Chemometrics and Intelligent Laboratory Systems 217 (2021) 104386

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Multivariate calibration on heterogeneous samples

Bin Li [a], Brian D. Marx [a,*], Somsubhra Chakraborty [b], David C. Weindorf [c]

[a] Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA, 70803, USA
[b] Agricultural and Food Engineering Department, IIT Kharagpur, 721302, India
[c] Department of Earth and Atmospheric Sciences, Central Michigan University, Mount Pleasant, MI, 48859, USA

ARTICLE INFO

Keywords:
Multivariate calibration
P-splines
Signal regression
Varying-coefficient model

ABSTRACT

Data heterogeneity has become a challenging problem in modern data analysis. Classic statistical modeling methods, which assume the data are independent and identically distributed, often show unsatisfactory performance on heterogeneous data. This work is motivated by a multivariate calibration problem from a soil characterization study, where the samples were collected from five different locations. Newly proposed and existing signal regression models are applied to the multivariate calibration problem, where the models are adapted to handle such spatially clustered structure. When compared to a variety of other methods, e.g. kernel ridge regression, random forests, and partial least squares, we find that our newly proposed varying-coefficient signal regression model is highly competitive, often out-performing the other methods, in terms of external prediction error.

1. Introduction

In some applications, regressors come in the form of a spectra, such that they resemble a signal or curve. This form of regressors can be problematic to "traditional" statistical modeling, in that one is often faced with an ill-conditioned estimation problem. Our work is motivated by a recent study in environmental soil science, where hyperspectral sensors have been widely used for rapid, non-invasive, and cost-effective natural resource management. In our application, visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) is used to predict multiple soil parameters, which we refer to as response variables.

Certain modeling challenges arise, as the distribution of the calibration samples is not always homogeneous. Such heterogeneity can arise for a variety of reasons, including clustered sampling from a range of geographic locations. Both global and local model fitting approaches often show unsatisfactory prediction performance. The former, which ignores the data heterogeneity, can be too rigid to model the non-homogeneous data. The latter, also called *clusterwise regression* [1], ignores the joint information across different clusters and fits each cluster separately. Refs. [2,3] proposed the partial least squares (PLS) regression and principal component regression (PCR) approaches for clusterwise regression on functional data. The latent mixed effect model [4] is another possible approach to model heterogeneous data. However, the EM-type algorithm is typically computationally intensive and requires multiple steps to converge. Ref. [5] developed a locally weighted modeling approach using the perdetermined ordinal group information to handle the heterogeneous data problem.

In this paper, newly proposed and existing signal regression models are applied to the multivariate calibration problem, where the models are adapted to handle such spatially clustered structure. We further compare these models to other approaches, including: random forests, kernel ridge regression, partial least squares, and principal component regression.

In the next section, we provide our motivating example that outlines the data structure and notation. We then move to Section 3, which provides a survey of various existing and newly proposed penalized signal regression models. Section 4 furnishes the details of the design parameters and optimal tuning aimed at quality external prediction, whereas Section 5 fully explores the comparison of the models presented. We close with a Discussion, elaborating on the reasons why signal regression models are competitive, while in some cases outperforming some machine learning approaches.

2. Motivating example

The dataset contains a total of 900 soil samples collected from five different locations: Seward County (Nebraska, 225 samples), Kern County (California, 225 samples), Lubbock County (Texas, 225 samples), Clay County (Minnesota, 75 samples), and the country of France (150

* Corresponding author.
E-mail addresses: bli@lsu.edu (B. Li), bmarx@lsu.edu (B.D. Marx), somsubhra@agfe.iitkgp.ernet.in (S. Chakraborty), weind1dc@cmich.edu (D.C. Weindorf).

https://doi.org/10.1016/j.chemolab.2021.104386
Received 6 March 2021; Received in revised form 28 June 2021; Accepted 6 July 2021
Available online 14 July 2021

samples). In France, all 150 composited soil samples were collected in Milly-la-Forêt. Sampling sites were randomly selected across the field using ArcGIS software. Fig. 1 shows the geographic map of the five sampling locations, also providing an impression of the vast distances between the sampling locations. Eight physicochemical properties were measured for all 900 soil samples. They are: Cation Exchange Capacity (CEC, mEq/l), Electrical Conductivity (EC, μS/cm), % Total Nitrogen Level, % Total Carbon Level, % Soil Organic Matter (SOM), % Clay, % Sand, and % Silt. The additional soil parameter Loss of Ignition (LOI) was removed from the study, as it was found to be highly collinear with SOM.

All samples were scanned using a portable PSR 3500 VisNIR spectroradiometer (same one for all the samples) with a spectral range of 350–2500 nm. After smoothing and taking first-order derivatives (differences), the processed reflectance spectra were resampled from 360 to 2490 nm, by 10 nm, resulting in 214 wavelength channels. Often first differences are used to remove uninformative vertical shifts among the spectra. A portion of the dataset was originally used in Ref. [6]; in which the details of the sample collection, preparation and data measurement are described. We can use the same spectral regressor information to separately model each of the eight soil responses; said differently, the regressors remain the same, while the response variable changes. The left panel of Fig. 2 is a visual representation of the spectra regressor matrix X, of dimension $m = 900$ rows by $p = 214$ columns, summarizing the digitized spectra by 10 nm intervals.

The right panel of Fig. 2 will serve as our proxy spatial representation for traditional longitude and latitude geographical coordinates. Given the distances between the various locations of the sampled states, compounded by France being on another continent, we choose spatial representation using the first two principal components of X (capturing over 70% of the variation) and found separation among the sampling locations. The ratio of between clusters sum of squares (BSS) and total sum of squares (TSS) is 80.9% indicating a good separation for the five clusters and evidence for the consideration for some form of a spatial effect.

3. Penalized signal regression and model variations

In soil science, the applicability of VisNIR DRS-based prediction of soil properties is dependent on robust calibration models. The regression goal is to relate a scalar response (y) to the signal regressors (X) yielding a multivariate calibration model that has quality external (future) prediction. We begin with the basic model, which stems from the foundational work found in Refs. [7,8] and can be expressed as

$$E(y_i) = \mu_i = \int x_i(v)\beta(v)dv, \tag{1}$$

where $x_i(v)$ is the functional regressor, and $\beta(v)$ is the coefficient function, each associated with the continuous index v (wavelength). We see that through the coefficients, prediction related to the signal regressors varies with v. We can get an idea of such from Fig. 2 (left panel), which provides a visual for functional regressors. The notion of a coefficient function (sometimes referred to as a contrast template) is further displayed ahead in Fig. 3 (also the left panel). In fact, our example does not have continuous $x_i(v)$ but rather a discrete digitization of the signal on $p = 214$ evenly spaced channels, resulting in a high-dimensional regression model

$$\mu_i = \sum_{j=1}^{214} x_{ij}\beta_j. \tag{2}$$

We next summarize our main modeling approaches, which at their core address regression on signals or spectra. Despite the fact that we have a $m > p$, and thus an apparent "full rank" regression setting, the model $\mu = X\beta$, for high dimensional β, is compromised. There is severe collinearity among the columns of X. Additionally, there is also an important ordering index for the spectra, and it may be sensible to explore statistical models that utilize this additional structure. The approaches taken here constrains β: our choice is a smoothness. We emphasize that the coefficient vector is being smoothed, not the spectra, in a way that either assumes the smoothness constraint for the coefficients is a reasonable assumption or non-detrimental toward prediction. For a complete overview, refer to Eilers and Marx (2021, Chapter 7). Ref. [9] also provides an excellent survey of functional regression approaches.

