Corollary 5.3—Under the conditions in Theorem 5.2,

1. when $a > b > 0$, $\tilde{\beta}_{a,b}$ is a sum of weighted ridge estimates on $\mathcal{Q}$ and $\mathcal{Q}^{\perp}$;
2. when $a > 0$ and $b = 0$, $\tilde{\beta}_{a,0}$ is given by (9), which is a sum of PCR and ridge estimates on $\mathcal{Q}$ and $\mathcal{Q}^{\perp}$, respectively;
3. for each d, the PCR estimate $\tilde{\beta}^{d}_{PCR}$ is the limit of $\tilde{\beta}_{a,0}$ as $a \to \infty$.

In item 2, this estimate is similar to PCR except that a ridge penalty is placed on the least-dominant singular vectors. Under the assumptions here, $w_k \equiv v_k$ are the ordinary singular vectors of X and the ordinary singular values appear as $\gamma_k = \sigma_k/\mu_k$, for $\mu_k > 0$. In the second term of (9), the singular vectors are in the null space of L (since $b = 0$), and so $\mu_k = 0$ and $\sigma_k = 1$, for $k = n - d + 1, \ldots, p$. Regarding item 3, although a PCR estimate is not obtained from equation (3) for any L, it is a limit of such estimates.

Other decompositions may be obtained simply by using a permutation, such as $Q = \Pi V$, for some $n \times n$ permutation matrix Π. Stein's estimate, $\tilde{\beta}_{a,\mathrm{S}}$, also fits into this framework as follows. When $X'X$ is nonsingular, then $\tilde{\beta}_{a,\mathrm{S}} = (X'X+\alpha X'X)^{-1}X'y$ (see, e.g., the class 'STEIN' in [10]), and $X'X = VD'DV'$. Hence this estimate arises from the penalty $L_{\mathrm{S}} = DV'$. This is a re-weighted version of $L = a(I - P_{\mathcal{Q}})$ where $d = n$, $Q = V$ and the parameter a is replaced by the matrix D. The result is a constant filter factor $F = \mathrm{diag}\{1/(1 + \alpha)\}$. Using $d < n$ and $Q = V_d$ is a natural extension of this idea. More generally, $\mathcal{Q}$ may be enriched with functions that span a wider range of structure potentially relevant to the estimate. This concept is illustrated in Section 7.3 where instead of V_d, we use a d-dimensional set of experimentally-derived "template" spectra supplemented with their derivatives to define $\mathcal{Q}$.

As an aside, we note that in a different approach to regularization one can define a general family of estimates arising from the SVD by way of $\tilde{\beta}_{h,\phi} = V\Sigma_h U'y$, where $\sum_h=\mathrm{diag}\{\frac{\sigma_k}{h}\phi(\frac{\sigma_k^2}{h^2})\}$, and ϕ: $\mathbb{R}_+ \to \mathbb{R}$ is an arbitrary continuous function [33]. A ridge estimate is obtained for $\phi(t) = 1/(1 + t)$, and PCR obtained for $\phi(t) = 1/t$ if $t > 1$, $\phi(t) = 0$ if $t \leq 1$ (an L^2-limit of continuous functions). This is similar to item 3 in Corollary 5.3, but the family of estimates $\tilde{\beta}_{h,\phi}$ is formulated in terms of functional filter factors rather than explicit penalty operators. Related to this is the fact that the optimal (with respect to MSE) estimate using SVD filter factors is, in the case $C = \sigma_{\varepsilon}I$, expressed as $\tilde{\beta}_{\mathrm{OH}} = VFD^{\dagger}V'y$, where $F=\mathrm{diag}\{\sigma_k^2/(\sigma_k^2+\sigma_{\varepsilon}^2(v_k'\beta)^{-2})\}$; see the "ideal filter" of O'Brien and Holt [34]. In fact, it's easy to check that this optimal estimate can be obtained as $\tilde{\beta}_{\mathrm{OH}} = \tilde{\beta}_{\alpha,L}$ for some $L \neq I$. Since $\tilde{\beta}_{\mathrm{OH}}$ involves knowledge of β, it is not directly obtainable but it points to the optimality of a PEER estimate.

5.3. The MSE of some penalized estimates

Theorem 5.2 is used here to show that $\tilde{\beta}_{a,b}$ can have smaller MSE than the ridge or PCR estimates for a wide range of values of a and/or b. The MSE is potentially decreased further when L is defined by a more general $\mathcal{Q}$. In that case, a general statement is difficult to formulate but Proposition 5.6 confirms that any improvement in MSE is robust to perturbations in L (e.g., general $\mathcal{Q}$) and errors in x.

An immediate consequence of Theorem 5.2 is that the mean squared error for an estimate in this family (16) decomposes into easily-identifiable terms for the bias and variance:

$$\mathrm{MSE}(\tilde{\beta}_{a,b})=\sum_{k=1}^{n-d}\left(\frac{a^2}{\sigma_k^2+a^2}\right)^2(v_k'\beta)^2+\sum_{k=n-d+1}^{n}\left(\frac{b^2}{\sigma_k^2+b^2}\right)^2(v_k'\beta)^2+\sum_{j=1}^{p-n}(v_{(j)}'\beta)^2+\sigma_\varepsilon^2\left(\sum_{k=1}^{n-d}\frac{\sigma_k^2}{(\sigma_k^2+a^2)^2}+\sum_{k=n-d+1}^{n}\frac{\sigma_k^2}{(\sigma_k^2+b^2)^2}\right). \tag{18}$$

The influence of $b = 0$ on the estimate is now clear: when the numerical rank of X is small relative to d, the σ_k's in the last term decrease and the contribution to the variance from this term increases—the estimate fails for the same reason that ordinary least-squares fails. Any nonzero b stabilizes the estimate in the same way that a nonzero α stabilizes a standard ridge estimate; the decomposition (16) merely re-focuses the penalty. This is illustrated in Section 7 (Table 1) and in the Appendix (Table 4). Although there are three parameters to consider, the MSE of $\tilde{\beta}_{a,b}$ is relatively insensitive to $b > 0$ for sufficiently large d. This could be optimized (similar to efforts to optimize the number of principal components) but here we assume approximate knowledge regarding $\mathcal{Q}$, hence d. Relationships between ridge, PCR and PEER estimates in this family $\{\tilde{\beta}_{a,b}\}_{a,b>0}$ can be quantified more specifically as follows.

Proposition 5.4—Suppose $\beta \in \mathcal{Q}$ and fix $\alpha > 0$. Then for any $a > \sqrt{\alpha}$, the ridge estimate satisfies

$$MSE(\tilde{\beta}_{\alpha,I}) \equiv MSE(\tilde{\beta}_{\sqrt{\alpha},\sqrt{\alpha}}) > MSE(\tilde{\beta}_{a,\sqrt{\alpha}}).$$

__Proof:__ This follows from the fact that if $\beta \in \mathcal{Q}$, then $V_0{}'\beta = 0$ and so the first term in (18) is zero. Therefore, the contribution to the MSE by the fourth term is decreased whenever $a > \sqrt{\alpha}$.

