Published in final edited form as:

J Chemom. 2017 April ; 31(4): . doi:10.1002/cem.2850.

Adaptive penalties for generalized Tikhonov regularization in statistical regression models with application to spectroscopy data

Timothy W. Randolph^{†,*}, Jimin Ding[#], Madan G. Kundu[‡], and Jaroslaw Harezlak[‡]

[†]Fred Hutchinson Cancer Research Center, Biostatistics and Biomathematics, Seattle, WA 98109

[#]Washington University in Saint Louis, Department of Mathematics, Saint Louis, MO 63130

[‡]Novartis Pharmaceuticals Corporation, Oncology, East Hanover, NJ 07936

[‡]Indiana University, Epidemiology and Biostatistics, Bloomington, IN 47405

Abstract

Tikhonov regularization was proposed for multivariate calibration by Andries and Kalivas [1]. We use this framework for modeling the statistical association between spectroscopy data and a scalar outcome. In both the calibration and regression settings this regularization process has advantages over methods of spectral pre-processing and dimension-reduction approaches such as feature extraction or principal component regression. We propose an extension of this penalized regression framework by adaptively refining the penalty term to optimally focus the regularization process. We illustrate the approach using simulated spectra and compare it with other penalized regression models and with a two-step method that first pre-processes the spectra then fits a dimension-reduced model using the processed data. The methods are also applied to magnetic resonance spectroscopy data to identify brain metabolites that are associated with cognitive function.

Keywords

penalized regression; Tikhonov regularization; calibration; adaptive penalty; generalized singular value decomposition

1 Introduction

This paper provides a statistical perspective on regression models for associating a vector of spectroscopy data, $\mathbf{x}$, with a scalar outcome, y , such as a patient's disease status or phenotype. In the case of a continuous outcome, a multiple linear regression model is of the form $y = \beta^T \mathbf{x} + \varepsilon$, where, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ denotes the random error corresponding to $y - \beta^T \mathbf{x}$; i.e., $E(y) = \beta^T \mathbf{x}$, where β is the (unknown) $p \times 1$ coefficient vector whose k^{th} coordinate models the association of the corresponding coordinate in $\mathbf{x}$ ($k = 1, \dots, p$) with the observed outcome y . Or, in matrix notation, this takes the form $y = \mathbf{X}\beta + \varepsilon$, where $\mathbf{X}$ denotes an $n \times p$

* trandolp@fredhutch.org.

matrix whose i^{th} row is the p -vector of spectroscopy measures, y is the $n \times 1$ vector of observed outcomes, and ε the vector of model errors. If, for example, a small set of $p < n$ analyte features has been extracted, then the ordinary least-squares (OLS) solution $\hat{\beta}_{\text{OLS}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$ is the minimum variance unbiased estimate for the linear association between y and x . In this paper we describe the estimation of β via a process that does not first undertake a wavelength- or variable-selection step, instead forming a model using the full spectra in which, typically, $p \gg n$, and hence has infinitely many solutions.

The topic of regularizing ill-posed linear systems has a long history and is well studied in a variety of contexts, including image analysis [3], statistics [26, 13], mathematics [25, 11, 9], and calibration [5, 15], to name a few. Our approach is very similar in spirit to calibration methods of Andries and Kalivas [1] (hereafter referred to as A&K) but with several important conceptual differences. In particular, rather than focusing on spectral calibration we formulate a statistical framework aimed at estimating the potential association between a set of subject-specific outcomes (something external to the collection of spectra, such as disease status) with the cohort's set of spectroscopy data. There may, in fact, not be any association (the null hypothesis) but in the event that there is, it is useful to further identify *which* spectral properties (wavelengths or analyte features) are associated. That is, our primary goal is the estimation of β rather than the prediction of y . With this in mind, we have taken care to establish notation that distinguishes between a statistical multiple regression model and the conceptually similar problem of multivariate calibration. In the latter, $\{y_i\}_{i=1}^n$ denotes a set of known, non-random analyte concentrations corresponding to calibration spectra $\{x_i\}_{i=1}^n$. Using notation analogous to the above, this may be represented in matrix form as $\mathbf{Xb} = y + e$, where the aim is to estimate a model vector, b , which may then be used to predict analyte concentrations in future samples. In this setting, e represents a vector of unmodeled residuals where the source of error comes from X (as opposed to y , above). We note that in either setting, the minimum-norm least-squares estimate (of b or β) can be obtained from among the infinitely many solutions [11]. This solution, however, can be highly unstable and even useless in a statistical setting when $\mathbf{X}^T \mathbf{X}$ is singular or ill-conditioned.

In summary, our presentation may be viewed as an offshoot of A&K in two ways. First, rather than approximating a model vector, b , in a calibration setting we focus on estimation of a coefficient function, β , in a statistical regression model for a biomedical study (section 2). Second, although we perform the estimation of β within the same Tikhonov regularization framework used by A&K, our penalty term serves to focus the estimation process *toward* a "favored subspace" of relevant signal rather than *filtering out* signal in the "interferent space" (Section 3.2). Finally, we propose a novel approach to the estimation of β by adaptively refining the favored space.

Conceptually, the favored space $\mathcal{Q} \subset \mathbb{R}^p$ contains the "true" β and is penalized lightly while its orthogonal complement is penalized heavily. Note, however, if $\mathcal{Q}$ is large and underpenalized, the variance of the estimate will be too large and so we develop the concept of an adaptive penalty to account for the fact that one cannot know, a priori, the precise

subspace of features in x that are associated with y . Therefore, we describe an estimation process that begins with a large favored space and iteratively pares down the size of this space, converging to a more stable estimate.

In order to evaluate the potential reduction in variance by this proposed process, Section 2 reviews some basic concepts on variation in regression. Further, because there is some discord among the way terminology is used in calibration versus statistical regression, we provide a glossary of how several potentially conflicting terms are used. In Section 3, we review the concept of Tikhonov regularization as used in A&K and in our methods. Section 4 describes the details of our adaptive penalization approach while Section 5 provides the results of several simulations designed to illustrate the properties of this approach. Also included is an application of this method to a clinical study involving magnetic resonance spectroscopy data and their potential association with cognitive decline in HIV patients.

The following notation is used. Lower- and upper-case boldface letters are used to denote a vector ($\mathbf{a}$) or matrix ($\mathbf{A}$). Non-boldface letters denote a scalar or scalar-valued random variable, y . The transpose, inverse and generalized inverse of a matrix $\mathbf{A}$ are denoted by $\mathbf{A}^T$, $\mathbf{A}^{-1}$ and $\mathbf{A}^\dagger$, respectively. $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\|\mathbf{a}\|$ denotes the Euclidean 2-norm of $\mathbf{a}$.

2 Calibration versus regression models

Multivariate calibration involves spectral processing for the removal of non-analyte signal and quantification of analyte concentrations. Multiple regression in a biomedical setting ideally begins with well-calibrated measurements — e.g., from metabolites quantified in magnetic resonance spectroscopy (MRS) data — and aims to infer potential associations of analyte concentrations with a clinically relevant outcome. For this, raw spectra containing a mix of analyte and interferent signal may be pre-processed with the goal of removing interferent signal and the analyte concentrations calibrated for subsequent use as covariates in a multiple regression model (see, e.g., Provencher [20]). Our approach is different: rather than pre-processing the spectra to remove noise, the estimation of β is performed within a penalized regression framework [18] that focuses the estimation of β away from non-analyte signal and toward analyte signal. For this, pure analyte spectra $\{\mathbf{q}_k\}_{k=1}^m$ (forming the rows a new matrix $\mathbf{Q}$) are used in conjunction with the observed spectra in $\mathbf{X}$ to construct the regression estimate $\hat{\beta}$ in such a way that the associations between x and y (as represented by non-zero coefficients of $\hat{\beta}$) are informed by both x and the pure metabolite spectra. As described here (and noted by A&K), this use of both $\mathbf{X}$ and $\mathbf{Q}$ can be made explicit using the generalized singular value decomposition (GSVD); see Section 3. In the remainder of this section we establish terminology on quantifying uncertainty in estimating a regression coefficient vector, $\hat{\beta}$ [18].