3.1. Penalized signal regression (PSR)

This model fits a smooth signal regression model onto the soil responses, using P-splines. The main idea of the Penalized Signal Regression (PSR) approach [10] is to constrain (smooth) β, while optimizing tuning for good external prediction performance. The regression model reexpresses the coefficient vector as $\beta = B\alpha$:

Fig. 1. Geographical map of where samples are taken in the United States and France.

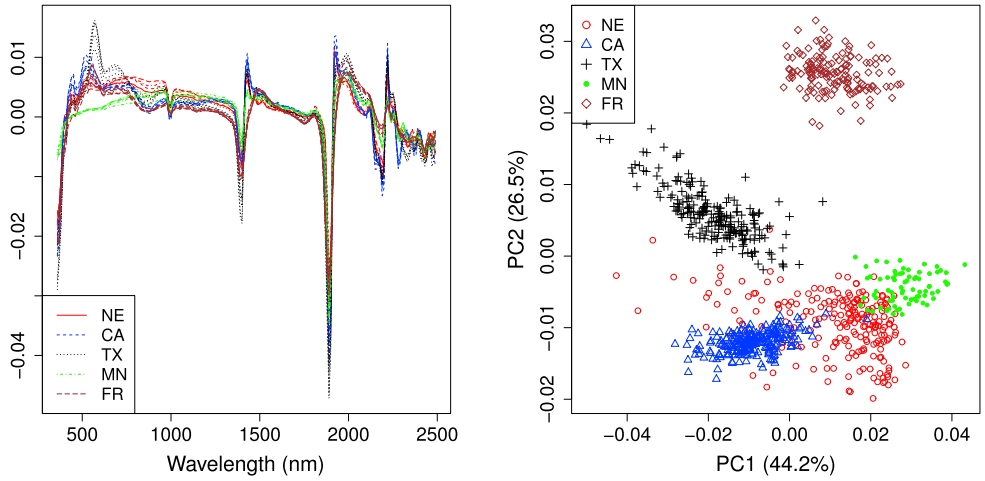

Fig. 2. Left: five sample spectra (first derivative) from each location. Right: PCA plot of 900 samples from 5 locations with different symbols and colors, capturing over 70% of the spatial variation and depicting good cluster separation. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

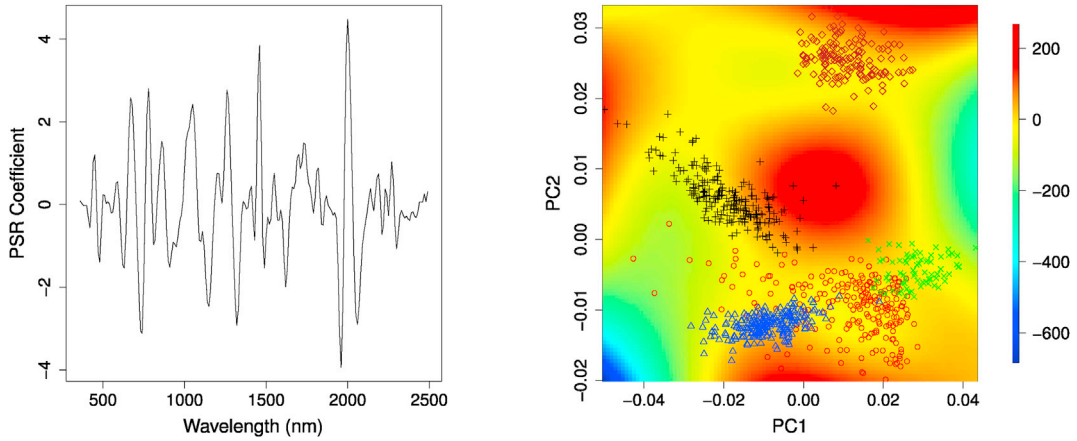

Fig. 3. For the response EC: PSR coefficient curve (left) and the fitted 2D response surface f_{geo} in PSR-geo.

$$\mu_i = \sum_{j=1}^{p} x_{ij}\beta_j = \sum_{j=1}^{p} x_{ij} \sum_{r=1}^{n} B_r(v_j)\alpha_r, \tag{3}$$

where $X = \{x_{ij}\}$ $(i = 1{:}\, m; j = 1{:}\, p)$, the (rich) B-spline basis matrix B (of dimension $p \times n$) is constructed along the index of the spectra (v), and α represents the B-spline coefficients of length n (usually of lower dimension than β). According to P-spline protocol, ample and evenly spaced knots are used. The above expression can be written in matrix representation through

$$\mu = X\beta = XB\alpha = U\alpha, \tag{4}$$

with modest dimensional *effective* regressors U of dimension $m \times n$.

The penalization presents itself through the minimization of the following objective:

$$S(\alpha) = ||y - U\alpha||^2 + \lambda||D\alpha||^2, \tag{5}$$

with the difference matrix D having banded structure such that its rows consist of polynomial contrasts of order d. Notice that the penalties are directly on α (the B-spline coefficients) with positive tuning parameter λ regularizing smoothness in α, and thus β. See Ref. [11] for more details about the penalty. The explicit solution for PSR is thus:

$$\hat{\alpha} = (U'U + \lambda D'D)^{-1}U'y, \tag{6}$$

and the corresponding estimated coefficient vector for the spectra is then $\hat{\beta} = B\hat{\alpha}$. Fig. 3 (left panel) shows the optimal PSR estimate $\hat{\beta}$ for the response EC. Section 4 provides details of optimal tuning based on cross-validation measures.

3.2. Additive signal and spatial model (PSR-geo)

This model brings in "spatial" effects. First, PSR is applied by regressing the soil parameters onto the spectra, and then constructing the residuals. The PC1&2 scores from the signal regressors are computed as the orthogonal proxy spatial variables. Using the PC1&2 scores as inputs, we fit a two-dimensional smooth surface, called f_{geo}, on the scattered PSR residuals with anisotropic penalization of tensor product P-splines. For details on two-dimensional smoothing, see Eilers and Marx (2021, Chapter 4). The model has the form

$$\mu = U\alpha + f_{geo}(PC1, PC2), \tag{7}$$

where tuning is performed sequentially on each term in the model. For the response EC, Fig. 3 shows the optimal PSR coefficient curve (left) and the corresponding optimally fitted two-dimensional response surface on

the residuals, using tensor product P-splines. This right panel highlights a desirable feature associated with the PSR-geo model, i.e. the spatial information of the samples is built into the model through PC1&2. Further the user can better understand the relative magnitudes for various regions of the PC plane, as well as its interactive features.