If β is exactly a sum of the d dominant right singular vectors, A PCR estimate using d terms may perform well, but in general it is not optimal:

Proposition 5.5—If $\beta \in \mathcal{Q}$, a sufficient condition for the PCR estimate to satisfy

$$MSE(\tilde{\beta}_{PCR}^{d}) \equiv MSE(\tilde{\beta}_{\infty,0}) > MSE(\tilde{\beta}_{\infty,b})$$

is

$$\sigma_\varepsilon^2\left(\sum_{k=n-d+1}^{n}\frac{1}{\sigma_k^2}+\frac{2d}{b^2}\right)>\left\|V_\varphi{}'\beta\right\|^2. \tag{19}$$

Note that the left side of (19) increases without bound as $\sigma_k \to 0$. Since $\left\|V_\varphi{}'\beta\right\|^2=\sum_{k=n-d+1}^{n}(v_k'\beta)^2$, and since the premise of PCR is that $v_k'\beta$ decreases with decreasing σ_k, this sufficient condition is entirely plausible.

Proof: If $\beta\in\mathcal{Q}$, then the first and third terms in (18) are zero and the MSE of $\tilde{\beta}_{\mathrm{PCR}}^{d}$ consists of the second and last terms of (18):

$$\mathrm{MSE}(\tilde{\beta}_{\mathrm{PCR}}^{d})=\sum_{k=1}^{n-d}(v_k'\beta)^2+\sigma_\varepsilon^2\sum_{k=n-d+1}^{n}\frac{1}{\sigma_k^2}.$$

In particular, a sufficient condition for this to exceed $\mathrm{MSE}(\tilde{\beta}_{\infty,b})$ is for the variance term to exceed the second and last terms of (18):

$$\sigma_\varepsilon^2\sum_{k=n-d+1}^{n}\frac{1}{\sigma_k^2}>\sigma_\varepsilon^2\sum_{k=n-d+1}^{n}\frac{\sigma_k^2}{(\sigma_k^2+b^2)^2}+\sum_{k=n-d+1}^{n}\left(\frac{b^2}{\sigma_k^2+b^2}\right)^2(v_k'\beta)^2.$$

One can check that this is satisfied when (19) holds.

A comment by Bingham and Larntz [3] on Dempster et al.'s intensive simulation study of ridge regression in [10] notes that "it is not at all clear that ridge methods offer a clear-cut improvement over [ordinary] least squares except for particular orientations of β relative to the eigenvectors of $X'X$." Equation (18) repeats this observation relating these two classical methods as well as the minor extensions contained in (16). If, on the other hand, the orientation of β relative to the v_k's is not favorable, i.e., if β is nowhere near the range of V, then a PEER estimate as in (17) is more desirable than the estimate in (16) (assuming sufficient information is available to form $\mathcal{Q}$).

In summary, the family of estimates $\{\tilde{\beta}_{a,b}\}_{a,b>0}$ in (16) represents a hybrid of ridge and PCR estimation. This family—based on the ordinary singular vectors of X—is introduced here to provide a framework within which these two familiar estimates can be compared to (slightly) more general PEER estimates. Direct analytical comparison between general PEER estimates is more difficult since there's no simple relationship between the generalized singular vectors for two different L (including $L = I$ versus $L \neq I$). However, it is important that the estimation process be stable with respect to changes in L and/or X. I.e., in going from an estimate in (16) to one in (17), the performance of the estimate should be predictably altered. Given an estimate in Proposition 5.4, if $\mathcal{Q}$ is modified and/or X is observed with error, the MSE of the corresponding estimate, $\tilde{\beta}_{\alpha,L}^{E}$, should be controlled: for sufficiently small perturbation E, the corresponding estimate $\mathrm{MSE}(\tilde{\beta}_{\alpha,L}^{E})$ should be close to $\mathrm{MSE}(\tilde{\beta}_{\alpha,I})$. This "stability" is true in general. To see this recall $Z=\begin{bmatrix} X' & \sqrt{\alpha}L' \end{bmatrix}'$, (of rank p) and $\underline{y}=\begin{bmatrix} y' & 0 \end{bmatrix}'$. Then another way to represent the estimate (3) is $\tilde{\beta}_{\alpha,L} = Z^{\dagger}\underline{y}$. Let $E=\begin{bmatrix} E_1{}' & E_2{}' \end{bmatrix}'$ for some $n\times p$ and $m\times p$ matrices E_1 and E_2. Set $Z_E = Z + E$ and denote the perturbed estimate by $\tilde{\beta}^{E}_{\alpha,L} = Z_E^{\dagger} y$. By continuity of the generalized inverse (e.g., [4], Section 1.4), $\lim_{\|E\|\to 0} Z_E^{\dagger} = Z^{\dagger}$ if and only if $\lim_{\|E\|\to 0}$ rank(Z_E) = rank(Z). Therefore, provided the rank of Z is not changed by E,

$$\lim_{\|E\|\to 0}\left| \left\|\tilde{\beta}_{\alpha,L} - \tilde{\beta}^{E}_{\alpha,L}\right\| \le \lim_{\|E\|\to 0} \left\|Z^{\dagger} - Z_E^{\dagger}\right\| \|y\| \right| = 0,$$

and hence $\mathrm{MSE}(\tilde{\beta}^{E}_{\alpha,L}) \to \mathrm{MSE}(\tilde{\beta}_{\alpha,L})$ as $\|E\| \to 0$. A more specific bound on the difference of estimates can be obtained under the condition $\|Z^{\dagger}\|\|E\| < 1$ which implies that $\left\|Z_E^{\dagger}\right\| < \frac{\|Z^{\dagger}\|}{1 - \|Z^{\dagger}\|\,\|E\|}$. This can be used to obtain the following bound.

Proposition 5.6—Assume $\|Z^{\dagger}\|\|E\| < 1$ and let $r = \underline{y} - Z\tilde{\beta}_{\alpha,L}$. Then

$$\left\|\tilde{\beta}_{\alpha,L} - \tilde{\beta}^{E}_{\alpha,L}\right\| \le \frac{\|Z^{\dagger}\|\,\|E\|}{1 - \|Z^{\dagger}\|\,\|E\|}\left(\left\|\tilde{\beta}_{\alpha,L}\right\| + \|Z^{\dagger}\|\,\|r\|\right).$$

See [4] and [18].

6. Tuning parameter selection

Despite our focus on the GSVD, the computation of a PEER estimate in (1) does not, of course, require that this decomposition be computed. Rather, the role of the GSVD has been to provide analytical insight into the role a penalty operator plays in the estimation process. For computation, on the other hand, we have chosen to use a method in which the tuning parameter, α, is estimated as part of the coefficient-function estimation process.

Because the choice of tuning parameter is so important, many selection criteria have been proposed, including generalized cross-validation (GCV) [9], AIC and its finite sample corrections [55]. As an alternative to GCV and AIC, a recently-proven equivalence between the penalized least squares estimation and a linear mixed model (LMM) representation [6] can be used. In particular, the best linear unbiased predictor (BLUP) of the response y is composed of the best linear unbiased estimator of the fixed effects and BLUP of the random effects for the given values of the random component variances (see [47] and [6]). Within the LMM framework, restricted maximum likelihood (REML) can be used to estimate the variance components and thus the choice of the tuning parameter, α, which is equal to the ratio of the error variance and the random effects variance [42]. REML-based estimation of the tuning parameter has been shown to perform at least as well as the other criteria and, under certain conditions, it seen to be less variable than GCV-based estimation [41]. In our case, the penalized least-squares criterion (1) is equivalent to

$$\tilde{\beta}_{\alpha,L} = \arg\min_{\beta}\left\{\left\|y - X_{unp}\beta_{unp} - X_{pen}\beta_{pen}\right\|^2 + \alpha\left\|L\beta\right\|^2\right\} \qquad (20)$$

where $\beta=[\beta'_{unp}\beta'_{pen}]'$, the X_{unp} corresponds to the unpenalized part of the design matrix, and X_{pen} to the penalized part.