2.1 Variance and bias in linear regression

While calibration models often focus on approximating the model vector, $\mathbf{b}$, a statistical linear regression model aims to estimate a coefficient vector, β , as well as estimating the conditional mean and conditional variance of the response, y , given the covariates, $\mathbf{X}$. For

example, the variability of the response in OLS is modeled by a separate parameter which is independent of the covariates and homogenous across observations. Specifically, $E(y|X) = X\beta$ and $V \equiv \text{Var}(y|X) = \sigma^2 I_n$, and so the goal is to estimate both the regression vector, β , and the variance of response, σ^2 . One may estimate β first and then estimate σ^2 under the assumption that the variance does not depend on X . This variance is essential for subsequent inference regarding $\hat{\beta}$. For example, when $n \gg p$, OLS is not only unbiased, $E(\hat{\beta}|X) = \beta$, but it has the smallest variance among all unbiased estimators, $\text{Var}(\hat{\beta}_{\text{OLS}}|X) = \sigma^2(X^T X)^{-1}$. More generally, when the observations y_i are not independent or have different variances, i.e., $\text{Var}(y|X) = V \neq \sigma^2 I_n$, then $\hat{\beta}_{\text{OLS}}$ no longer has the smallest variance, although it is still unbiased. In this case, a weighted least square estimate $\hat{\beta}_{\text{WLS}} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ is unbiased and has the smallest variance among all unbiased estimators, hence may be preferred over OLS.

The variance of $\hat{\beta}$ has both statistical and computational relevance. In situations for which $X^T X$ (or $X^T \hat{V} X$) is ill-conditioned or singular (e.g., $p > n$), the estimate $\hat{\beta}_{\text{OLS}}$ is, respectively, highly variable or undefined. This computational instability is a reflection of large variance, $\text{Var}(\hat{\beta}|X)$, which is illustrated in the examples of Section 5. A more stable, but biased, estimate may be preferred and ridge regression achieves this:

$\hat{\beta}_R = (X^T X + \lambda I)^{-1} X^T y = \arg \min_{\beta} \{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \}$. The tuning parameter λ controls the balance between the bias and the variance, $\text{bias}(\hat{\beta}_R) = (I - X^{\#} X^{\#T})\beta$ and $\text{Var}(\hat{\beta}_R) = \sigma^2 X^{\#} X^{\#T}$, respectively, where $X^{\#} = (X^T X + \lambda I)^{-1} X^T$.

This tradeoff appears in the mean square error (MSE) of the estimate:

$\text{MSE}(\hat{\beta}) = \|\text{bias}(\hat{\beta}_R)\|^2 + \text{trace}(\text{Var}(\hat{\beta}_R))$. Ridge regression will reduce the variance while increasing the bias. In contrast, while the penalty learning process proposed in Section 4 also aims to reduce the variance, it aims to do so with a smaller increase in bias. In our estimation process, as well as in ridge regression or the more general Tikhonov regularization discussed in the next section, the choice of tuning parameter λ is important. For this, one may use a subjective method, e.g. the trace plot [14], or a more objective method, such as minimizing cross-validation prediction error [10]. For this, and for the more general Tikhonov regularized estimate discussed in the next section, we will describe a random-effects model for obtaining an objective, model-based estimate of λ .

Finally, for evaluating the performance of our approach in the simulations, we are interested in the pointwise squared error, $(\beta_j - \hat{\beta}_j)^2, j = 1, \dots, p$. This can be summarized (pointwise) as: $E[(\beta_j - \hat{\beta}_j)^2] = \text{Bias}^2(\hat{\beta}_j) + \text{Var}(\hat{\beta}_j)$, which measures the accuracy and the precision of $\hat{\beta}$ at each coordinate. A vector summary of this is the “integrated” squared error, $\text{ISE}(\hat{\beta}) = \sum_j (\beta_j - \hat{\beta}_j)^2$. And, viewing $\hat{\beta}$ as a *random vector*, the mean of this is $E(\text{ISE}(\hat{\beta})) = E\|\beta - \hat{\beta}\|^2$, i.e., the $\text{MSE}(\hat{\beta})$, as defined above. In the interest of a robust

summary in the simulations, we will also consider the *median ISE*($\hat{\beta}$) and the interquartile range of *ISE*($\hat{\beta}$); see, e.g., [27, 7].

2.2 Terminology in calibration and regression

We have emphasized the role of variation in $\hat{\beta}$ because it is inherent in the estimation process for the statistical regression model $y = X\beta + \varepsilon$ in which y is a sample-specific trait that is assumed to be measured with error. In contrast, a calibration model $Xb = y + e$ assumes y is known physical property that is quantified without error. In particular, these models focus on different *sources* of error: in a calibration model X is random while in a (frequentist) regression model y is random and the measurements in X (the “variables”) are typically assumed to be measured without error. Consequently, when an estimate $\hat{\beta}$ is used to infer which measurements in X that are associated with y , the uncertainty in this estimation process is used to define statistical significance.

The *mathematical* framework for these two problems, however, are the same. The work of A&K, and references therein, provide valuable insight and mathematical clarity on the various perspectives and approaches used for calibration models. For additional perspective, see [16]. One of our goals is to provide parallel insight and options for the context of estimation in a statistical regression model. Because research on calibration and statistical regression often operate in parallel with limited crossover in the literature, the emphases and terminologies sometimes differ. Therefore, in order to facilitate communication it is useful to make explicit the definitions and concepts as they are used in this paper. This is done in Table 1.

3 Regularization methods in calibration and regression

We briefly review concepts specific to multivariate calibration, as discussed by A&K, who provide an excellent mathematical unification of various spectral processing methods: generalized net-analyte signal (GNAS), generalized least-squares (GLS), generalized Tikhonov regularization (GTR). A&K highlight the difference between the concepts of “spectral preprocessing” versus “inverse processing”. Spectral pre-processing refers to two-step methods in which the spectra in X are first pre-processed via post-multiplication, $\tilde{X} = XP$, and then the linear system $y = \tilde{X}b$ is solved. In contrast, inverse processing methods allow for concurrent spectral processing and model building.

More specifically, if $P = I - L^\dagger L$, where L is a matrix of interferent spectra — non-signal output corresponding to spectral interference — then XP is the projection of the calibration spectra onto the orthogonal complement of the interferent space. As shown by A&K, the methods of NAS, GNAS and GLS are distinguished simply by how this projection is weighted. A&K then proceed to define spectral processing via augmentation where the interferent operator, L , serves the role of penalty operator in GTR:

$$\hat{b}_{\text{GTR}} = \arg \min_b \left\{ \|Xb - y\|^2 + \lambda \|Lb\|^2 \right\}. \quad (1)$$

In particular, the filtering out of interferent signal is performed concurrently with the formation of the model vector $\hat{\mathbf{b}}$ via the GTR model (1). The terminology “inverse modeling” arises from the property that the model vector comes from a closed-form inverse expression of the form $\hat{\mathbf{b}} = \mathbf{P}^\dagger \mathbf{g}$, for some $\mathbf{g}$. A&K have shown how this is true for each of GNAS, GLS and GTR. GTR is of particular interest for us as one can show that expression in (1) is equivalent to $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{L}^\top \mathbf{L})^{-1} \mathbf{X}^\top \mathbf{y}$; see, e.g., [11].