3.3. Varying-coefficient penalized signal regression (PSR-VC)

Interactive structure is now allowed through varying-coefficient penalized signal regression model. The soil responses are each regressed onto the spectra resulting in a smooth coefficient vector along v. However we now allow this contrast template to vary over another measured covariate t, for example the first PC. The next section will bring in other PCs. In essence a two-dimensional coefficient surface is produced, and then sliced at a particular level of the covariate t. With some of the technical details described below, such a varying coefficient PSR approach is the topic of [12]. The problem is effectively solved by using a modified tensor product basis constructed with $U = XB$ (for the spectra) and $\breve{B}$ (as the B-spline basis for the covariate t). For the ith observation, with signal x_{ij} ($j = 1{:}\ p$) and covariate t_i, the varying signal coefficient construction stems from the following:

$$\begin{aligned}\mu_i &= \sum_{j=1}^{p} x_{ij} \sum_{r=1}^{n} \sum_{s=1}^{\breve{n}} B_r(v_j)\breve{B}_s(t_i)\alpha_{rs} \\ &= \sum_{j=1}^{p} \sum_{r=1}^{n} \sum_{s=1}^{\breve{n}} x_{ij} b_{jr} \breve{b}_{is}\alpha_{rs} \\ &= \sum_{r=1}^{n} \sum_{s=1}^{\breve{n}} \left(\sum_{j=1}^{p} x_{ij} b_{jr} \right) \breve{b}_{is}\alpha_{rs} \\ &= \sum_{r=1}^{n} \sum_{s=1}^{\breve{n}} u_{ir} \breve{b}_{is}\alpha_{rs}.\end{aligned} \tag{8}$$

The above expression (8) can be written in matrix representation using Kronecker products. Consider "unfolding" the coefficient surface and expressing the mean in a standard multiple regression of the form

$$\mathrm{vec}(\mu) = \mathbf{U}\boldsymbol{\alpha}, \tag{9}$$

where $\mathbf{U} = (U \otimes \mathbf{1}_{\breve{n}}') \odot (\mathbf{1}_n' \otimes \breve{B})$, $\boldsymbol{\alpha}$ is the $n\breve{n}$ − vector of tensor coefficients, and the notation $\otimes$, $\odot$ denotes Kronecker and element-wise products, respectively. The dimension of $\mathbf{U}$ is $m \times n\breve{n}$.

The PSR-VC approach takes a penalized least squares approach by minimizing the following objective:

$$\begin{aligned}Q_P(\alpha) &= \text{Residual SS} + \text{Row Penalty} + \text{Column Penalty} \\ &= ||y - \mathbf{U}\alpha||^2 + \lambda\alpha' P\alpha + \breve{\lambda}\,\alpha'\breve{P}\,\alpha.\end{aligned}$$

Notice that there are two penalties placed directly on α (tensor product coefficients) with (anisotropic) tuning parameters: $\lambda,\ \breve{\lambda}$. The compact representation of the difference penalties associated with the rows and columns of the tensor basis is $P = (D'D) \otimes I_{\breve{n}}$ and $\breve{P} = I_n \otimes (\breve{D}'\breve{D})$, which is needed for the model form given in (9). We suppress the notation that reflects the penalty order, but in practice we usually use second or third order differences. The explicit PSR-VC solution results as

$$\hat{\alpha} = (\mathbf{U}'\mathbf{U} + \lambda P + \breve{\lambda}\,\breve{P})^{-1}\mathbf{U}'y,$$

which naturally has to be reshaped to construct the two-dimensional surface.

Using the PSR-VC model for the EC response, Fig. 4 shows the optimal varying signal coefficient surface (left) and six slices of the smooth coefficient vectors at fixed levels of the varying index (PC1). We find that estimated coefficient curves for the signal vary wildly across t, indicating a strong interactive spatial effect.

3.4. Two-dimensional varying-coefficient PSR models (PSR-2DVC)

Building on the PSR-VC model, we present a new model that is not published to date. We fit a varying-coefficient penalized signal regression model onto the soil responses, but now allow the spectra coefficients to vary along a two-dimensional surface defined by two covariates $\breve{t}$ and $\dot{t}$. In our application, these covariates are chosen to be the spatial proxy variables, PC1&2. Extending (8) to higher dimensions, we have:

$$\begin{aligned}\mu_i &= \sum_{j=1}^{p} x_{ij} \overbrace{\sum_{r=1}^{n} \sum_{s=1}^{\breve{n}} \sum_{q=1}^{\dot{n}} B_r(v_j)\breve{B}_s(\breve{t}_i)\dot{B}_q(\dot{t}_i)\alpha_{rsq}}^{\alpha(v_j,\ \breve{t}_i,\ \dot{t}_i)} \\ &= \sum_{r=1}^{n} \sum_{s=1}^{\breve{n}} \sum_{q=1}^{\dot{n}} u_{ir} \breve{b}_{is}\dot{b}_{iq}\alpha_{rsq}.\end{aligned} \tag{10}$$

The bracketed formula in (10) clearly shows that the (smooth) signal coefficient vector varies smoothly with "spatial" coordinates associated with sample i, i.e. at location $(\breve{t}_i, \dot{t}_i)$, and as such is then applied as the contrast template for its corresponding digitized signal regressors. As with PSR-VC, the PSR-2DVC expressions can also be written in matrix

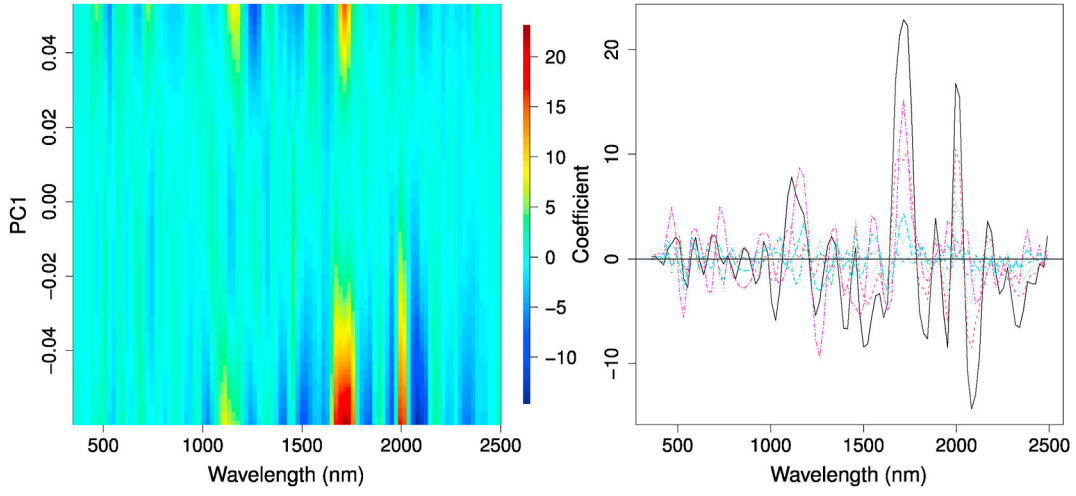

Fig. 4. PSR-VC model for response EC: Optimal varying signal coefficient surface (left) and six slices of the smooth coefficient vectors at fixed levels of the varying index (PC1).