For simplicity of presentation, we describe the transformation with an invertible L. However, a generalized inverse can be used in case L is not of full rank; see equation (14). Also, to facilitate a straightforward use of existing linear mixed model routines in widely available software packages (e.g., R [37] or SAS software [43]), we transform the coefficient vector β using the inverse of the matrix L. Let $X^\star = XL^{-1}$ and $\beta^\star = L\beta$. Then equation (20) can be modified as follows

$$\tilde{\beta}^\star_{\alpha,L}=\arg\min_{\beta}\{\left\|y-X^\star\beta^\star\right\|^2+\alpha\left\|\beta^\star\right\|^2\}.$$

This REML-based estimation of tuning parameters is used in the application of Section 7.3.

For estimation of the parameters a, b and α involved in the decomposition-based penalty of equation (16), we view a and b as weights in a tradeoff between the subspaces and assume ab = const. In the current implementation, we use REML to estimate α for a fixed value of a, and do a grid search over the a values to jointly select the tuning parameters which maximize the REML criterion.

7. Numerical examples

To illustrate algebraic properties given in Section 5, we consider PEER estimation alongside some familiar methods in several numerical examples. Section 7.1 elaborates on the simple example in Section 2.1. These mass spectrometry-like predictors are mathematically synthesized in a manner similar to the study of Reiss and Ogden [40] (see also a numerical study in [48]). Here, β is also synthesized to represent a spectrum, or specific set of bumps. In contrast, Section 7.3 presents a real application to Raman spectroscopy data in which a set of spectra $\{x_i\}$ and nanoparticle concentrations $\{y_i\}$ are obtained from sets of laboratory mixtures. This laboratory-based application is preceded in section 7.2 by a simulation that uses these same Raman spectra. In both Raman examples, targeted penalties (13) are defined using discretized functions q_j chosen to span specific subspaces, $\mathcal{Q}=\text{span}\{q_j\}_{j=1}^{d}$. As before, let $Q = \text{col}[q_1, \ldots, q_d]$ and $P_{\mathcal{Q}} = QQ^\dagger$.

Each section displays the results from several methods, including derivative-based penalties. Implementing these requires a choice of discretization scheme and boundary conditions which define the operator. We use $\mathcal{D}^2$ where $\mathcal{D} = [d_{i,j}]$ is a square matrix with entries $d_{i,i} = 1$, $d_{i,i+1} = -1$ and $d_{i,j} = 0$ otherwise. In addition to some standard estimates, sections 7.3 and 7.2 also consider FPCR$_R$, a functional PCR estimate described in [40]. This approach extends the penalized B-spline estimates of [8] and assumes $\beta = B\eta$ where B is an $p \times K$ matrix whose columns consist of K B-spline functions and η is a vector of B-spline coefficients. The estimation process takes place in the coefficient space using the penalty $L = \mathcal{D}^2$ applied to η. The FPCR$_R$ estimate further assumes $\beta = BV_d\,\eta$ (V_d as defined in section 2).

Estimation error is defined as mean squared error (MSE) $\|\beta - \tilde{\beta}_{\alpha,L}\|^2$, and the prediction error defined similarly as $\Sigma_i\,|y_i - \tilde{y}_i|^2$, where $\tilde{y}_i = \langle x_i, \tilde{\beta}\rangle$. Each simulation incorporates response random errors, $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$, added to the ith true response, $y_i^{\text{true}}=\langle x_i,\beta\rangle$. Letting S_Y^2 denote the sample variance in the set $\{y_i^{\text{true}}\}_{i=1}^{n}$, the response random errors, ε_i, are chosen such that $R^2{:=}S_Y^2/[S_Y^2{+}\sigma_\varepsilon^2]$(the squared multiple correlation coefficient of the true model) takes values 0.6 and 0.8. In sections 7.1 and 7.2, tuning parameters are chosen by a grid search. In section 7.3, tuning parameters are chosen using REML, as described in section 6.

7.1. Bumps simulation

Here we elaborate on the simple example of section 2.1. This simulation involves bumpy predictor curves $x_i(t)$ with a response y_i that depends on the amplitudes $x_i(t)$ at some of the bump locations, $t = c_k$, via the regression function β. In particular,

$$x_i(t)=\sum_{j\in J_X} a_{ij}\exp[-b_j(t-c_j)]+e_i(t), \qquad \beta(t)=\sum_{j\in J_\beta} a_j\exp[-b_j(t-c_j)],$$

for $t \in [0, 1]$, where $J_X = \{2, 6, 10, 14, 20, 26, 30\}$ and $J_\beta = \{6, 14, 26\}$; $a_\star$ are magnitudes, $b_\star$ are spreads, and $c_\star$ are the locations of the bumps. In the first simulation, we set $b_j = 10000$ and $c_j = 0.004(8j - 1)$, the same for each curve x_i. This mimics, for instance, curves seen in mass spectrometry data. The assumption $J_\beta \subset J_X$ simulates a setting in which the response is associated with a subset of metabolite or protein features in a collection of spectra. The a_{ij}'s are from a uniform distribution, and $a_j = 3, 5, 2$ for $j = 6, 14, 24$, respectively. We consider discretized curves, $x_i(t)$, evaluated at $p = 250$ points, t_j, $j = 1, \ldots, p$. The sample size is fixed at $n = 50$ in each case.

Penalties—We consider a variety of estimation procedures: ridge ($L = I$), second-derivative ($\mathcal{D}^2$), a more general derivative operator ($\mathcal{D}^2 + a\,I$) and PCR. We also define two decomposition-based penalties (13) formed by specific sub-spaces $\mathcal{Q} = \text{span}\{q_j\}_{j\in J}$ for q_j of the form $q_j(t) = a_j \exp[b_j(t-c_j)]$, with c_j at all locations seen in the predictors, $J_V = \{2, 6, 10, 14, 20, 26, 30\}$, or at uniformly-spaced locations, $J_U = \{2, 4, \ldots, 30\}$; denote these penalties by L_V and L_U, respectively.

Simulation results—The simulation incorporates two sources of noise: (i) response random errors, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, added to the ith true response so that $R^2 = 0.6, 0.8$; (ii) measurement error, $e_i \sim N_p(0, \sigma_e^2 I)$, added to the ith predictor, x_i. To define a signal-to-noise ratio, S/N, set $S_i^2 := \left\|x_i - \mu_i\right\|^2/(p-1)$, where μ_i is the mean value of x_i, and set $S_x^2 := 1/n\sum_i S_i^2$. The e_i are chosen so that S/N: $= S_X/\sigma_e = 2, 5, 10$.