3.1 The GSVD in Tikhonov regularization

As noted by A&K, there is a rigorous mathematical underpinning for this provided by the generalized singular value decomposition (GSVD). Recall, first, that a ridge estimate ($\mathbf{L} = \mathbf{I}$)

can be expressed explicitly in terms of the singular vectors of $\mathbf{X}$ as $\sum_{k=1}^n \left(\frac{\sigma_k}{\sigma_k^2 + \lambda^2} \right) \mathbf{u}_k^\top \mathbf{y} \mathbf{v}_k$,

where σ_k is the k -th largest singular value of $\mathbf{X}$, and $\mathbf{u}_k$ and $\mathbf{v}_k$ are the corresponding left and right singular vectors, respectively, assuming $\text{rank}(\mathbf{X}) = n$. Now, given $\mathbf{X}$ and $\mathbf{L}$, the GSVD assures a simultaneous diagonalization of these two matrices as

$$\begin{aligned} \mathbf{X} &= \mathbf{U} \mathbf{S} \mathbf{W}^{-1}, \quad \mathbf{S} = [\mathbf{0} \quad \mathbf{S}], \quad \mathbf{S} = \text{diag}\{\sigma_k\} \\ \mathbf{L} &= \mathbf{V} \mathbf{M} \mathbf{W}^{-1}, \quad \mathbf{M} = [\mathbf{M} \quad \mathbf{0}], \quad \mathbf{M} = \text{diag}\{\mu_k\}. \end{aligned} \quad (2)$$

where $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns, $\mathbf{S}$ and $\mathbf{M}$ are diagonal and $\mathbf{W}$ is nonsingular. By convention, the generalized singular values are ordered as: $0 < \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0, 1 > \mu_1 \geq \mu_2 \geq \dots \geq \mu_r > 0$, where $\sigma_k^2 + \mu_k^2 = 1, k = 1, \dots, r$; r depends on the rank of $\mathbf{X}$ and $\mathbf{L}$. Denote the columns of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ by $\mathbf{u}_k$, $\mathbf{v}_k$ and $\mathbf{w}_k$, respectively. In this convention for ordering the GS values and vectors, the last few columns of $\mathbf{W}$ span the subspace $\text{Null}(\mathbf{L}) = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{L}\mathbf{x} = \mathbf{0}\}$. If $\text{Null}(\mathbf{L}) = \{0\}$, these correspond to the smallest generalized singular values, μ_k . We set $d = \dim(\text{Null}(\mathbf{L}))$ and note that $\mu_k = 0$ for $k > n - d$. See [11] or [4] for details.

From now on we will focus on the estimation of β in the regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$, although all the properties discussed above apply to the mathematically equivalent process of approximating a calibration model vector $\mathbf{b}$ in (1). The $\mathbf{L}$ -penalized estimate can be expressed as a series in terms of the GSVD as

$$\begin{aligned} \hat{\beta}(\lambda, \mathbf{L}) &= \arg \min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\mathbf{L}\beta\|^2 \right\} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{L}^\top \mathbf{L})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= \sum_{k=1}^{n-d} \left(\frac{\sigma_k}{\sigma_k^2 + \lambda \mu_k^2} \right) \mathbf{u}_k^\top \mathbf{y} \mathbf{w}_k + \sum_{k=n-d+1}^n \mathbf{u}_k^\top \mathbf{y} \mathbf{w}_k. \end{aligned} \quad (3)$$

This is an expansion with respect to the generalized singular vectors which correspond to a new basis for the estimation process: the estimate is expressed in terms of the generalized singular vectors $\{\mathbf{w}_k\}$ determined jointly by $\mathbf{X}$ and $\mathbf{L}$. This property motivated the

terminology of “partially empirical eigenvectors for regression” (PEER) [22] in order to distinguish it from the purely empirical (from X) eigenvectors that make up the regression estimates from ordinary least squares, partial least squares, ridge and principal component regression. We emphasize that PEER is mathematically identical to GTR, but the two contexts in which they are applied—calibration versus regression—are actually addressing different questions and so we will adopt both terminologies here in order to distinguish between the two settings.

In Section 4 we introduce an adaptive version of this regularization process. For brevity, we will refer to this proposed method for obtaining a refined penalty L and its corresponding estimate $\hat{\beta}(\lambda, L)$ as “supervised PEER”, or SuPEER.

3.2 Two perspectives on penalized regularization

A&K discuss general approaches to removing interferent contributions from spectra and focus on the theoretical relationships between GNAS processing, GLS and GTR. As noted, the proposal for GTR is mathematically equivalent to our statistical modeling approach, but in addition to the different contexts of calibration versus regression, the premise is different: GTR aims to *filter out* non-analyte or interferent signal with a penalty term whereas we *focus the estimation process toward* analyte-specific signal. Moreover, A&K work in the context of calibration where y is known (presumably without error) and b is viewed as a deterministic vector whose true structure is inherent in X , albeit contaminated with some interferent structure. In particular, b is presumed to be a non-zero vector. Our goal, on the other hand, is ultimately to infer which properties of spectroscopy signal, x , are associated with the outcome, y . The difference between these two settings is highlighted by the fact that there may be no association; indeed, our null hypothesis is $\beta \equiv 0$.

In spirit, our goal is much the same as in A&K, but the goal is accomplished from a reverse perspective. To quote A&K, *the constraint $Lb = 0$ attempts to immunize the model vector against noise (interferent) by orthogonally pointing $\hat{b}$ away from the interferent space*. We adopt the alternative perspective that the penalty term serves to focus the estimation process in a manner that encourages $\hat{\beta}$ to be in or near a “favored” subspace $\mathcal{Q}$. This, and a variety of perspectives on regularization, can be found in the references cited in the introduction.

In general, the least-dominant eigenvectors of a penalty operator L have the largest effect on the estimate $\hat{\beta}(\lambda, L)$. This observation can be used to construct a “favored subspace” by defining a penalty L whose least-dominant eigenvectors (those corresponding to the smallest eigenvalues) are most relevant to the estimation process. One possible approach is as follows: (1) Identify a subspace $\mathcal{Q}$ where β is likely to belong and treat this as a favored space; (2) define a decomposition-based penalty, L that penalizes more when the estimate falls into the orthogonal complement of this favored space $\mathcal{Q}$.

For intuition about how a penalty operator can be used in this way, it is useful to recall the familiar example of a Laplacian, or second-derivative, $L = D^2$. One heuristic for this penalty arises from viewing β as a function whose local “smoothness” is presumed. In this case, the term $\|D^2\beta\|^2$ penalizes sharp changes in $\hat{\beta}$. Alternatively, recall that the dominant

eigenvectors of D^2 are sharply oscillatory while the least dominant eigenvectors are very smooth. Hence, a linear-algebraic view of this is that rather than penalizing sharp changes, smoothness in the estimate is inherited from the eigenproperties of D^2 . Specifically, structure in $\hat{\beta}(\lambda, L)$ arises from the joint eigenproperties of X and D^2 . This heuristic is formalized by the GSVD representation in (2)–(3).

4 Favored spaces and adaptive penalties

As described in Section 2.1, an important motivation for adding a penalty term to a least-squares loss function is to stabilize or reduce the variance of the estimation process. A ridge penalty imposes no structure and simply shrinks all coordinates in $\hat{\beta}$ equally, producing a biased but more stable estimate of β . More generally, a user-defined penalty in terms of L in (3) can be employed to guide the estimation process toward a subspace in $\mathcal{Q} \subset \mathbb{R}^p$ constructed to be orthogonal to spectral structure that is irrelevant to the problem. We refer to such a subspace $\mathcal{Q}$ as “favored” based on the idea that it should contain the “true” β . Using prior knowledge and/or data to inform the construction of $\mathcal{Q}$, one can construct a penalty operator L as follows. Let $P_{\mathcal{Q}}: \mathbb{R}^p \rightarrow \mathcal{Q}$ denote a projection onto $\mathcal{Q}$. If the true regression vector β resides in or near $\mathcal{Q}$ then the orthogonal complement, $\mathcal{Q}^\perp$, should be more heavily penalized than $\mathcal{Q}$. Therefore, we also consider the projection, $I - P_{\mathcal{Q}}$, onto the orthogonal complement, $\mathcal{Q}^\perp$. For practical reasons, an invertible penalty operator is often preferred, so we define