representation. Now (8) requires higher dimensional tensor expressions as follows:

$$\mathbf{U} = [U \otimes (\mathbf{1}'_{\breve{n}} \otimes \mathbf{1}'_{\dot{n}})] \odot [\mathbf{1}'_{n} \otimes (\breve{B} \otimes \mathbf{1}'_{\dot{n}})] \odot [\mathbf{1}'_{n} \otimes (\mathbf{1}'_{\breve{n}} \otimes \dot{B})],$$

and $\boldsymbol{\alpha}$ is the $n\breve{n}\,\dot{n}-$ vector of tensor coefficients. With this additional dimension, the tensor representation of penalties becomes

$$P = D'D \otimes (I_{\breve{n}} \otimes I_{\dot{n}}), \quad \breve{P} = I_n \otimes (\breve{D}'\breve{D} \otimes I_{\dot{n}}) \quad \text{and} \quad \dot{P} = I_n \otimes (I_{\breve{n}} \otimes \dot{D}'\dot{D}),$$

with the explicit PSR-2DVC solution $\hat{\alpha} = (\mathbf{U}'\mathbf{U} + \lambda P + \breve{\lambda}\,\breve{P} + \dot{\lambda}\dot{P})^{-1}\mathbf{U}'y$. We will see in Section 5 that this regression model is highly competitive, even when compared to machine learning approaches aimed toward quality external prediction.

Using the PSR-2DVC model for the EC response, Fig. 5 shows the optimal varying signal coefficients in with PC1 and PC2 fixed at zero through the 4D plot (left) and the corresponding smooth coefficient curve (right). The smooth spectral coefficient vectors are now allowed to vary continuously and smoothly along the entire PC1&2 surface. The 4D plot uses the color to represent the level of the 4th dimension (i.e. the size of the coefficient). Like both Figs. 3 and 4, the red/green/blue represent the large/medium/small coefficient size, respectively. Comparing the estimated signal coefficient vector for the two varying-coefficient approaches, it appears that those from PSR-VC may be generally rougher than those from PSR-2DVC. This is seen when contrasting the right panels of Figs. 4 and 5. An explanation may be that the changes across the one-dimensional index t in the former needs much more flexibility than those changes allowed in across a more generous two-dimensional indexing surface $(\breve{t}\ ,\dot{t})$, in the latter. The lighter smoothness required with a more restricted index is likely a compensation for the loss of index dimension.

4. Design and optimal tuning

The above modeling approaches require some design parameters, such as the size of the B-spline basis and the order of the penalty. Since we are using P-spline models, as mentioned, all of the basis knots are evenly spaced. Further, for simplicity in presentation above, we have not included an intercept term in any of the above PSR expressions. In practice, we find that an intercept term improves the model fit, and we included it in all of our models. Thus a column of ones has been augmented to the spectral X matrix, with proper adjustments to the penalty term to ensure that this associated intercept remains unpenalized. See Ref. [13] (Chapter 7) for such details.

Table 1 provides the signal regression design parameters, including the size of the basis and the order of the difference penalty. In all cases, we used cubic B-splines and third order penalties. Technically, the spline degree and penalty order can also be viewed as additional hyper-parameters, but our choices are commonly used as defaults. A rich basis was implemented, depending on the model, while trying to keep the number of columns (of $\mathbf{U}$) low enough for computational feasibility (e.g. keeping $n\breve{n}\,\dot{n} < 3000$).

The effective model dimension, which quantifies the model complexity, is defined as the trace of the "hat matrix" $\mathbf{H}$, where $\hat{y} = \mathbf{H}y$. Since $\mathbf{H}$ is not idempotent, it is not a projection matrix. For example, in PSR-2DVC model, the effect dimension (ED) can be computed as

$$\begin{aligned} \mathrm{ED} = \mathrm{trace}(\mathbf{H}) &= \mathrm{trace}[\mathbf{U}(\mathbf{U}'\mathbf{U} + \lambda P + \breve{\lambda}\,\breve{P} + \dot{\lambda}\dot{P})^{-1}\mathbf{U}'] \\ &= \mathrm{trace}[(\mathbf{U}'\mathbf{U} + \lambda P + \breve{\lambda}\,\breve{P} + \dot{\lambda}\dot{P})^{-1}\mathbf{U}'\mathbf{U}]. \end{aligned} \tag{11}$$

The second equation in (11) avoids direct computation of the diagonal elements of $\mathbf{H}$ through cyclic permutation and uses a matrix of lower dimension. Similar effective dimension calculations can be made for the other PSR approaches by substituting the $\mathbf{U}$ and the penalty term(s) in an obvious way. Despite the very rich B-spline bases used for each approach, we find from Fig. 6 that the penalty is indeed working. For most soil responses, we find ED to be dramatically reduced compared to the model size. ED roughly rest between 80 and 130 for PSR and PSR-geo (right panel) and is found to be somewhat larger (roughly between 100 and 200) for the varying coefficient PSR models (left panel).

Since external prediction quality is a primary goal, choosing the tuning parameters to minimize cross-validation error is a reasonable choice. For each penalized signal regression approach, the optimal value(s) of λ(s) was (were) found by minimizing the leave-one-out cross-validation (LOOCV) error on the training set. Let y_i be the one observation to be left out and $\hat{y}_{-i}$ be the predicted value at the "left out" location, both used in computation of the LOOCV measure. By repeating this for each observation in turn, the LOOCV prediction error can be computed as

Table 1
Design parameters for various signal regression models.

Design parameters	PSR	PSR-geo	PSR-VC	PSR-2DVC
Basis size	203	203 × 13	203 × 8	53 × 8 × 7
B-spline degree	cubic	cubic	cubic	cubic
Penalty order	3	(3, 3)	(3, 3)	(3, 3, 3)

