Electronic Journal of Statistics
Vol. 11 (2017) 4151–4178
ISSN: 1935-7524
DOI: 10.1214/17-EJS1301

Structured regression models for high-dimensional spatial spectroscopy data

Arash A. Amini, Elizaveta Levina and Kerby A. Shedden

Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1107
e-mail: aaamini@umich.edu; elevina@umich.edu; kshedden@umich.edu

Abstract: Modeling and analysis of spectroscopy data is an active area of research with applications to chemistry and biology. This paper focuses on modelling high-dimensional spectra for the purpose of noise reduction and prediction in problems where the spectra can be used as covariates. We propose a functional representation of the spectra as well as functional regression model that accommodates multiple spatial dimensions. Both steps emphasize sparsity to reduce the number of parameters and mitigate overfitting. The motivating application for these models, discussed in some detail, is predicting bone-mineral-density (BMD), an important indicator of fracture healing, from Raman spectra, in both the in vivo and ex vivo settings of a bone fracture healing experiment. To illustrate the general applicability of the method, we also use it to predict lipoprotein concentrations from spectra obtained by nuclear magnetic resonance (NMR) spectroscopy.

Keywords and phrases: Structured regression, functional data, spatial data, spectroscopy.

Received September 2016.

1. Introduction

Spectroscopic imaging is a technique for understanding the spatial variation in the composition of a material. It obtains a chemical spectrum at each pixel in the imaging field. As in traditional non-imaging spectroscopy, these spectra reflect the prevalence of different types of atomic bonds, which can ultimately be used to make inferences about the chemical composition of the material. Since a spectrum is obtained for every pixel, using this technique we are able to recover spatial variation in the materials chemical composition. This can be used, for example, to identify different types of bone in a tissue specimen that would appear to be homogeneous when using non-spectroscopic imaging. The data produced by spectroscopic imaging comprise a data cube that can be viewed either as a collection of 1d spectra indexed by two or three spatial dimensions, or as a stack of 2d (or 3d) images indexed by the position along the spectral axis (wave number or frequency). This results in a hybrid of functional and spatial data.

Our goal in this paper is to propose models for dealing with these challenging data types, especially in the context where the spectral image data is to be used as a predictor of an outcome (e.g. the mineral content of the bone). The work is applicable to any spectroscopic approach (e.g. infrared, nuclear magnetic resonance), but is motivated by Raman spectroscopy. Raman spectra arise from light and can be used transcutaneously as a non-invasive biomedical test. However Raman spectra are quite weak and therefore sensitive analytic methods are needed to recover meaningful changes in the spectra.

More specifically, we assume that, for each subject, we observe a collection of functional data of the forms $X_{ij}(t)$, where i and j index the spatial dimensions, and $t \in T \subset \mathbb{R}$ indexes the continuous functional dimension (e.g., spectral wavelengths). Further, for each subject we observe a response y and our objective is to effectively predict y based on the entire data $\{X_{ij}(t)\}$. Ignoring the spatial dimension, the problem can be formulated in the context of functional data analysis (FDA) [1, 2]. A line of work [3, 4, 5, 6, 7, 8, 9, 10] considers the natural situation where both the predictors and the response are functional, e.g., values of quantities that vary over time (longitudinal data). In this paper, we consider the case where the response is a scalar while the predictors are functional; we refer to [11, 12, 13, 14, 15] for a sample of work in this area. From the FDA perspective, our setting is similar to that of [15], where a functional regression model, with real-valued response, of the form $y = \int X(t)\beta(t)dt + \varepsilon$ is considered. Here $\beta(t)$ is the coefficient function and ε is the additive noise. It is clear that one needs some restrictions on β, otherwise one can obtain a perfect fit to any scalar y, based on a finite sample. In [15], the authors point out that there are in general two approaches to dealing with a functional coefficient vector to avoid overfitting: modeling $\beta(t)$ via a truncated basis function representation, or adding smoothness penalties of the form $\int \beta^{(r)}(t)dt$ to the least-squares problem, where $\beta^{(r)}$ is the rth derivative of β. They propose a hybrid approach based on a basis function representation and sparsity constraints on the derivatives $\beta^{(r)}$, over a fine grid of points. See [16] for a nice survey of scalar-on-function regression.