$$L_{\mathcal{Q}} = a(I - P_{\mathcal{Q}}) + (1 - a)P_{\mathcal{Q}}, \quad (4)$$

for some $1/2 \leq a \leq 1$. This can be viewed as a weighted average of two orthogonal projections. Note that when $a = 1/2$, $L_{\mathcal{Q}} = I$ is simply a ridge penalty. On the other hand, when $a = 1$, then $L_{\mathcal{Q}} = I - P_{\mathcal{Q}}$ penalizes only the orthogonal complement of the favored space. In this case, if $\beta \in \mathcal{Q}$, then $L_{\mathcal{Q}}\beta = (I - P_{\mathcal{Q}})\beta = 0$ and no penalty is applied to the true β . Hence an estimate $\hat{\beta}(\lambda, L)$ derived from using $L = L_{\mathcal{Q}}$ in (3) is unbiased. However, if $\mathcal{Q}$ is too large, such an unbiased estimate will have a large variance, making the estimate useless in practice. For example, in the most extreme case of $\mathcal{Q} = \mathbb{R}^p$, we have $L_{\mathcal{Q}} = 0$, which penalizes nothing and, when $p > n$, the OLS estimate is undefined. For this reason, we may wish to initially cast a wide net by defining $\mathcal{Q}$ to be large, then iteratively refine this favored space so as to more parsimoniously represent the preferred structure.

4.1 Penalty learning

To implement this type of penalty, we begin with a large collection of pure analyte spectra or some other informed, pre-collected set denoted by $S^{(0)} = \{q_1, \dots, q_d\}$, where q_s s are linearly independent p -dimensional vectors. Let Q_d be the $d \times p$ matrix whose rows are formed by the vectors in $S^{(0)}$. Then the initial favored space $\mathcal{Q}$ is d -dimensional and

$\mathcal{Q} = \text{Range}(\mathbf{Q}_d^\top) = \text{span}S^{(0)}$. By a slight abuse of notation, we will use $\mathbf{Q}_d$ for both the matrix and the favored space. Then an orthogonal projection onto $\mathbf{Q}_d$ can be defined as

$$\mathbf{P}_{\mathbf{Q}_d}(f) = \mathbf{Q}_d^\dagger \mathbf{Q}_d f, \quad f \in \mathbb{R}^p, \quad (5)$$

Here, the Moore-Penrose inverse $\mathbf{Q}_d^\dagger$ is $\mathbf{Q}_d^\top (\mathbf{Q}_d \mathbf{Q}_d^\top)^{-1}$ since the rows of $\mathbf{Q}_d$ are linearly independent. An important special case is when $S^{(0)}$ contains orthonormal q.s. Then $\mathbf{Q}_d \mathbf{Q}_d^\top = \mathbf{I}$ and the projection can be simplified to $\mathbf{P}_{\mathbf{Q}_d}(f) = \mathbf{Q}_d^\top \mathbf{Q}_d f, f \in \mathbb{R}^p$. Replacing $\mathcal{Q}$ by $\mathbf{Q}_d$ in (4), we denote the initial penalty term by $\mathbf{L}(a, \mathbf{Q}_d)$. When $1/2 < a < 1$, $\mathbf{L}(a, \mathbf{Q}_d)$ is invertible with

$$\mathbf{L}^{-1}(a, \mathbf{Q}_d) = a^{-1}(\mathbf{I} - \mathbf{P}_{\mathbf{Q}_d}) + (1 - a)^{-1} \mathbf{P}_{\mathbf{Q}_d}. \quad (6)$$

Furthermore, the initial estimate for β using (3) is

$$\hat{\beta}^{(0)}(a, \lambda, \mathbf{Q}_d) = [\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{L}(a, \mathbf{Q}_d)]^{-1} \mathbf{X}^\top \mathbf{y}. \quad (7)$$

To simplify notation, we suppress the parameters in $\hat{\beta}$ when the context is clear.

Next, we examine the favored space for similarity between the vectors q_j and the regression coefficient estimate $\hat{\beta}$. Define

$$\begin{aligned} \delta_j &= \langle \hat{\beta}^{(0)}, q_j \rangle, \quad j = 1, \dots, d, \\ \xi_{ij} &= \langle \mathbf{x}_i, q_j \rangle, \quad i = 1, \dots, n, \quad j = 1, \dots, d \\ \sigma_j^2 &= \sum_{i=1}^n (\xi_{ij} - \bar{\xi}_{.j})^2 / (n-1), \end{aligned} \quad (8)$$

where $\bar{\xi}_{.j} = \sum_{i=1}^n \xi_{ij} / n$. Now we define a similarity coefficient between $\hat{\beta}^{(0)}$ and q_j in $S^{(0)}$ as

$$c_j^{(0)} = |\sigma_j \delta_j|, \quad j = 1, \dots, d, \quad (9)$$

and use it to sort q_j decreasingly. Denoted the ordered vector as $q_{(j)}$, which has the j -th largest $c_j^{(0)}$. To trim the initial favored space and improve the penalty, we remove the $q_{(d)}$ with the smallest $c_j^{(0)}$ and denote the set of remaining vectors as $S^{(1)} = \{q_{(1)}, \dots, q_{(d-1)}\}$.

Heuristically, if a q_j is orthogonal to $\hat{\beta}^{(0)}$ (so $\delta_j=0$) then β is less likely to reside in the linear space that contains that q_j and one should exclude it from Q_d . Because the absolute value of δ_j is affected by the size of the j -th term, we normalize it by the standard deviation of X in the direction of q_j which is quantified by σ_j .

With the remaining $d - 1$ vectors in $S^{(1)}$, we refine the favored space

$$Q_{d-1} = \text{span} S^{(1)} \subset Q_d$$

and use this trimmed space to reconstruct the penalty $L(a, Q_{d-1})$ via (4) and (5). Then, we update the estimate of β using the new penalty $L(a, Q_{d-1})$ as in (7), and denote it by $\hat{\beta}^{(1)}(a, \lambda, Q_{d-1})$. In (8)–(9), we recalculate the similarity coefficients $c_j^{(1)}$ between $\hat{\beta}^{(1)}$ and q_j in $S^{(1)}$, for $j = 1, \dots, d - 1$. The q vectors in $S^{(1)}$ can be then reordered based on $c_1^{(1)}, \dots, c_{d-1}^{(1)}$. Removing the q_j with the smallest $c_j^{(1)}$, we have $S^{(2)}$ and $Q_{d-2} = \text{span} S^{(2)}$, a further refined favored space.

Iterating the above procedure and trimming off one q vector at a time, we obtain

$$S^{(0)} \supset S^{(1)} \supset S^{(2)} \supset \dots \supset S^{(d)} = \phi, \\ Q_d \supset Q_{d-1} \supset Q_{d-2} \supset \dots \supset Q_0,$$

which produces a sequence of adaptive penalties. Consequently, we can update the adaptively penalized regression coefficient estimates as

$$\hat{\beta}^{(k)}(a, \lambda, Q_{d-k}) = [X^T X + \lambda L(a, Q_{d-k})^T L(a, Q_{d-k})]^{-1} X^T y, \quad (10)$$

for $k = 0, \dots, d$. Here the tuning parameter a , which balances the fidelity to the observed data and the loyalty to the subjective prior information in q vectors, can be chosen subjectively and may increase as the iteration proceeds. From our experience in simulations, we find little difference in estimation when a is sufficiently close to 1. Therefore, a fixed large a is used in our simulations and data analysis. The detailed stopping rule will be discussed in Section 4.3.