Figure 1 shows a few partial sums of (7) for estimates arising from three penalties: $\mathcal{D}^2$, $L = I$ and L_V, when $R^2 = 0.8$ and $S/N = 2$. Table 1 gives a summary of estimation errors. The penalty L_V, exploiting known structure, performs well in terms of estimation error. Not surprisingly, a penalty that encourages low-frequency singular vectors, $\mathcal{D}^2$, is a poor choice although $\mathcal{D}^2 + a\,I$ easily improves on $\mathcal{D}^2$ since the GSVs are more compatible with the relevant structure. PCR performs well with estimation errors that can be several times smaller than those of ridge. The number of terms used in PCR ranges here from 8 ($S/N = 10$) to 25 ($S/N = 2$).

Predictably, PCR performance degrades with decreasing S/N, a property that is less pronounced, or not shared, by other estimates. Performances of L_V and L_U illustrate properties described in Section 5.3. As $S/N \to 0$, the ordinary singular vectors of X (on which ridge and PCR rely) decreasingly represent the structure in β. The GS vectors of (X, L_V) and (X, L_U), however, retain structure relevant for representing β.

Table 2 summarizes prediction errors. When S/N is large, performance of PCR is comparable with L_V and L_U, but degrades for low S/N. Here, even $\mathcal{D}^2 + aI$ provides smaller prediction errors, in most cases, than ridge, $\mathcal{D}^2$ or PCR. This illustrates the GS vectors role in (12) and reiterates observations in [14].

7.2. Raman simulation

We consider Raman spectroscopy curves which represent a vibrational response of laser-excited co-organic/inorganic nanoparticles (COINs). Each COIN has a unique signature spectrum and serves as a sensitive nanotag for immunoassays; see [27, 44]. Each spectrum consists of absorbance values measured at $p = 600$ wavenumbers. By the Beer-Lambert law, light absorbance increases linearly with a COIN's concentration and so a spectrum from a mixture of COINs is reasonably modeled by a linear combination of pure COIN spectra. The data here come from experiments that were designed to establish the ability of these COINs to measure the existence and abundance of antigens in single-cell assays.

Let $P_1, \ldots, P_{10}$ denote spectra from nine pure COINs and one "blank" (no biochemical material), each normalized to norm one. We form in-silico mixtures as follows: $x_i = \sum_{k=1}^{10} c_{i,k} P_k,\ i = 1, \ldots, n$, $i = 1, \ldots, n$, with coefficients $\{c_{i,k}\}$ generated from a uniform distribution. Figure 2 shows representative spectra from all nine COINs superimposed on a collection of mixture spectra, $\{x_i\}_{i=1}^{50}$. Included in Figure 2 is the β (dashed curve) used to defined the simulation: $y_i = \langle x_i, \beta \rangle + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$.

In this simulation, we have created a coefficient function which, instead of being modeled mathematically, is a curve that exhibits structure of the type found in Raman spectra. Details on the construction of this β are in Appendix 9.1 so here we simply note that it arises as a ridge estimate from a set of in-silico mixtures of Raman spectra in which one COIN, P_9, is varied prominently relative to the others. See Figure 2. Motivation for defining β in this way is based on a view that it seems implausible for us to predict the structure of realistic signal in these data and recreate it using polynomials, Gaussians or other analytic functions.

Regardless of its construction, β defines signal that allows us to compute estimation and prediction error. The performances of five methods are summarized in Table 3. Note that although β was constructed as a ridge estimate (using a different set of in-silico mixtures; see Appendix 9.1), the ridge penalty is not necessarily optimal for recovering β. This is because the strictly empirical eigenvectors associated with the new spectra may contain structure not informative regarding y. Also, in these data, the performance of FPCR$_R$ is adversely affected by a tendency for the estimate to be smooth; cf., Figure 3. The PEER penalty used here is defined by a decomposition-based operator (16) in which $\mathcal{Q}$ is spanned by a 10-dimensional set of pure-COIN spectra (including a blank). The success of such an estimate obviously depends on an informed formation of $\mathcal{Q}$, but as long as the parameter-selection procedure allows for $a = b$, then the set of possible estimates includes ridge as well as estimates with potentially lower MSE than ridge; see Proposition 5.4.

We note that this simulation may be viewed as inherently unfair since the PEER estimate uses knowledge about the relevant structure. However, this is a point worth reemphasizing: when prior knowledge about the structure of the data is available, it can be incorporated naturally into the regression problem.

7.3. Raman application

We now consider spectra representing true antibody-conjugated COINs from nine laboratory mixtures. These mixtures contain various concentrations of eight COINs (of the nine shown in Figure 2). Spectra from four technical replicates in each mixture are included to create a set of $n = 36$ spectra $\{x_i\}_{i=1}^{n}$. We designate P_1 as the COIN whose concentration within each mixture defines y. Assuming a linear relationship between the spectra, $\{x_i\}$, and the P_1-concentrations, $\{y_i\}$, we estimate P_1. More precisely, we estimate the structure in P_1 that correlates most with its concentrations, as manifest in this set of mixtures. The fLM is a simplistic model of this relationship between the concentration of P_1 and its functional structure, but the physics of this technology imply it is a reasonable starting point.

We present the results of three estimation methods: ridge, FPCR$_R$ and PEER. In constructing a PEER penalty, we note that the informative structure in Raman spectra is not that of low-frequency or other easily modeled features, but it may be obtainable experimentally. Therefore, we define L as in (13) in which $\mathcal{Q}$ contains the span of COIN template spectra: $\mathcal{Q}_1=\text{span}\{P_k\}_{k=1}^{8}$. However, since a single set of templates may not faithfully represent signal in subsequent experiments (with new measurement errors, background and baseline noise etc), we enlarge $\mathcal{Q}$ by adding additional structure related to these templates. For this, set $\mathcal{Q}_2=\text{span}\{P_k', P_k''\}_{k=1}^{8}$, where P_k' denotes the derivative of spectrum P_k. (Note, to form $\mathcal{Q}_1$, scale-based approximations to these derivatives are used since raw differencing of non-smooth spectra introduces noise.) Then set $\mathcal{Q}$ = span{ $\mathcal{Q}_1 \cup \mathcal{Q}_2$} and define $L = a(I - P_{\mathcal{Q}}) + b\, P_{\mathcal{Q}}$.

The regularization parameters in the PEER and ridge estimation processes were chosen using REML, as described in Section 6. For the FPCR$_R$ estimate, we used the R-package refund [39] as implemented in [40].

Since β is not known (the model $y = X\beta + \varepsilon$ is only approximate), we cannot report MSEs for these three methods. However, the structure of P_1 is qualitatively known and by experimental design, y is directly associated with P_1. The goal here is that of extracting structure of the constituent spectral components as manifest in a linear model. This application is similar to the classic problem of multivariate calibration [5, 31] which essentially leads to a regression model using an experimentally-designed set of spectra from laboratory mixes.

The structure in the estimate here is expected to reflect the structure in P_1 that is correlated with P_1's concentrations, y. The estimate is not, however, expected to precisely reconstruct P_1 since P_1 shares structure with the other COIN spectra not associated with y. See Figure 2 where P_1 is plotted alongside the other COIN spectra. Now, Figure 3 shows plots of the PEER, FPCR$_R$ and ridge estimates of the fLM coefficient function. The PEER estimate, $\tilde{\beta}_Q$, provides an interpretable compromise between ridge, which involves no smoothing, and FPCR$_R$, which appears to oversmooth. For reference, the P_1 spectrum is also plotted along with a mean-adjusted version of $\tilde{\beta}_Q$, $\tilde{\beta}_Q + \mu$ (dashed line), where $\mu(t) = (1/36)\, \Sigma_i\, x_i(t)$, $t \in [400, 1800]$.