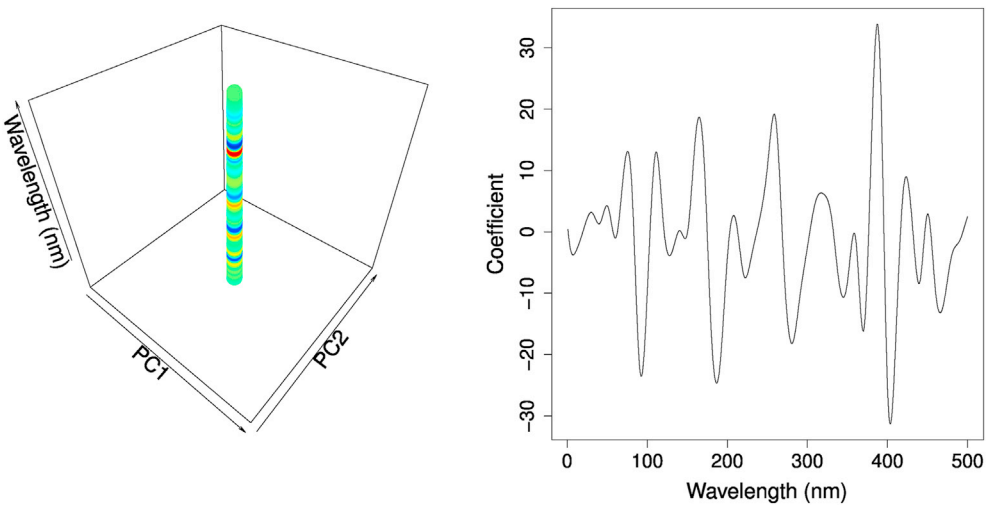

Fig. 5. PSR-2DVC model for response EC: Optimal varying signal coefficient in VCM2D with PC1 and PC2 fixed at zero: 4D plot (left) and the corresponding smooth coefficient curve (right).

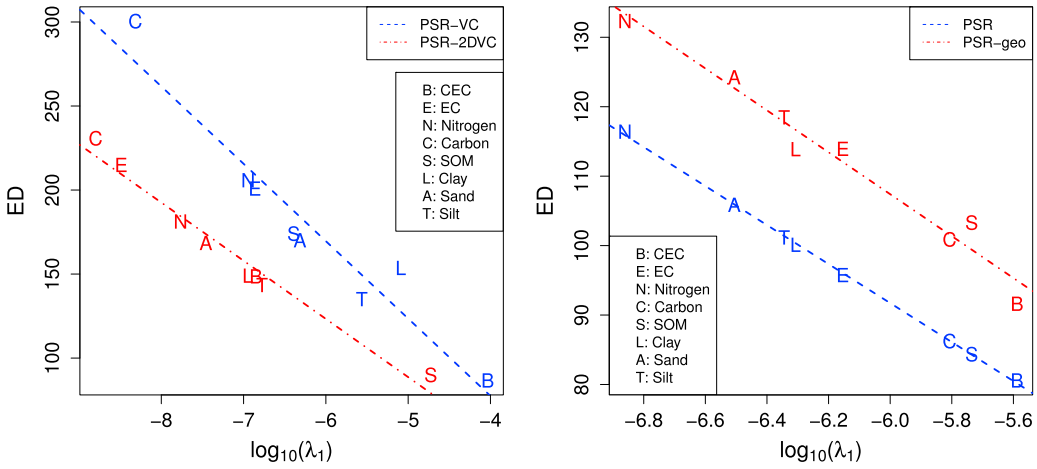

Fig. 6. Scatter plots of effective dimensions for PSR-VC and PSR-2DVC models (left), and PSR and PSR-geo models (right) against $\log_{10}(\lambda_1)$, the tuning parameter associated with the spectra, with fitted lines.

$$PE_{LOOCV} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_{-i})^2}. \tag{12}$$

For linear fitting models, where $\hat{y} = Hy$, one can show that

$$y_i - \hat{y}_{-i} = (y_i - \hat{y}_i)/(1 - h_{ii}),$$

where h_{ii} is the ith diagonal element of "hat" matrix, which is straightforward to compute in the standard regression setting. This type of adjustment from the residual to the "leave-one-out" residual also exactly holds true for all of our PSR approaches, now using the diagonals of **H** in the spirit of (11). We applied the PSR, PSR-geo, PSR-VC and PSR-2DVC on each of the soil responses. Models are fit on the entire dataset with the optimal tuning (λs) shown in Table 2, based on minimizing LOOCV for all models. These choices, along with the design parameters in Table 1, can be useful for users who wish to implement our models using this dataset.

Computation and storage is expensive with the Kronecker product expressions in the varying-coefficient approaches, especially with $\mathbf{U}'\mathbf{U}$. Unfortunately, despite the gridded appearance of the tensor products, any computational relief offered by array regression techniques [14] are not possible. This is because the data are pairs of a two-dimensional surface and a scalar response. The coefficient surface is modeled by tensor product P-splines and the response is not on a grid. However as seen in Fig. 6, and as mentioned, the ED values are generally less than 200 for the vast majority of models, indicating that we perhaps could have managed just as well with much less rich B-spline basis matrices. From Fig. 6, we also see an obvious and expected negative association between λ_1, the tuning parameter associated with the spectra, and the effective dimension.

Table 2
Optimal values of tuning parameters (log_{10} scale).

Method	PSR	PSR-geo	PSR-VC $(\lambda, \breve{\lambda})$	PSR-2DVC $(\lambda, \breve{\lambda}, \dot{\lambda})$
CEC	−5.6	0.37, 3.2	−4.0, 1.1×10^{-3}	−6.9, 1.2×10^{-4}, 2.0×10^{-4}
EC	−6.2	0.04, −0.41	−6.9, -1.7×10^{-2}	−8.5, -1.5×10^{-2}, -2.4×10^{-3}
Nitrogen	−6.9	−0.84, 1.6	−6.9, -9.5×10^{-5}	−7.8, -4.4×10^{-4}, -2.3×10^{-4}
Carbon	−5.8	−0.70, 2.1	−8.3, -7.9×10^{-3}	−8.8, -6.3×10^{-4}, -2.4×10^{-3}
SOM	−5.7	−0.19, −0.38	−6.4, -6.1×10^{-4}	−4.7, -6.9×10^{-4}, 1.6×10^{-4}
Clay	−6.3	−0.66, 7.5	−5.1, −6.4	−6.9, 7.9×10^{-4}, -3.4×10^{-4}
Sand	−6.5	−0.84, 0.57	−6.3, 2.2×10^{-4}	−7.5, -1.0×10^{-4}, -8.7×10^{-4}
Silt	−6.3	−0.77, 0.84	−5.6, -6.0×10^{-4}	−6.8, -7.7×10^{-5}, -1.0×10^{-6}

5. Comparison of models

This section is devoted to comparative studies applied to the soil data set when using the PSR, PSR-geo, PSR-VC, and PSR-2DVC models as described in Section 3. We also make further comparisons to Random Forests, Kernel Ridge Regression, Partial Least Squares (PLS), and Principal Component Regression (PCR). In the study, we randomly split the dataset into a training (sample size is 720, 80% of the entire dataset) and a test set (the remaining 180 observations). The models are fit on the training set and then used for external prediction on test samples for all eight soil variables.