4.2 A Bayesian interpretation of the penalty learning process

Another way to interpret the adaptive penalty and derive the penalized least square estimate $\hat{\beta}$ in (10) is to view the penalty term as controlling variation of β [8, 2, 17, 19]. From a Bayesian perspective, the estimate depends not only on data but also the prior distribution of the parameter. Heuristically, the prior distribution may shrink the estimate toward a smaller region of nominally large parameter space, hence reducing its variability. This is particularly useful when the unknown model parameters lie in a large sample space and the observed data do not contain sufficient information to accurately estimate the parameters. This is

exactly what we need for the examples we are considering. More precisely, if we assume that

$$\mathbf{y}|\mathbf{X}, \beta \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}), \quad \beta \sim N(0, (2\lambda \mathbf{L}^\top \mathbf{L})^{-1}), \quad (11)$$

then the posterior mean of β given $\mathbf{y}$, $E(\beta|\mathbf{y})$, is a Bayes estimate for β under square loss function. Furthermore, using the properties of normal distribution, one can derive the posterior distribution of β and find $E(\beta|\mathbf{y}) = [\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{L}^\top \mathbf{L}]^{-1} \mathbf{X}^\top \mathbf{y}$, which is same as the penalized least square solutions in (3) or (10).

In our setting, the choice of the prior distribution is equivalent to choosing $\mathbf{L}$ in (11). To gain insight on the influence of $\mathbf{L}$ on β , consider the simple case of $\mathbf{L} = \mathbf{I}$, so that $\beta \sim N(0, (2\lambda)^{-1} \mathbf{I})$. This is equivalent to a ridge penalty which assumes that all components of β are independent with variances proportional to $1/\lambda$. A smaller λ will lead to larger variance of all components of β , and cause the estimator to depend more on the observed spectra, $\mathbf{X}$, and the response, $\mathbf{y}$. On the other hand, a larger λ will force β closer to its prior mean, 0. More generally, assuming equal variance for all components of β and shrinking them all toward 0 might not be the best practice, especially if we have some prior information about β .

We now incorporate an informative prior via $\mathbf{L}$. Let $\mathbf{Q}_d$ denote a favored space presumed to contain β . We decompose $\beta = (\mathbf{I} - \mathbf{P}_{\mathbf{Q}_d})\beta + \mathbf{P}_{\mathbf{Q}_d}\beta$, where the first term is the portion of β outside of $\mathbf{Q}_d$ and the latter portion is in $\mathbf{Q}_d$. Intuitively, we should assign a smaller variance for the first term and larger for the second, constraining the portion of β outside of $\mathbf{Q}_d$ more heavily. The $\mathbf{L}$ defined in (4) accomplishes this goal:

$$\begin{aligned} \text{Var}(\beta) &= (2\lambda \mathbf{L}^\top \mathbf{L})^{-1} \\ &= (2\lambda)^{-1} a^{-2} (\mathbf{I} - \mathbf{P}_{\mathbf{Q}_d}) + (2\lambda)^{-1} (1-a)^{-2} \mathbf{P}_{\mathbf{Q}_d} \\ &= \text{Var}((\mathbf{I} - \mathbf{P}_{\mathbf{Q}_d})\beta) + \text{Var}(\mathbf{P}_{\mathbf{Q}_d}\beta). \end{aligned}$$

Evaluating the overall variance in $\mathbf{Q}_d$ and $\mathbf{Q}_d^\top$, we obtain

$$\begin{aligned} \text{trace}(\text{Var}(\mathbf{P}_{\mathbf{Q}_d}\beta)) &= \text{trace}((2\lambda)^{-1} (1-a)^{-2} \mathbf{P}_{\mathbf{Q}_d}) = (2\lambda)^{-1} (1-a)^{-2} d, \\ \text{trace}(\text{Var}((\mathbf{I} - \mathbf{P}_{\mathbf{Q}_d})\beta)) &= \text{trace}((2\lambda)^{-1} a^{-2} (\mathbf{I} - \mathbf{P}_{\mathbf{Q}_d})) = (2\lambda)^{-1} a^{-2} (n-d). \end{aligned}$$

This implies that the average variance of β in each direction of $\mathbf{Q}_d$ is $(1-a)^{-2}$, and in each direction of $\mathbf{Q}_d^\top$ is a^{-2} . Note $a^{-2} < (1-a)^{-2}$ for $1/2 < a < 1$. So, loosely speaking, we start with a large favored space $\mathbf{Q}_d$ and allow a larger variance for the portion of β in the favored space, i.e., $\mathbf{P}_{\mathbf{Q}_d}\beta$; this allows $\hat{\beta}$ to depend more on the data, $\mathbf{X}$ and $\mathbf{y}$. On the other hand, for the portion of β outside of $\mathbf{Q}_d$ i.e., $(\mathbf{I} - \mathbf{P}_{\mathbf{Q}_d})\beta$, we assign smaller variance and shrink this more toward 0; this is the prior mean of β . As we iteratively trim $\mathbf{Q}_d$ and reduce its

dimension, we allow $\hat{\beta}$ to depend on the data in a few supervised directions and restrict its variability in other directions.

4.3 REML estimation and stopping rule

The Bayesian perspective of adaptive penalties also provides a way to compute the proposed $\hat{\beta}$. For an invertible $\mathbf{L}$ defined in (4), if we define $\theta := \mathbf{L}\beta$, then the model (11) becomes $y|X$, $\theta \sim \mathcal{N}(X^* \theta, \sigma^2 \mathbf{I})$, $\theta \sim (0, (2\lambda)^{-1} \mathbf{I})$, where $X^* = \mathbf{X}\mathbf{L}^{-1}$ and $\mathbf{L}^{-1}$ is given in (6). Hence we may first estimate θ through a standard ridge regression $\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}^* \theta\|^2 + \lambda \|\theta\|^2$, and then transform back to β using $\beta = \mathbf{L}^{-1} \theta = a^{-1}(\mathbf{I} - \mathbf{P}_{Q_d})\theta + (1 - a)^{-1}\mathbf{P}_{Q_d}\theta$

Furthermore, the regression coefficients in a ridge estimate can be obtained by using an equivalence with the linear mixed model formulation; see, for example, [23] or [19]. Such a representation facilitates a straightforward use of existing linear mixed model routines widely available in software packages (e.g., R [21] or SAS [24]). Importantly, this formulation also provides an estimate of the tuning parameter λ via the restricted maximum likelihood (REML), in which λ is estimated as a ratio of the variances of the errors, ε , and the regression coefficients, β . This REML-based estimation of tuning parameter λ is used in the application of Section 5. In addition, a value of the estimated λ can be used for stopping rule: stop the iteration when λ is minimized.

4.4 Some properties of SuPEER

To understand the theoretical properties of SuPEER, we investigate the difference between a SuPEER estimate and the target or “true” β . Let Q_{d^*} be the final, selected favored space with dimension d^* and $\mathbf{L} = \mathbf{L}(a, Q_{d^*})$. Then

$$\begin{aligned} \hat{\beta}(a, \lambda, Q_{d^*}) - \beta &= \left(\frac{1}{n} \mathbf{X}^T \mathbf{X} + \frac{\lambda}{n} \mathbf{L}^T \mathbf{L} \right)^{-1} \left(\frac{1}{n} \mathbf{X}^T \varepsilon - \frac{\lambda}{n} \mathbf{L}^T \mathbf{L} \beta \right) \\ &= \mathbf{A}^{-1}(\mathbf{B} - \mathbf{C}), \end{aligned} \quad (12)$$

Where

$$\mathbf{A} = \frac{1}{n} \mathbf{X}^T \mathbf{X} + \frac{\lambda}{n} \mathbf{L}^T \mathbf{L}, \quad \mathbf{B} = \frac{1}{n} \mathbf{X}^T \varepsilon, \quad \mathbf{C} = \frac{\lambda}{n} \mathbf{L}^T \mathbf{L} \beta.$$