Finally, we consider prediction for these methods by forming a new set of spectra from different mixture compositions (different concentrations of each COIN) and, additionally, taken from different batches. This “test” set consists of spectra from four technical replicates in each of 15 mixtures forming a set of $n = 60$ spectra, $\{x_i^{\text{test}}\}_{i=1}^{n}$. As before, P_1 is the COIN whose concentration within each mixture defines the values $\{y_i^{\text{test}}\}_{i=1}^{n}$. For the estimates from each of the three methods (shown in Figure 3) we compute the prediction error: $(1/n)\sum_i (y_i^{\text{test}} - \langle x_i^{\text{test}}, \tilde{\beta}\rangle)^2$. The errors for PEER, ridge, and FPCR$_R$ estimates are 0.770, 0.752, 2.139, respectively. The ridge estimate here illustrates how low prediction error is not necessarily accompanied by interpretable structure in the estimate (or low MSE) [7].

8. Discussion

As high-dimensional regression problems become more common, methods that exploit a priori information are increasingly popular. In this regard, many approaches to penalized regression are now founded on the idea of “structured” penalties which impose constraints based on prior knowledge about the problem’s scientific setting. There are many ways in which such constraints may be imposed, and we have focused on the algebraic aspects of a penalization process that imposes spatial structure directly into a regularized estimation. This approach fits into the classic framework of L^2-penalized regression but with an emphasis on the algebraic role that a penalty operator plays to impart structure on the estimate.

The interplay between a structured regularization term and the coefficient-function estimate may not be well understood in part because it is not typically viewed in terms of the generalized singular vectors/values, which is fundamental to this investigation. In particular, any penalized estimate of the form (1) with $L \neq I$ is intrinsically based on GSVD factors in the same way that many common regression methods (such as PCR, ridge, James-Stein, or partial least squares) are intrinsically based on SVD factors. Just as the basics of the ubiquitous SVD are important to understanding these methods, we have aspired to established the basics of the GSVD as it applies to a this general penalized regression setting and to illustrate how the theory underlying this approach can be used inform the choice of penalty operator.

Toward this goal the presentation emphasizes the transparency provided by the partially-empirical eigenvector expansion (7). Properties of the estimate’s variance and bias are determined explicitly by the generalized singular vectors whose structure is determined by the penalty operator. We have restricted attention to additive constraints defined by penalty operators on L^2 in order to retain the direct algebraic connection between the eigenproperties of the operator pair (X, L) and the spatial structure of $\tilde{\beta}_{\alpha,L}$. Intuitively, the structure of the penalty’s least-dominant singular vectors should be commensurate with the informative structure of β. The actual effect a penalty has on the properties of the estimate can be quantified in terms of the GSVD vectors/values.

This perspective differs from popular two-stage signal regression methods in which estimation is either preceded by fitting the predictors to a set of (external) basis functions or is followed by a step that smooths the estimate [8, 21, 30, 38, 40]. Instead, structure (smoothness or otherwise) is imposed directly into the estimation process. The implementation of a penalty that incorporates structure less generic than smoothness (or sparseness) requires some qualitative knowledge about spatial structure that is informative. Clearly this is not possible in all situations, but our presentation has focused on how a functional linear model may provide a rigorous and analytically tractable way to take advantage of such knowledge when it exists.

Acknowledgments

The authors wish to thank three anonymous referees and the associate editor for several corrections and suggestions that greatly improved the paper. We also gratefully acknowledge the patience, insight and assistance from Dr. C. Schachaf who provided the Raman spectroscopy data for the examples in Sections 7.2 and 7.3. Madan Gopal Kundu carefully read the manuscript and corrected several errors. Research support was provided by the National Institutes of Health grants R01-CA126205 (TR), P01-CA053996 (TR, ZF), U01-CA086368 (TR, ZF), R01-NS036524 (JH) and U01-MH083545 (JH).

References

1. Belge M, Kilmer ME, Miller EL. Wavelet domain image restoration with adaptive edge-preserving regularization. IEEE Transactions On Image Processing. 2000; 9(4):597–608. [PubMed: 18255433]
2. Bertero, M.; Boccacci, P. Introduction to Inverse Problems in Imaging. Institute of Physics; Bristol, UK: 1998. MR1640759
3. Bingham C, Larntz K. A simulation study of alternatives to ordinary least squares: Comment. Journal of the American Statistical Association. 1977; 72(357):97–102.
4. Björck, Å. Numerical Methods for Least Squares Problems. SIAM; Philadelphia: 1996.
5. Brown, P. Measurement, Regression and Calibration. Oxford University Press; Oxford, UK: 1993. MR1300630
6. Brumback BA, Ruppert D, Wand MP. Comment on “Variable selection and function estimation in additive nonparametric regression using a data-based prior”. Journal of the American Statistical Association. 1999; 94:794–797.
7. Cai TT, Hall P. Prediction in functional linear regression. The Annals of Statistics. 2006; 34(5): 2159–2179. MR2291496.
8. Cardot H, Ferraty F, Sarda P. Spline estimators for the functional linear model. Statistica Sinica. 2003; 13:571–591. MR1997162.
9. Craven P, Wahba G. Smoothing noisy data with spline functions. Numerische Mathematik. 1979; 31:377–403. MR0516581.
10. Dempster AP, Schatzoff M, Wermuth N. A simulation study of alternatives to ordinary least squares. Journal of the American Statistical Association. 1977; 72(357):77–912.
11. Eldén L. A weighted pseudoinverse, generalized singular values, and contstrained least squares problems. BIT. 1982; 22:487–502. MR0688717.
12. Engl, HW.; Hanke, M.; Neubauer, A. Regularization of inverse problems. Kluwer; Dordrecht, Germany: 2000.
13. Golub, GH.; Van Loan, C. Matrix computations. Johns Hopkins University Press; Baltimore: 1996. MR1417720
14. Goutis C. Second-derivative functional regression with applications to near infrared spectroscopy. Journal of the Royal Statistical Society B. 1998; 60(1):103–114. MR1625656.
15. Groetsch, CW. Research Notes in Mathematics. Vol. 105. Pitman; Boston, MA: 1984. The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind. MR0742928
16. Hall P, Horowitz JL. Methodology and convergence rates for functional linear regression. The Annals of Statistics. 2007; 35(1):70–91. MR2332269.
17. Hall P, Poskitt DS, Presnell B. A functional data-analytic approach to signal discrimination. Technometrics. 2001; 43(1):1–9. MR1847775.
18. Hansen PC. Perturbation bounds for discrete Tikhonov regularisation. Inverse Problems. 1989; 5:L41–L44. MR1009032.
19. Hansen, PC. Rank-Deficient and Discrete Ill-Posed Problems. SIAM; Philadelphia, PA: 1998. MR1486577
20. Hastie T, Buja A, Tibshirani R. Penalized discriminant analysis. The Annals of Statistics. 1995; 23(1):73–102. MR1331657.
21. Hastie T, Mallows C. Discussion of “A statistical view of some chemometrics regression tools”. Technometrics. 1993; 35:109–148.
22. Heckman NE, Ramsay JO. Penalized regression with model based penalties. Canadian Journal of Statistics. 2000; 28:241–258. MR1792049.
23. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1):55–67.
24. Huang JZ, Shen H, Buja A. Functional principal components analysis via penalized rank one approximation. Electronic Journal of Statistics. 2008; 2:678–695. MR2426107.
25. Kilmer ME, Hansen PC, Español MI. A projection-based approach to general-form Tikhonov regularization. Siam J Sci Comput. 2007; 29(1):315–330. MR2285893.

26. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008; 24(9):1175. [PubMed: 18310618]
27. Lutz BR, Dentinger CE, Nguyen LN, Sun L, Zhang J, Allen AN, Chan S, Knudsen BS. Spectral Analysis of Multiplex Raman Probe Signatures. ACS nano. 2008; 2(11):2306–2314. [PubMed: 19206397]
28. MacLeod AJ. Finite-dimensional regularization with nonidentity smoothing matrices. Linear Algebra and its Applications. 1988; 111:191–207. MR0974054.
29. Marquardt DW. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. Technometrics. 1970; 12(3):591–612.
30. Marx BD, Eilers PHC. Generalized linear regression on sampled signals and curves: A P-spline approach. Technometrics. 1999; 41(1):1–13. MR2135789.
31. Marx BD, Eilers PHC. Multivariate calibration stability: a comparison of methods. Journal of Chemometrics. 2002; 16(3):129–140.
32. Müller HG. Functional modelling and classification of longitudinal data. Scandinavian Journal of Statistics. 2005; 32:223–240. MR2188671.
33. Neumaier A. Solving ill-conditioned and singular linear systems: A tutorial on regularization. SIAM Review. 1998; 10(3):636–666. MR1642811.
34. O'Brien DM, Holt JN. The extension of generalized cross-validation to a multi-parameter class of estimators. Austral Math Soc (Series B). 1981; 22:501–514.
35. Paige CC, Saunders MA. Towards a generalized singular value decomposition. SIAM J Numerical Analysis. 1981; 18(3):398–405. MR0615522.
36. Phillips DL. A technique for the numerical solution of certain integral equations of the first kind. Journal of the ACM. 1962; 9(1):84–97. MR0134481.
37. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2011.
38. Ramsay, JO.; Silverman, BW. Functional Data Analysis. Springer-Verlag; New York: 2005. MR2168993
39. Reiss, Philip; Huang, Lei; Goldsmith, Jeff. R package version 0.1-2. 2010. refund: Regression with functional data.
40. Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. Journal of the American Statistical Association. 2007; 102(479):984–986. MR2411660.
41. Reiss PT, Ogden RT. Smoothing parameter selection for a class of semiparametric linear models. Journal of the Royal Statistical Society B. 2009; 71(2):505–523. MR2649608.
42. Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric regression. Cambridge University Press; New York: 2003. MR1998720
43. SAS Institute Inc. SAS/STAT software, version 9.2. Cary, NC: 2008.
44. Shachaf CM, Elchuri SV, Koh AL, Zhu J, Nguyen LN, Mitchell DJ, Zhang J, Swartz KB, Sun L, Chan S, et al. A Novel Method for Detection of Phosphorylation in Single Cells by Surface Enhanced Raman Scattering (SERS) using Composite Organic-Inorganic Nanoparticles (COINs). PLoS ONE. 2009; 4(4)
45. Silverman BW. Smoothed functional principal components analysis by choice of norm. The Annals of Statistics. 1996; 24:1–24. MR1389877.
46. Slawski M, Zu Castell W, Tutz G. Feature selection guided by structural information. Annals of Applied Statistics. 2010; 4(2):1056–1080. MR2758433.
47. Speed T. Comment on "that BLUP is a good thing: The estimation of random effects", by G.K. Robinson. Statistical Science. 1991; 6(1):42–44. MR1108815.
48. Stout F, Kalivas JH. Tikhonov regularization in standardized and general form for multivariate calibration with application towards removing unwanted spectral artifacts. Journal of Chemometrics. 2006; 20:22–33.
49. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society B. 2005; 67(1):91–108. MR2136641.
50. Tibshirani RJ, Taylor J. The solution path of the generalized lasso. The Annals of Statistics. 2011; 39(3):1335–1371. MR2850205.

51. Tikhonov AN. On the stability of inverse problems. Dokl Akad Nauk SSSR. 1943; 39:176–179. MR0009685.

52. Tutz G, Ulbricht J. Penalized regression with correlation-based penalty. Statistics and Computing. 2009; 19(3):239–253. MR2516217.

53. Varah JM. A practical examination of some numerical methods for linear discrete ill-posed problems. SIAM Review. 1979; 21(1):100–111.

54. Wahba G. Spline models for observational data. Society for Industrial Mathematics. 1990; 59 MR1045442.

55. Wood, SN. An introduction to generalized additive models with R. Chapman and Hall; 2006. MR2206355

9. Appendix

9.1. Defining β for the simulation in Section 7.2

This simulation is motivated by an interest in constructing a plausibly realistic β whose structure is naturally derived by the scientific setting involving Raman signatures of nanoparticles. Although one could model a β mathematically using, say, polynomials or Gaussian bumps (cf., Appendix A.2), such a simulation would be detached from the physical nature of this problem. Instead, we construct a coefficient function that genuinely comes from a functional linear model with Raman spectra as predictors.

We first generate in-silico mixtures of COIN spectra as $x_i^o=\sum_{k=1}^{9} c_{i,k}P_k, i=1,\ldots,50$, where $c_{i,k}$ ~ unif[0, 1]. Designating P_9 as the COIN of interest, we define response values that correspond to the "concentration" of P_9 by setting $y_i^o:=3\,c_{i,9}$, $i = 1, \ldots, n$. The factor of 3 imposes a strong association between P_9 and the response.

Now, the example in section 7.2 aims to estimate a coefficient function that truly comes from a solution to a linear model. However, the equation $y^o = X^o\beta$ has infinitely many solutions (where X^o is the matrix whose ith row is x_i^o), so we must we must regularize the problem to obtain a specific β. For this, we simply use a ridge penalty and designate the resulting solution to be β. This is shown by the dotted curve in Figure 2 and is qualitatively similar to P_9.

We note that the simulation in section 7.2 uses the same set of COINs, but a new set of in-silico mixture spectra (i.e., a new set of $\{c_{i,k}\}$ ~ Unif[0, 1]). In addition, a small amount of measurement error was added, as in section 7.1, to each spectrum during the simulation.

9.2. Frequency domain simulation

We display results from a study that mimics the scenario of simulations studied by Hall and Horowitz [16]. We illustrate, in particular, properties of the MSE discussed following equation (18) in section 5.3 relating to $b = 0$. In fact, we consider the more general scenario in which $\mathcal{Q}$ is not constructed from empirical eigenvectors (as in PCR and ridge), but is defined by a prespecified envelope of frequencies.