The prediction results are evaluated using *root mean square error* (RMSE) on the test set. The RMSE is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m^{test}}\sum_{i=1}^{m^{test}}(y_i^{\star} - \hat{y}_i^{\star})^2}, \tag{13}$$

where m^{test} is the number of observations on the test set and $\hat{y}_i^{\star}$ is the predicted response for the observation $y_i^{\star}$ in the external test set, using the parameter estimates from the training set with the optimal λ. The results are based on 50 random splits of the dataset, and models were separately optimized in each split using LOOCV. To directly compare the prediction performance, we use the *comparative* test errors, defined by

$$c_{i,j} = \frac{\mathrm{d}_{ij}}{min\{\mathrm{d}_{i,l}\}_{l=1,\ldots,4}},\ i = 1, \ldots, 50;\quad j = 1, 2, 3, 4,$$

where d_{ij} is a performance measure (i.e. RMSE) over 50 replications for each of the four methods: PSR, PSR-geo, PSR-VC and PSR-2DVC. This quantity facilitates individual comparisons by using the test error of the best method for each dataset to calibrate the difficulty of the problem.

Fig. 7 shows the boxplots of the comparative RMSE among PSR, PSR-geo, PSR-VC, and PSR-2DVC on each of the eight soil responses. From this figure, we see that PSR-2DVC models achieve better overall performance in six of the eight soil responses, when compared to other signal regression approaches. Even simpler models perform best on soil responses CEC and EC, i.e. using PSR-geo and PSR-VC, respectively.

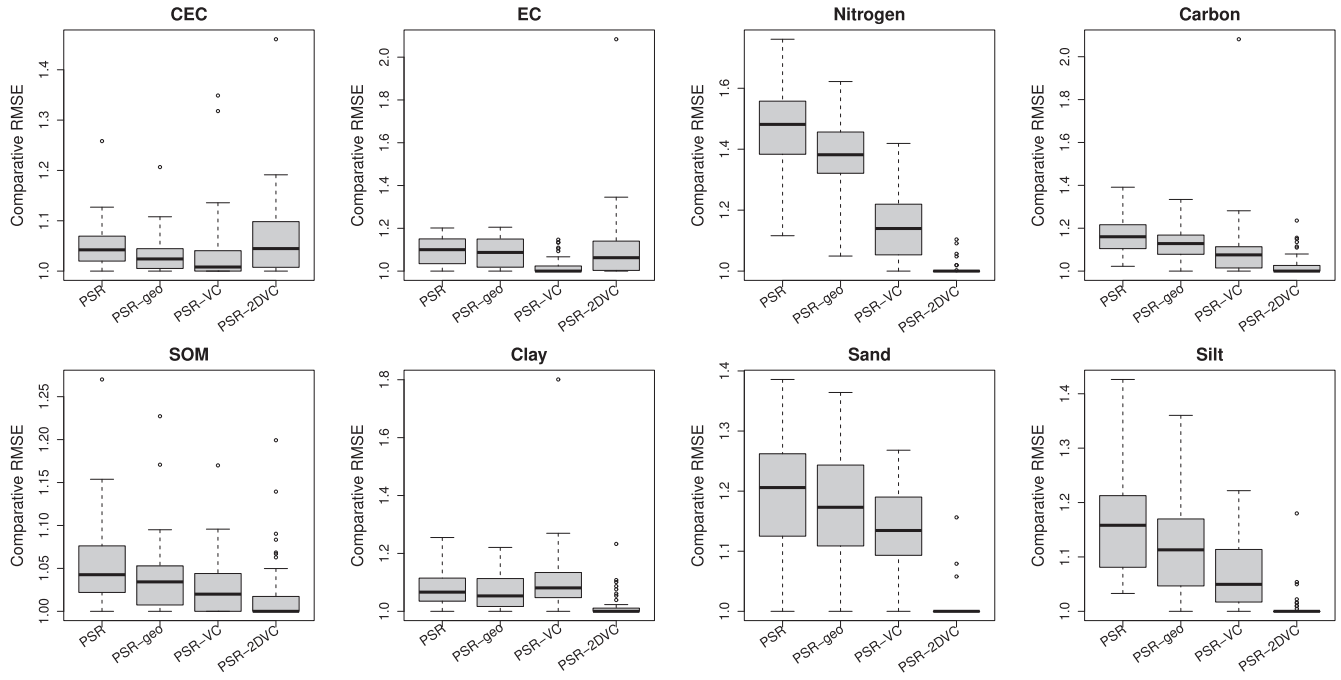

Fig. 7. Boxplots of comparative RMSE for PSR, PSR-geo, PSR-VC and PSR-2DVC based on 50 replications, by soil response.

We next bring in results for four other methods: Random Forests (RF; [15]); and Kernel Ridge Regression (KRR; [16]), Partial Least Squares (PLS; [17]), and Principal Component Regression (PCR; [18]). PCR is a classical linear regression method that overcomes the multicollinearity problem by using linear combinations of the signal regressors as orthogonal constructed regressors. Like PCR, the PLS approach also constructs orthogonal latent variables, but with the correlations of the (deflated) response to the (deflated) wavelength channels as the loadings. The choice of RF and KRR as methods can provide some insight of performance for techniques that are closer in line with machine learning realm. RF is a popular ensemble tree method that enjoys good predictive performance with relatively little hyperparameter tuning. The KRR, which combines ridge regression with "kernel trick," enables the model to approximate nonlinear and non-additive functions. Some further details related to model fitting for these four methods are described in Section 5.1.

Fig. 8 shows the boxplots of the comparative RMSE among RF, KRR, PLS, and PSR-2DVC, again on the eight soil response variables with the results based on 50 random split of the dataset. From Fig. 8, we find that the PSR-2DVC model remains highly competitive, achieving better overall performance in five of the eight soil responses, now when also compared to these other approaches. RF outperforms PSR-2DVC in CEC, Nitrogen and SOM. However, KRR performs best of all in these same: CEC, Nitrogen and SOM. Table 3 shows the average test RMSE of each of the four PSR approaches, as well as RF, KRR, PLS, and PCR on the eight soil responses, with the best result highlighted in each column. This table further confirms the above statements, and in general, our newly proposed PSR-2DVC performs extremely competitively when compared to machine learning methods. In terms of RMSE, the RF, PCR and PLS methods showed mediocre results for all responses. It is interesting that PLS and PCR both outperformed the RF for EC, Carbon, and Clay responses. KRR performs better than RF in seven responses, although the differences are generally very close. This is probably because the RF prediction surface is not smooth. Since RF and KRR perform better than the other linear methods for Nitrogen, we believe the input variables have some nonlinear effects on Nitrogen.

5.1. R software and details for model fitting

All of the code for these models is available from the authors, who used the `R` packages `JOPS` (CRAN) and `JOPSplus` (psplines.bitbucket.io). Within these packages, there are built-in functions `psSignal` for PSR, `psVCSignal` for PSR-VC, and `ps2DVCSignal` for PSR-2DVC. The PSR-geo model used a combination of `psSignal` (signal regression) and `ps2DNormal` (bivariate smoothing for scattered data) functions. The `ucminf` function in `ucminf R` package is used to search the optimal values of λs, which minimize the LOOCV error, in all of the various penalized signal regression models.