Next, we will show that the difference in (12) is actually very small, provided λ and a are appropriately chosen. First, $\frac{1}{n} \mathbf{X}^T \mathbf{X} \rightarrow \text{Var}(\mathbf{X}) \geq 0$ as $n \rightarrow \infty$. Then, assuming that there is a constant c such that $0 < c = \lim \lambda/n < \infty$, we have

$$\frac{1}{n} \lambda \mathbf{L}^T \mathbf{L} \rightarrow c \left[a^2 (\mathbf{I} - \mathbf{P}_{Q_{d^*}}) + (1 - a)^2 \mathbf{P}_{Q_{d^*}} \right] > 0, \quad n \rightarrow \infty$$

for $1/2 < a < 1$. Hence $\mathbf{A} > 0$ and has a finite inverse. (For the limiting case $a = 1$, $\mathbf{A}$ is only invertible when $\text{Null}(\mathbf{X}) \cap \text{Null}(\mathbf{L}) = \{0\}$, which implies $d^* > p - n$.) Second, for the random term $\mathbf{B}$, note that

$$E(\mathbf{B}) = 0, \text{ and } \text{Var}(\mathbf{B}) = \left(\frac{1}{n} \mathbf{X}^T \mathbf{X}\right) \frac{\sigma^2}{n} = O_p\left(\frac{\sigma^2}{n}\right), n \rightarrow \infty$$

under the model $E(y|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{Var}(y|\mathbf{X}) = \sigma^2 \mathbf{V}$. We use notation $O_p(\cdot)$ and $o_p(\cdot)$ to denote convergence in probability. Finally, the term $\mathbf{C} \rightarrow \frac{\lambda}{n} \left[a^2 (\mathbf{I} - \mathbf{P}_{\mathbf{Q}_{d^*}}) \boldsymbol{\beta} + (1 - a)^2 \mathbf{P}_{\mathbf{Q}_{d^*}} \boldsymbol{\beta} \right]$.

Assuming that $a \rightarrow 1$ faster than $1/\sqrt{n}$, we have $\mathbf{C} \rightarrow a(\mathbf{I} - \mathbf{P}_{\mathbf{Q}_{d^*}}) \boldsymbol{\beta}$ as $n \rightarrow \infty$. If the favored space is correctly identified, i.e., $\boldsymbol{\beta} \in \mathbf{Q}_{d^*}$, then $\mathbf{C} = 0$. Putting the three terms together, we have

$$\text{bias}(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1} (E(\mathbf{B}) - \mathbf{C}) = o_p\left(\frac{1}{\sqrt{n}}\right) n \rightarrow \infty,$$

and

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1} \text{Var}(\mathbf{B}) \mathbf{A}^{-1} = O_p\left(\frac{\sigma^2}{n}\right) n \rightarrow \infty.$$

Hence the SuPEER estimate $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, which implies that the estimate approaches the true coefficient vector as the sample size increases.

5 Numerical examples

The SuPEER approach is illustrated on simulated data and using real spectroscopy data from a cohort study. In the two simulation studies, we compared it to ordinary ridge regression, generalized ridge (PEER) and a two-step estimation process in which features in spectra are first extracted and then used as regressors in an ordinary linear regression model. In the first simulation study, we assessed performance by comparing SuPEER with the two-step method by looking at the number of correct basis functions that each method chose from a predefined favored space. In the second simulation study, we compared SuPEER, PEER and ordinary ridge regression in terms of squared error of estimation, $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^2$.

5.1 Simulation study

In subsection 5.1.1 we describe how we simulated the spectroscopy data and how we used various forms of a predefined $\boldsymbol{\beta}$ to construct a “true” association with an outcome, y . Spectroscopy curves, $\mathbf{x}$, were simulated to loosely resemble noisy MRS spectra. In our regression setting, these are sometimes referred to as “predictor functions” and may be as used to predict an outcome such as neurocognitive impairment of HIV patients. Our interests are focused on discovery of biomarkers (e.g., brain metabolites) associated with cognitive decline and hence in our simulations we do not evaluate prediction performance, but rather focus on the error observed while estimating the “true” coefficient vector, $\boldsymbol{\beta}$. Subsection

5.1.2 outlines the various methods used to estimate β while a comparison of the results is discussed in subsection 5.1.3.

5.1.1 Data simulation—Spectra were simulated as an affine combination of “bump functions” with varying support at a number of pre-specified locations, with gaussian noise added to simulate measurement noise. The “true” coefficient vector was defined to correspond to some, but not all, “bumps” in these simulated spectra. Specifically, we constructed each spectrum x to consist of narrow, wide, and moderate bumps centered at: $H_{narrow} = \{0.40, 0.90\}$, $H_{wide} = \{0.50\}$, and $H_{moderate} = \{0.05, \dots, 0.95\}$ (increments of 0.05). We generated the subject-specific spectra $x_i(s)$ to be of the form

$$\begin{aligned} x_i(s) = & \sum_{h \in H_{wide}} (\xi_{w,h} + a_{w,h}) \exp \left[-c_{w,h} * \left(\frac{s-h}{100} \right)^2 \right] \quad (13) \\ & + \sum_{h \in H_{moderate}} (\xi_{m,h} + a_{m,h}) \exp \left[-c_{m,h} * \left(\frac{s-h}{100} \right)^2 \right] \\ & + \sum_{h \in H_{narrow}} (\xi_{n,h} + a_{n,h}) \exp \left[-c_{n,h} * \left(\frac{s-h}{100} \right)^2 \right], \end{aligned}$$

where $\{a_{w,h}, c_{w,h}\}$, $\{a_{m,h}, c_{m,h}\}$ and $\{a_{n,h}, c_{n,h}\}$ correspond to the amplitude and degree of curvature at the wide, moderate and narrow bumps, respectively. The values are described in Table 2. For both the predictor and regression vectors, we sampled s at either $p = 200$ or $p = 400$ equispaced sampling points in the interval $[0, 1]$. Note that the amount of curvature in $x_i(s)$ differs at some locations, s , from the amount in β . $\xi_{w,h}$, $\xi_{m,h}$ and $\xi_{n,h}$ were drawn independently from $\text{Uniform}(0, 0.1)$. Figure 1 exhibits a few examples of these x_i .

The following model was used to generate the outcome data for n independent samples:

$$y_i = \beta_0 + \sum_{j=1}^p x_i(S_j) \beta(S_j) + \epsilon_i, \quad i = 1, \dots, n. \quad (14)$$

$$\beta(s) = \sum_{h \in H_\beta} a_{0,h} \exp \left[-c_{0,h} * \left(\frac{h-s}{100} \right)^2 \right], \quad \text{for } s \in [0, 1].$$

Also, $\beta_0 = 0.0$, $\epsilon_i \sim N[0, \sigma^2]$, where σ^2 corresponded to the signal to noise ratio defined as the ratio of variances of the true responses (noiseless version of $y = X\beta$) to the total variance of observed y (R^2) of 0.8 and 0.9.

5.1.2 Model fitting—We apply both the SuPEER and PEER methods using the decomposition based penalty L_Q in (4). We define the discretized functions $q_1, \dots, q_d$ spanning the favored subspace (PEER or a starting point for SuPEER), where each q_j is defined to have a single bump ($H_{moderate}$). We compute P_Q using (5). Note that the rows of Q

need not be orthogonal, just linearly independent. For a ridge penalty, $L = I_p$. The two-step procedure is defined as: (i) extract $d = 19$ “features” from each x_j by regressing each onto the span of $\{\mathbf{q}_j\}_{j=1}^d$ vectors; and (ii) regress the outcome y on the resulting set of $d = 19$ covariates.

In the first simulation study, proper selection of the penalized space was of primary interest. We summarize the performance of SuPEER by presenting the percentage of correctly chosen q_j vectors and compare it with the two-step procedure over 250 simulation runs.

In the second simulation study, our primary interest lies in estimation of the regression vector β . We summarize the estimation error using a scalar summary, namely integrated squared error (ISE). We report its median and interquartile range over 250 simulation runs.

The estimates of β from the SuPEER, PEER and ridge methods are obtained as best linear unbiased predictors (BLUPs) from the linear mixed model equivalent formulation of the penalized least squares problem (see Section 4.3).