In this simulation both β and x_i, $i = 1, \ldots, n$, are generated as sums of the cosine functions

$$x_i(t)=\sum_{j=1}^{40}\gamma_j Z_{ij}\varphi_j(t)+e_i(t), \qquad \beta(t)=0.75\varphi_5(t)+1.5\varphi_{11}(t)+1\varphi_{17}(t),$$

$t \in [0, 1]$; here $\gamma_j = (-1)^{j+1} j^{-0.75}$, Z_{ij} is uniformly distributed on $[-3^{1/2}, 3^{1/2}]$ ($E(Z_{ij}) = 0$ and $\text{var}(Z_{ij}) = 1$), $\varphi_1 \equiv 1$ and $\varphi_j(t) = 2^{1/2} \cos(j\pi t)$ for $j \geq 1$, and $e_i(t) \sim N(0, \sigma_X^2)$, and $\text{cov}(e_i(t), e_{i'}(t')) = 0$ for either $i \neq i'$ or $t \neq t'$. The response y_i is defined as $y_i = \langle \beta, x_i \rangle + \varepsilon_i$, where $\varepsilon_i \sim N(0\sigma^2)$, i.i.d.. The simulations involve discretizations of these curves evaluated at $p = 100$ equally spaced time points, t_j, $j = 1, \ldots, p$, that are common to all curves.

Penalties

We consider properties of estimates from a variety of penalties: ridge ($L = I$), $\mathcal{D}^2$, $\mathcal{D}^2 + aI$, and PCR[1]. In addition, targeted penalties of the form $L = I - P_{\mathcal{Q}}$, are defined by the specified subspaces $\mathcal{Q} = \text{span}\{\varphi_j\}_{j \in J}$, for φ_j defined above. Specifically, we use $J = J_F = \{j = 5, \ldots, 17\}$ (a tight envelope of frequencies) to define L_F, and $J = J_G = \{j = 4, \ldots, 20\}$ (a less focused span of frequencies) to define L_G. The operator $\mathcal{D}^2 + aI$ simply serves to illustrate the role of higher-frequency singular vectors as discussed in Section 4.1. In the simulations, the coefficient a in $\mathcal{D}^2 + aI$ was chosen simultaneously with α via a two-dimensional grid search.

Simulation results

Table 4 summarizes estimation results for all six penalties and two sample sizes, $n = 50$, 200. The prediction results for these estimates are in Table 5. These are reported for $S/N = 10$, 5 and $R^2 = 0.8$, 0.6. The number of terms in the PCR estimate was optimized and ranged from 19 to 25 when $R^2 = 0.8$ and decreased with decreasing R^2. Analogously, one could optimize over the dimension of $\mathcal{Q}$ (to implement a truncated GSVD), but the purpose here is illustrative while in practice a more robust approach would emply a penalty of the form (13).

Errors obtained with ridge and PCR are small, as expected, since the structure of β in this example is consistent with the structure represented in the singular vectors, v_k. Therefore, even though the relationship between the y_i and x_i degrades (indeed, even as $R^2 \to 0$), these estimates are comprised of vectors that generally capture structure in β since it is strongly represented by the dominant eigenstructure of X. The second-derivative penalty, $\mathcal{D}^2$, produces the worst estimate in each of the scenarios due to oversmoothing. Note $\mathcal{D}^2 + a\,I$ improves on $\mathcal{D}^2$, yet it is still not optimal for the range of frequencies in β.

Regarding L_G, the MSE gets worse as S/N increases. Indeed, here $\mathcal{Q}$ is fixed and relatively large and since the σ_k decay faster when S/N is big, this leads to rank deficiency and large variance; see equation (18) (note, this only applies approximately since $\mathcal{Q}$ does not consist of ordinary SVs). In our previous examples, this is stabilized by a $b > 0$.

The problems of estimation and prediction have different properties [7]; good prediction may be obtained even with a poor estimate, as seen in Table 5. The estimate from L_{D_a} is generally poor relative to others (as measured by the L^2-norm), but its prediction error is comparable with other methods and is best among the non-targeted penalization methods. This is consistent with the outcome described by C. Goutis [14] where (derivatives of) the predictor curves contain sharp features and so standard least-squares regularization (OLS, PCR, ridge, etc.) perform worse than a PEER estimate which imposes a greater emphasis on "regularly oscillatory but not smooth components"; see section 4.1.

[1] PCR is not obtained explicitly from a penalty operator, but see Corollary 5.3.

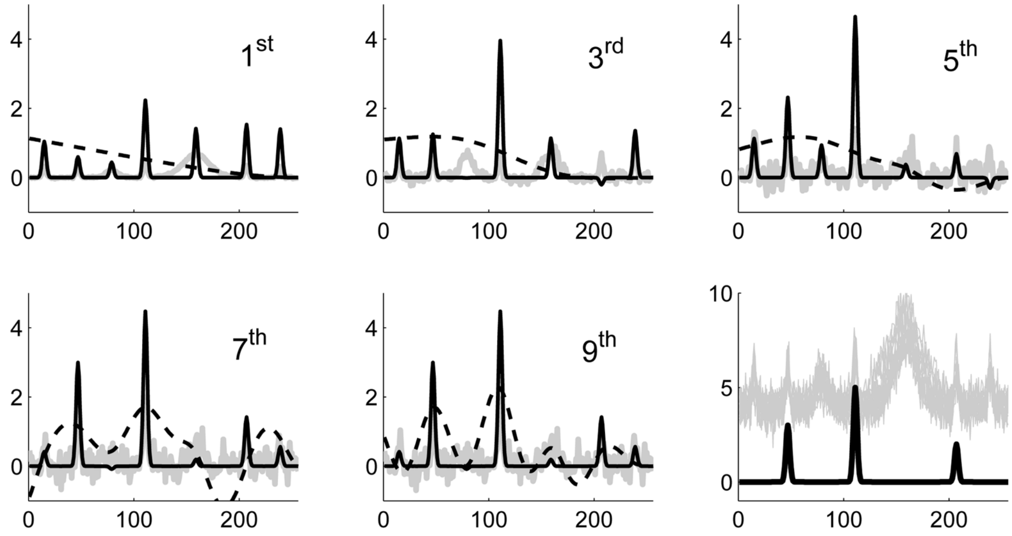

Fig 1. Partial sums of penalized estimates

The first five odd-numbered partial sums from (7) for three penalties: 2nd-derivative (dashed), ridge (gray), targeted PEER (black; see text in sections 2.1 and 7.1). The last panel displays β (black) and 15 predictors, x_i (gray), from the simulation.

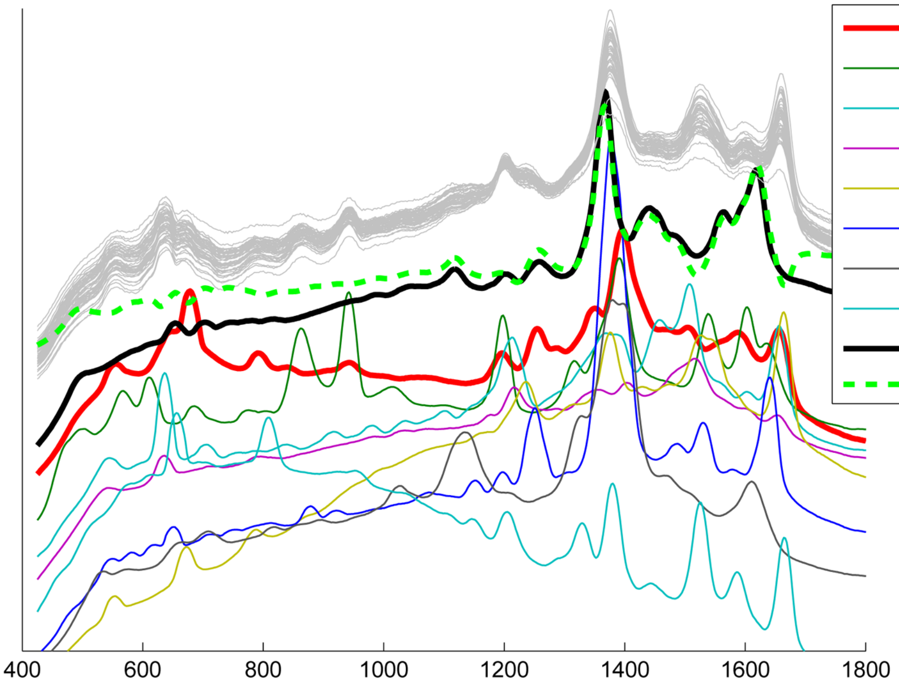

Fig 2.