Random forest was run by using the `randomForest` package with the default setting in R. The kernel ridge regression used `krr` function in `listdtr R` package. The `plsr` and `pcr` functions in `pls` package were used to fit PLS and PCR models. Note, for RF, the spectra channels and also the PC1&2 are the input variables. The first eight PCs, which explains over 97.5% of the total variance, are used as the inputs for KRR. The spectra channels are the input variables for both PLS and PCR.

The optimal tuning parameters in KRR were also chosen to minimize exact LOOCV error, which is further used to select the optimal number of components in the PLS and PCR models (up to 100 components). For PLS (PCR), we found the optimal number of components ranges between 15 and 30 (60 and 100) components. The reason the optimal number of components in PCR is much larger than the one in PLS is that PCR uses unsupervised features (i.e. PCs are generated without using the response information) in the models.

6. Discussion

At its core, the idea that we present is a relatively simple one, in that we regularize the high dimension regression problem by constraining the β vector to be smooth. In some cases, we further allow this vector to vary smoothly across spatial proxy variables or by boosting prediction through smoothing spatially patterns in the residuals. In our opinion, it is refreshing to find that signal regression statistical models have a competitive nature and outperform some of the ensemble or machine

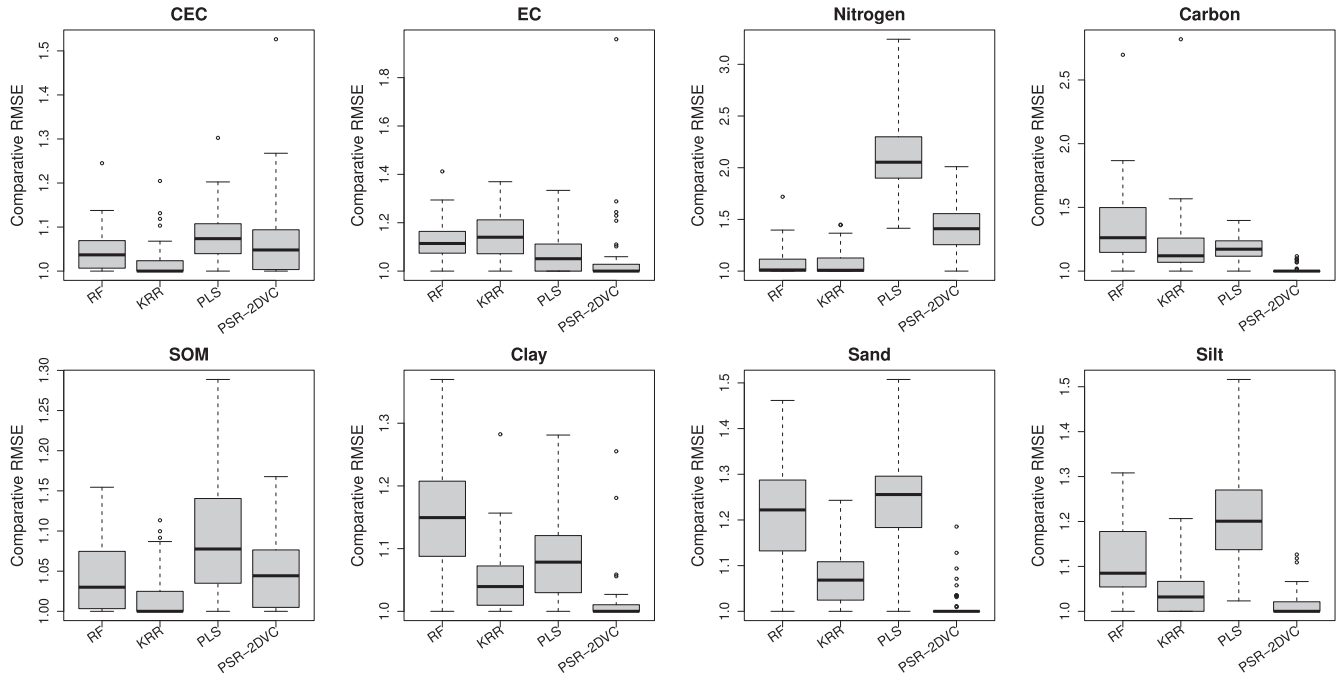

Fig. 8. Boxplots of comparative RMSE for Random Forests, Kernel Ridge Regression, Partial Least Squares, and PSR-2DVC based on 50 replications, by soil response.

Table 3
Average RMSE for the test sets on eight soil responses.

RMSE	CEC	EC	Nitrogen	Carbon	SOM	Clay	Sand	Silt
PSR	3.512	271.397	0.201	0.186	0.411	3.329	6.403	5.660
PSR-geo	3.454	269.400	0.188	0.180	0.406	3.280	6.297	5.473
PSR-VC	3.467	**252.500**	0.157	0.178	0.403	3.392	6.124	5.223
PSR-2DVC	3.559	273.352	0.139	**0.165**	0.398	**3.134**	**5.417**	**4.955**
RF	3.469	292.723	0.110	0.219	0.397	3.552	6.462	5.377
KRR	**3.398**	298.135	**0.107**	0.198	**0.389**	3.242	5.737	5.082
PLS	3.580	278.649	0.207	0.190	0.413	3.351	6.591	5.837
PCR	3.588	277.647	0.207	0.191	0.414	3.358	6.709	5.912

learning approaches, like random forests and kernel ridge regression. Other machine learning approaches, like the Neural network (NN), present little salvation for signal structures. This is because, at its heart, the NN is an over-parameterized scheme, which results in a single layer NN with hundreds (or even thousands) of weights. Some preprocessing for NN can be done using, e.g., principal components, but our experience is that it is sensitive to initial values of weights and unstable, and also optimal tuning is difficult.

Although statistical models may offer some meaningful scientific interpretation or insight into high dimensional regression, they often have difficulties competing with the predictive ability of a broader class of machine learning approaches. Here, we rather find that such "machine learning" approaches, not only offer very little interpretability, but often poorer prediction. Traditionally, there perhaps exists a general and unsatisfying trade-off between competitive prediction and scientific interpretability; gains in one appear to come at a compromise of the other. We hope that our various smoothing approaches can possibly provide a step forward toward a promising dual role of a model: (i) one of regularization, and in some cases interpretability, and (ii) another of leveraging prediction quality through low-dimensional adjustments to the model structure itself, here through tensor varying coefficient or other structure.

Partial Least Squares (Principal Component Regression) is not really competitive in terms of RMSE, producing supervised (unsupervised) linear combinations that likely yield signal coefficients that are too rough to steer competitive prediction. In terms of RMSE, it may be surprising that the newly proposed PSR-2DVC model is outperforming Random Forests in seven of the eight soil variables (apart from Nitrogen). In fact, even the most basic global PSR approach out performs the RF for the responses: EC, Carbon, Clay, and Sand (refer to Table 3). We believe that there may be three reasons for these findings:

1. The Random Forest estimator is not smooth. For situations where the underlying function is smooth, then RF may be suboptimal to estimate the underlying function.
2. In the RF ensemble, each tree uses only a few wavelengths to predict the response. Therefore, RF does not fully use the autocorrelated data structure in prediction, and as mentioned, variable selection is likely not to have a sharp optimal choice.
3. Due to the high-dimensional nature, penalization (i.e. regularization) is the key to avoid over-fitting. Although the RF uses averages to reduce variance, each tree is fully grown without pruning. Therefore, RFs can easily over-fit the data.