5.1.3 Performance evaluation—In the first simulation study, we evaluated the fidelity of the chosen basis functions and compared it with the two-step estimation procedure. Table 3 summarizes the dimensionality of the Q space and the number of significant features chosen in the two-step procedure, while Figure 2 shows the relative frequencies of the features selected in one specific simulation setting of $n = 100$, $p = 400$ and $R^2 = 0.9$. In all settings considered, the SuPEER method chooses the correct dimension of the Q space most of the time; median = 6 and a very narrow interval for the quartiles ranging in most cases from either 5 to 6 or from 6 to 7. Two-step procedure selects too many features in the first step; median between 12 and 14 with a higher variability.

In the second simulation study, we compared the estimation using ISE. Table 4 gives a summary of estimation errors. In all settings we considered, the methods based on the incorporation of external information (PEER and SuPEER) outperform the ordinary ridge regression by orders of magnitude.

In some settings PEER outperforms SuPEER (last 2 lines of Table 4). Here, the favored space with PEER is all of Q_d , whereas the SuPEER favored space is reduced to Q_{d^*} (of dimension $d^* < d$). If β is not contained in the original Q_d , then the reduction to Q_{d^*} may result in an accumulation of error in some regions of $\hat{\beta}$ (those not represented by Q_d), and this error may exceed the gains obtained from a better SuPEER fit to β than PEER in other regions. This is exhibited in the tradeoff between the pointwise bias and variance as displayed in Figure 4.

5.2 Spectroscopy data example

We consider data from a cross-sectional study of $n = 114$ individuals chronically infected with HIV in which metabolite spectra were obtained from the basal ganglia. In this study, magnetic resonance spectroscopy (MRS) curves were sampled at $p = 399$ distinct frequencies and a marker of neurocognitive impairment, called the global deficit score (GDS) [6], was measured for each subject. An increased GDS score indicates a high degree

of impairment. A detailed study description can be found in [12]. Here we are interested in finding metabolite markers of neurocognitive impairment by modelling the association of the metabolite spectra, $\{\mathbf{X}_i\}_{i=1}^n$, with the global deficit score, $\{y_i = \text{GDS}\}_{i=1}^n$.

The MRS spectra are comprised of pure metabolite spectra, instrument noise and a background profile. Figure 5 shows a sample of 10 subjects' spectra. Spectral information from the following $d = 9$ pure metabolites was used as to define an initial favored subspace for the SuPEER estimation procedure: Creatine (Cr), Glutamate (Glu), Glucose (Glc), Glycerophosphocholine (GPC), *myo*-Inositol (Ins), N-Acetylaspartate (NAA), N-Acetylaspartylglutamate (NAAG), *scyllo*-Inositol (Scyllo) and Taurine (Tau). All pure metabolite spectra are displayed in Figure 6. It is worth observing that there are several prominent features (peaks) in the pure metabolite spectra as well as regions which contain mostly measurement error. We applied SuPEER to select from the favored space and estimate a regression vector, β , that relates the metabolite spectra to the outcome, GDS. In the final solution, $d^* = 3$ pure metabolite spectra were selected for the Q_{d^*} space (see the top panel of Figure 7) and the estimated regression function β (bottom panel of Figure 7) corresponds, in most part, to the selected vectors spanning the Q space. Finally, we note that the two-step procedure provided a solution with no significant predictors; i.e., this procedure resulted in a null model.

6 Discussion

We have aimed to provide a bridge between two inverse problems: approximating a model vector in a calibration model, and estimating a coefficient vector in a statistical regression model. Our starting point was the mathematical equivalence between GTR (for calibration) and PEER (for regression). Although the mathematical formulations for these two methods are identical, their goals and motivations are different. In reviewing and connecting these concepts we have aimed to improve on this framework to be more flexible in formulating an adaptively modified penalization process that allows for a less rigid pre-specification of prior information. Our simulations and example illustrate this approach for estimation in a statistical regression model and we imagine that these concepts may find application in the context of approximating a model vector in a calibration model. Among the proposed tools is a method of selecting the tuning parameter that is implemented by way of the linear mixed effects model framework. This may also prove useful in calibration settings.

The empirical results from our simulations suggest our choices worked well. In particular, in the MRS example, the two-step method found no significant pre-selected features associated with GDS. There are, however, several potential limitations of our approach. These include its strong dependence on defining the initial favored space, Q. Although the decomposition-based penalty in (4) aims to alleviate a complete dependence, there is still the issue how these two terms should be weighted. Additionally, although the adaptively refined favored space, as described in Section 4, also reduces the dependence on this choice, the stopping criteria we present here is not backed up by theoretical properties. These and additional theoretical aspects of the procedure require further investigation.

Acknowledgements

Research support for TR, MK and JH was provided by the National Institutes of Health grant R01 MH108467 (Harezlak); TR was additionally supported by R01 GM114029 (Shojaie).

References

- [1]. Andries E and Kalivas JH, Interrelationships between generalized tikhonov regularization, generalized net analyte signal, and generalized least squares for desensitizing a multivariate calibration to interferences, *Journal of Chemometrics* 27 (2013), no. 5, 126–140.
- [2]. Rice JA Brumback BA, Smoothing spline models for the analysis of nested and crossed samples of curves, *Journal of the American Statistical Association* 93 (1998), no. 443, 961–976.
- [3]. Bertero M and Boccacci P, *Introduction to Inverse Problems in Imaging*, Institute of Physics, Bristol, UK, 1998.
- [4]. Björck A, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
- [5]. Brown P, *Measurement, Regression and Calibration*, Oxford University Press, Oxford, UK, 1993.
- [6]. Carey CL, Woods SP, Gonzalez R, Conover E, Marcotte TD, Grant I, and Heaton RK, Predictive validity of global deficit scores in detecting neuropsychological impairment in hiv infection, *Journal of clinical and experimental neuropsychology* 26 (2004), no. 3, 307–319. [PubMed: 15512922]
- [7]. Choi E, Hall P, and Rousson V, Data sharpening methods for bias reduction in nonparametric regression, *Annals of Statistics* (2000), 1339–1355.
- [8]. Smith AFM Lindley DV, Bayes estimates for the linear model, *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (1972), no. 1, 1–41.
- [9]. Engl HW, Hanke M, and Neubauer A, *Regularization of inverse problems*, Kluwer, Dordrecht, Germany, 2000.
- [10]. Golub GH, Heath M, and Wahba G, Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* 21 (1979), no. 2, 215–223.
- [11]. Hansen PC, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, PA, 1998.
- [12]. Harezlak J, Buchthal S, Taylor M, Schifitto G, Zhong J, Daar E, Alger J, Singer E, Campbell T, Yiannoutsos C, et al., Persistence of HIV-associated cognitive impairment, inflammation, and neuronal injury in era of highly active antiretroviral treatment, *AIDS* 25 (2011), no. 5, 625–633. [PubMed: 21297425]
- [13]. Hastie T, Tibshirani R, and Friedman J, *The elements of statistical learning: Data mining, inference, and prediction*, Springer-Verlag, New York, 2011.
- [14]. Hoerl AE and Kennard RW, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12 (1970), no. 1, 55–67.
- [15]. Kalivas JH, Overview of two-norm (l2) and one-norm (l1) tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance, *Journal of Chemometrics* 26 (2012), no. 6, 218–230.
- [16]. Kalivas JH and Palmer J, Characterizing multivariate calibration tradeoffs (bias, variance, selectivity, and sensitivity) to select model tuning parameters, *Journal of Chemometrics* 28 (2014), no. 5, 347–357.
- [17]. Lin X and Zhang D, Inference in generalized additive mixed models by using smoothing splines, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61 (1999), no. 2, 381–400.
- [18]. Neter J Li W, Kutner M, Nachtsheim C, 5th edition ed., McGraw-Hill/Irwin, 2004.
- [19]. Muñoz Maldonado Y, Mixed models, posterior means and penalized least-squares, *Lecture Notes–Monograph Series*, vol. Volume 57, pp. 216–236, Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2009.
- [20]. Provencher SW, Estimation of metabolite concentrations from localized in vivo proton NMR spectra, *Magnetic Resonance in Medicine* 30 (1993), no. 6, 672–679. [PubMed: 8139448]