Nine pure COIN spectra, $P_1, \ldots, P_9$, and a coefficient function, β (each shifted for display). β arises as a solution to the fLM in which y denotes concentrations of P_9 in an in silico mixture of 50 COIN spectra, x_i (light gray). This β is used in the simulation study of Section 7.2.

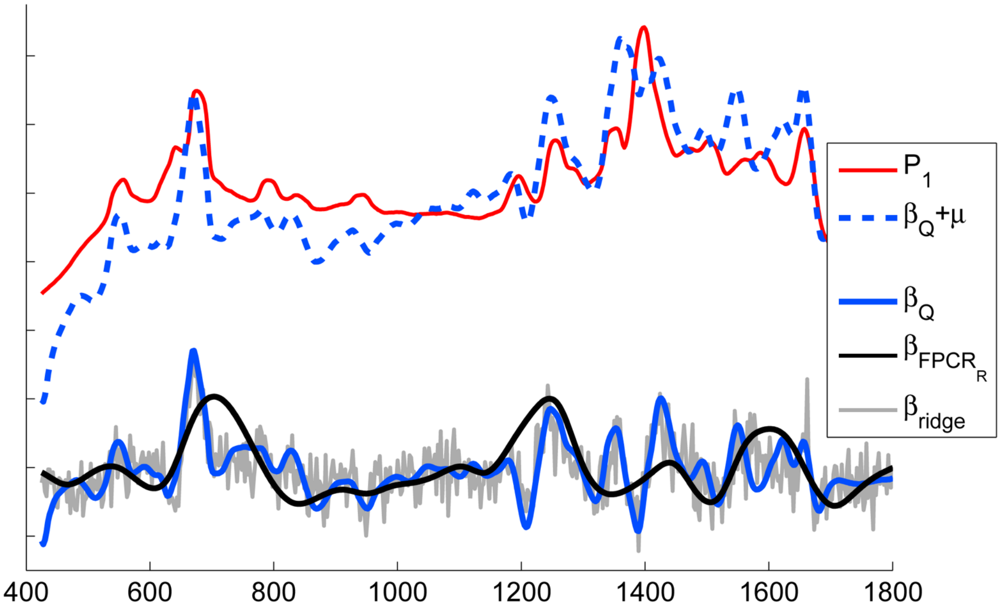

Fig 3.

Three estimates for a coefficient function that relates concentrations of P_1 to its signal in 8-COIN laboratory mixtures. Estimates shown: ridge ($\tilde{\beta}_{\text{ridge}}$; gray), $FPCR_R$ ($\tilde{\beta}_{FPCR_R}$; black) and PEER ($\tilde{\beta}_Q$; blue). For perspective, P_1 is plotted (in red) and the mean-adjusted PEER estimate, $\tilde{\beta}_Q + \mu$ (dashed blue); μ is the mean of the mixture spectra $\{x_i\}_{i=1}^{36}$ (not shown).

Table 1

Estimation errors (MSE) for simulation with selected bump locations. Sample size is $n = 50$

R^2	*S/N*	L_V	L_U	**PCR**	**ridge**	$\mathcal{D}^2$	$\mathcal{D}^2 + a\,I$
0.8	10	4.00	13.81	9.38	34.39	359.83	76.31
0.8	5	3.72	15.46	21.50	40.02	246.17	72.64
0.8	2	4.40	12.96	57.89	58.22	126.75	59.35
0.6	10	9.60	21.60	14.10	50.50	497.70	113.60
0.6	5	10.22	21.65	26.02	50.68	338.70	87.58
0.6	2	11.75	23.18	63.50	67.94	181.75	78.45

Table 2

Prediction errors for simulation with selected bump locations. Sample size is $n = 50$. Errors are multiplied by 1000 for display

R^2	S/N	L_V	L_U	PCR	ridge	$\mathcal{D}^2$	$\mathcal{D}^2 + a I$
0.8	10	9.0	10.5	10.8	16.6	19.3	12.9
0.8	5	8.4	11.0	12.2	26.7	27.9	17.8
0.8	2	12.9	19.0	53.2	55.7	50.3	40.1
0.6	10	21.4	23.0	23.9	33.0	39.0	26.2
0.6	5	23.9	25.0	29.5	49.2	54.6	34.4
0.6	2	34.4	42.5	90.4	110.4	104.4	77.9

Table 3

Estimation (MSE) and prediction (PE) errors of several penalization methods for the simulation described in Figure 2. Numbers represent the average error from 100 runs. PE errors are multiplied by 1000 for display

	L_Q	**PCR**	**ridge**	$\mathcal{D}^2 + aI$	**FPCR**$_R$
MSE	8.91	12.34	13.87	41.69	15.29
PE	0.0071	0.0179	0.0139	0.0131	0.0175

Table 4

Estimation errors (MSE) for the simulation with localized frequencies

n	R^2	S/N	L_F	L_G	**PCR**	**ridge**	$\mathcal{D}^2$	$\mathcal{D}^2 + a\,I$
50	0.8	10	42.42	77.31	123.60	132.50	1051.20	568.45
50	0.8	5	41.55	75.41	128.75	143.48	447.64	184.07
200	0.8	10	8.28	13.48	33.44	65.78	453.54	169.41
200	0.8	5	8.56	13.08	36.36	87.59	100.15	76.24
50	0.6	10	106.89	200.08	173.56	173.94	1098.20	631.05
50	0.6	5	109.51	178.05	178.15	196.62	612.12	259.92
200	0.6	10	25.30	38.73	58.90	98.13	847.59	350.46
200	0.6	5	22.08	33.79	59.92	119.48	240.09	127.52

Table 5

Prediction errors for the simulation with localized frequencies

n	R^2	S/N	L_F	L_G	PCR	ridge	$\mathcal{D}^2$	$\mathcal{D}^2 + aI$
50	0.8	10	0.848	1.134	1.490	1.423	1.292	1.246
50	0.8	5	0.840	1.124	1.427	1.390	1.304	1.222
200	0.8	10	0.200	0.273	0.432	0.497	0.444	0.418
200	0.8	5	0.211	0.276	0.460	0.547	0.466	0.455
50	0.6	10	2.165	2.900	3.051	2.705	2.621	2.472
50	0.6	5	2.171	2.832	3.158	2.938	2.912	2.724
200	0.6	10	0.621	0.801	1.058	1.160	1.044	0.990
200	0.6	5	0.584	0.766	1.062	1.186	1.069	1.014