Note that there does exist some P-spline methodology for automatically choosing the tuning parameters when such models are framed as mixed-models or in a Bayesian context. To our knowledge, no such tuning approaches have been extended to the penalized signal regression type models presented in this paper. As we are not necessarily driven for "pleasing" smoothness, the choice of tuning rather greedily minimizes external prediction error. Should such automated tuning be developed, it would be interesting to thoroughly investigate just how such a choice for

tuning fairs in terms of external RMSE. Separately, if one wishes to implement further nonlinear structure into the penalized signal regressions approaches, then single-index approaches can be taken, e.g. $f(\mu) = X\beta$, with $f(\cdot)$ denoting an unknown but explicit link function. Ref. [19] presents such modifications for PSR, while [20] extends such single-index models to PSR-VC. Further research could explore whether such additional nonlinear structure further improves prediction at the expense of having more tuning parameters.

Regularized optimization, which plays an important role in both statistical and machine learning problems, can often be described as:

$$\hat{\beta}(\lambda) = arg\min_{\beta} L(Y, X\beta) + \lambda J(\beta), \tag{14}$$

where $L(Y, X\beta)$ is a non-negative loss function, $J(\beta)$ is a non-negative penalty, and λ is the non-negative tuning parameter. Many popular methods fall into this category, such as: ridge regression [21], lasso [22], support vector machine [23] and elastic net [24]. It is known that many of these problems have a Bayesian interpretation, such that the loss function is interpreted as the negative log-likelihood, the penalty term is the negative log-prior density, and the regularized solution corresponds to the global maxima of the posterior distribution. Therefore, using different penalty terms corresponds to imposing different prior knowledge and/or assumptions on the estimated parameters and models. Ref. [25] proposed a generalized L_1 norm penalty framework which can be applied to PCA, PLS, canonical correlation analysis, and multivariate analysis of variance type of models. The generalized L_1 penalty enforces certain structural properties, such as sparsity, e.g. sparsity on pairwise differences between adjacent coefficients. One example of applying generalized L_1 penalty in PLS is to use the fused lasso penalty [26] to estimate the coefficients. The fused lasso contains both the lasso and fusion penalties. The former encourages sparsity in the coefficients, while the latter leads to interval selection, where the coefficients within each interval have the similar values. Instead of assuming sparsity on the coefficients and intervals, penalized signal regression (PSR) assumes or constrains smoothness on the coefficients vector. Unlike PSR, which has an explicit solution (and can be solved easily), the solution for the generalized L_1 penalty problems can be found by routines for inequality constrained quadratic programming and iterative procedures.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the referees for pointing out related work and for their constructive comments and suggestions that helped to improve the presentation of this paper. Portions of this research were conducted with high performance computational resources provided by the Louisiana Optical Network Infrastructure (http://www.loni.org).

This research was funded in part by the BL Allen Endowment in Pedology at Texas Tech University.

References

[1] W.S. DeSarbo, W.L. Cron, A maximum likelihood methodology for clusterwise linear regression, J. Classif. 5 (1988) 249–282.
[2] C. Preda, G. Saporta, Clusterwise PLS regression on a stochastic process, Comput. Stat. Data Anal. 49 (2005) 99–108.
[3] C. Preda, G. Saporta, "PCR and PLS for clusterwise regression on functional data," selected contributions in data analysis and classification, in: P. Brito, G. Cucumel, P. Bertrand, F. de Carvalho (Eds.), Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, Heidelberg, 2007, pp. 589–598.
[4] B.O. Muthén, Latent variable modeling in heterogeneous populations, Psychometrika 54 (1989) 557–585.
[5] P. Wang, Y. Liu, D. Shen, Flexible locally weighted penalized regression with applications on prediction of Alzheimer's disease neuroimaging initiative's clinical scores, IEEE Trans. Med. Imag. 38 (6) (2019) 1398–1408.
[6] D.D. Wang, S. Chakraborty, D.C. Weindorf, B. Li, A. Sharma, S. Paul, M.N. Ali, Synthesized use of VisNIR DRS and PXRF for soil characterization: total carbon and total nitrogen, Geoderma 243–244 (2015) 157–167.
[7] I.E. Frank, J.H. Friedman, A statistical view of some chemometric regression tools, Technometrics 35 (1993) 109–148.
[8] T. Hastie, C. Mallows, A statistical view of some chemometrics regression tools: Discussion, Technometrics 35 (1993) 140–143.
[9] J.S. Morris, Functional regression, Ann. Rev. Stat. Appl. 2 (2015) 321–359.
[10] B.D. Marx, P.H.C. Eilers, Generalized linear regression on sampled signals and curves: a P-spline approach, Technometrics 41 (1999) 1–13.
[11] P.H.C. Eilers, B.D. Marx, Flexible smoothing with B-splines and penalties (with comments and rejoinder), Stat. Sci. 11 (1996) 89–121.
[12] P.H.C. Eilers, B.D. Marx, Multivariate calibration with temperature interaction using two-dimensional penalized signal regression, Chemometr. Intel. Lab. 66 (2003) 159–174.
[13] P.H.C. Eilers, B.D. Marx, Practical Smoothing: the Joys of P-Splines, Cambridge University Press, Cambridge, 2021.
[14] I. Currie, M. Durbán, P.H.C. Eilers, Generalized linear array models with applications to multidimensional smoothing, J. Roy. Stat. Soc. B 68 (2006) 259–280.
[15] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[16] K.P. Murphy, Machine Learning: A Probabilistic Perspective, 14.4.3, The MIT Press, 2012, pp. 492–493.
[17] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, Soc. Ind. Appl. Math. 5 (3) (1984) 735–743.
[18] I.T. Jolliffe, A note on the use of principal components in regression, J. Roy. Stat. Soc. C. 31 (3) (1982) 300–303.
[19] P.H.C. Eilers, B. Li, B.D. Marx, Multivariate calibration with single-index signal regression, Chemometr. Intell. Lab. Syst. 96 (2009) 196–202.
[20] B.D. Marx, Varying-coefficient single-index signal regression, Chemometr. Intell. Lab. Syst. 143 (2015) 111–121.
[21] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.
[22] R. Tibshirani, Regression shrinkage and selection via the LASSO, J. Roy. Stat. Soc. B 58 (1996) 267–288.
[23] V. Vapnik, Statistical Learning Theory, John Wiley, New York, 1998.
[24] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. B 67 (2005) 301–320.
[25] M.A. Rasmussen, Generalized L_1 penalized matrix factorization, J. Chemometr. 31 (4) (2017), e2855.
[26] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, J. Roy. Stat. Soc. B 67 (2005) 91–108.