- [21]. R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3–900051–07–0.
- [22]. Randolph TW, Harezlak J, and Feng Z, Structured penalties for functional linear models—partially empirical eigenvectors for regression, *Electronic Journal of Statistics* 6 (2012), 323–353. [PubMed: 22639702]
- [23]. Ruppert D, Wand MP, and Carroll RJ, *Semiparametric regression*, Cambridge University Press, New York, 2003.
- [24]. SAS Institute Inc., *SAS/STAT software*, version 9.2, Cary, NC, 2008.
- [25]. Tikhonov AN and Arsenin VA, *Solution of ill-posed problems*, Winston and Sons, 1977.
- [26]. Wahba G, *Spline models for observational data*, vol. 59, Society for Industrial Mathematics, Philadelphia, 1990.
- [27]. Wand MP and Jones MC, *Kernel smoothing*, Chapman and Hall/CRC Press, 1994.

Figure 1:
 $n = 5$ sample predictor curves x sampled at $p = 400$ points.

Figure 2: Selection frequency (out of 250 simulations) of the features in Q for SuPEER (black bars) versus the number of the significant factors selected in the two-step procedure (red bars) for the simulated data with $n = 100$, $p = 400$ and $R^2 = 0.9$.

Figure 3:
The true regression vector (broken thick black line) with 10 sample estimates from: ridge (top), SuPEER (middle), PEER (bottom)).

Figure 4:
Top: a decomposition of pointwise MSE for SuPEER as: pointwise squared bias (black) and pointwise variance (red). Bottom: pointwise median ISE for SuPEER (black) and PEER (red), as summarized in Table 4.

Figure 5:
Sample spectra for 10 subjects.

Figure 6:
 $d = 9$ pure metabolite spectra.

Figure 7:
Estimate of the regression function β (bottom panel) with the 3 pure metabolite spectra spanning the favored Q_{d^*} space (top panel).

Table 1:

Summary of terms used for calibration versus regression modeling.

terminology	calibration model	statistical regression model
model	$\mathbf{x}^T \mathbf{b} = y + e$; $\mathbf{Xb} = \mathbf{y} + \mathbf{e}$	$y = \mathbf{x}^T \beta + \epsilon$; $\mathbf{y} = \mathbf{X}\beta + \epsilon$
response	y , a known physical property, quantified without error	y , a quantified trait, observed with error
estimate	$\hat{\mathbf{b}}$, model vector; a non-random approximation of $\mathbf{b}$	$\hat{\beta}$, regression coefficient vector. $\mathbf{A}$ random vector estimated with uncertainty
error	e , model approximation error; $E(\mathbf{x}^T \mathbf{b}) = y$	ϵ , random error in y ; $E(y \mathbf{x}) = \mathbf{x}^T \beta$
estimation variance	—	$Var(\hat{\beta})$
bias	$\ \mathbf{y} - \hat{\mathbf{y}}\ ^2$	$bias(\hat{\beta}) = E(\beta - \hat{\beta})$
mean squared error of estimation (MSE)	—	$MSE(\hat{\beta}) = E(\ \beta - \hat{\beta}\ ^2) = \ bias(\hat{\beta})\ ^2 + trace(Var(\hat{\beta}))$
prediction	$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$	$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$
mean squared error of prediction	$\frac{1}{n} \ \mathbf{y} - \hat{\mathbf{y}}\ ^2$	$\frac{1}{n} \ \mathbf{y} - \hat{\mathbf{y}}\ ^2$
prediction variance	proportional to $\ \hat{\mathbf{b}}\ ^2$	$Var(\mathbf{y} - \hat{\mathbf{y}})$
interferent	non-analyte signal or spectral noise	measurement noise
inference	—	testing regions for which $\beta = 0$

Table 2:

The amplitude and curvature used to generate the predictors x and regression vector β . Parameters defining features in equation (14) are specified in columns *Wide*, *Moderate* and *Narrow*. H_β represents the set of sampling points corresponding to bumps in β . Note that the degree of curvature in β may be different from the one assumed in x . We considered two scenarios for β curvature at $s = 0.5$: (i) $C = 1000$ (same curvature as x) and (ii) $C = 500$ (different curvature from x).

h	<i>Wide</i>		<i>Moderate</i>		<i>Narrow</i>		H_β	
	amp	curv	amp	curv	amp	curv	amp	curv
	$(a_{w,h})$	$(c_{w,h})$	$(a_{m,h})$	$(c_{m,h})$	$(a_{n,h})$	$(c_{n,h})$	$(a_{0,h})$	$(c_{0,h})$
0.05			1	1000				
0.10			1	1000				
0.15			1	1000				
0.20			1	1000			0.15	1000
0.25			1	1000				
0.30			1	1000				
0.35			1	1000			0.12	1000
0.40					0.25	2500		
0.45			1	1000				
0.50	1	500					-0.10	C
0.55			1	1000				
0.60			1	1000			0.08	1000
0.65			1	1000				
0.70			1	1000			0.06	1000
0.75			1	1000				
0.80			1	1000			0.06	1000
0.85			1	1000				
0.90					0.25	2500		
0.95			1	1000				

amp: Amplitude; curv: Curvature

Table 3:

Median dimension of $\mathbf{Q}_{d^*}$ with first and third quartiles in parentheses as selected by the SuPEER method and the number of significant covariates from the first stage of the two-step procedure, under various simulation scenarios: strengths of signal, R^2 , sampling points, p , and whether the curvature in simulated β is the same or different as in x at $s = 0.5$.

R^2	p	β	d^* in SuPEER	# covariates in two-step
0.9	400	same	6 (6, 7)	13 (13, 14)
0.9	200	same	6 (6, 6)	13 (13, 14)
0.8	400	same	6 (5, 6)	12 (10, 12)
0.8	200	same	6 (5, 6)	12 (11, 13)
0.9	400	different	6 (6, 7)	14 (13, 15)
0.9	200	different	6 (6, 7)	14 (14, 15)
0.8	400	different	6 (5, 6)	12 (11, 13)
0.8	200	different	6 (5, 7)	12 (11, 13)

Table 4:

Median integrated squared error, med ISE, with first and third quartiles for three methods in simulation with selected bump locations. The simulation scenarios include varied strengths of signal, R^2 , sampling points, p , and whether the curvature in simulated β is the same or different as in x at $s = 0.5$. The sample size is $n = 100$ in each case.

R^2	p	β	SuPEER med ISE	PEER med ISE	ridge med ISE
0.9	400	same	0.020 (0.011, 0.038)	0.061 (0.046, 0.082)	1.852 (1.806, 1.898)
0.9	200	same	0.018 (0.010, 0.032)	0.061 (0.046, 0.075)	1.359 (1.285, 1.441)
0.8	400	same	0.068 (0.035, 0.177)	0.123 (0.106, 0.159)	1.940 (1.894, 1.997)
0.8	200	same	0.063 (0.030, 0.175)	0.120 (0.099, 0.159)	1.526 (1.448, 1.609)
0.9	400	different	0.052 (0.041, 0.072)	0.068 (0.051, 0.083)	1.955 (1.898, 2.009)
0.9	200	different	0.050 (0.039, 0.074)	0.069 (0.052, 0.086)	1.441 (1.352, 1.528)
0.8	400	different	0.186 (0.069, 0.252)	0.135 (0.119, 0.173)	2.058 (2.012, 2.114)
0.8	200	different	0.169 (0.069, 0.253)	0.133 (0.109, 0.175)	1.613 (1.534, 1.705)