Our approach here can also be considered a hybrid, but with some notable differences from [15]. First, we model the functional covariates as elements of a reproducing kernel Hilbert space (RKHS) [17]. This allows us to obtain efficient representations of the spectra as sparse nonnegative combinations of the kernel functions. The nonnegativity assumption is natural when dealing with spectra; it helps interpretation of the coefficients and results in a very different representation than that based on basis functions. The RKHS can be chosen so that its natural norm measures the smoothness, and we can then control the smoothness of the representation by adding a norm penalty. Alternatively, the kernel choice can be motivated by physical considerations: for example, physics of scattering suggests the Lorentzian (cf. (4)) for Raman spectra. Our penalized approach to representation is also an efficient noise reduction mechanism, since $\{X_{ij}(t)\}$ are often very noisy. This is especially important for Raman spectroscopy, since Raman scattering is a very weak effect.

An important contribution of our approach is the joint modeling of spatial and spectral dimensions in a regression model. From this perspective, our approach is related to recent work on matrix or tensor regression with a low-rank coefficient matrix [18, 19, 20]. Specifically, we impose an explicit rank-one assumption on the coefficient viewed as a tensor indexed by (i, j, t), with a model of the form

$$y = \sum_{ij} \int_T \alpha_i \beta_j \widehat{X}_{ij}(t) + \varepsilon \tag{1}$$

where $\widehat{X}_{ij}(t)$ is our functional covariate representation. The explicit rank-one assumption dramatically reduces the dimension of the parameter space (i.e., model complexity), guarding against overfitting. Among the challenges posed by this model is the existence of cross-terms in the discrete version of (1) due to non-orthogonality of the components of $\widehat{X}_{ij}(t)$, and the nonconvex nature of the resulting least squares problem. Nevertheless, we show that a simple alternating minimization approach can be used to fit the model, and that our approach is effective in predicting responses based on various spectroscopy data, when a functional relation exists. We also provide a case study of an in vivo Raman experiment, which served as the original motivation for this model, and show that in this case the model can be used to reveal anomalies in the dataset, which were later traced back to an experimental flaw. In this experiment, $n = 37$ rats were considered, and the Raman spectra were obtained on $N = 544$ wavenumbers, for each rat, by rotating a source-detector pair over 5×10 possible positions around a ring. This led to a high-dimensional, very noisy, tensor of size $5 \times 10 \times 544$ per rat. The goal was to use the Raman tensor to build a predictive model of bone mineral density (BMD). The denoising and implicit dimension reduction of our approach helps reduce the potential for overfitting in building models based on such high-dimensional tensors, relative to the sample size (i.e., $p = 27200$ covariates relative to $n = 37$). We refer to Section 3.2 for more details on the in vivo experiment.

The indices i and j in model (1) need not literally reflect orthogonal spatial axes, which would imply a spatial structure with primarily horizontal and vertical patterning. Instead, i and j can index axes in any transformed coordinate system, for example resulting from 2d Fourier or wavelet transforms of the original image. Alternatively, the data may suggest a natural 2-dimensional parameterization, such as the ring of source/detector pairs used in the in-vivo Raman data considered in Section 3.2. In general, we only require that the spatial variation be captured through two ordered sequences of parameters, corresponding to α_i and β_j in model (1), with the mean structure being separately convex in (α_i) and (β_j).

We note that multilinear or tensor regression models have successfully been used in the past for modeling longitudinal data with matrix covariates [21] and in imaging applications where each image can be considered a matrix-valued covariate [22, 23, 24]. In general, a tensor model with reduced-rank coefficients is a fairly natural extension of a regular regression model to covariates of dimension ≥ 2 (see also Remark 2). Our model also takes this natural approach and combines it with the functional representation of the spectral dimension of the tensor. In other words, the novelty relative to a simple tensor regression setup is that we are dealing with tensors with both discrete and functional indices (mixed tenors) and provide an effective representation of the functional dimension suitable for dealing with nonnegative spectra. This latter aspect is also what separates our approach from the popular approaches in the family of partial-least squares (PLS) and principal component regression (PCR) [25]. Our representation of the spectra, via kernels of nonnegative weights, keeps the coefficients in a roughly one-to-one correspondence with original wavenumbers. In contrast, the coefficients in PLS or PCR are weights for the "learned" basis functions, and individually not very meaningful. The basis functions are usually orthonormal and not restricted to be nonnegative and have no spectral interpretations. While PLS family is good for prediction, we believe our approach has the added advantage of giving interpretable weights which in many examples are necessary for understanding which parts of the spectra affect certain phenomena. For this reason (i.e., very different representation of spectra), we have not directly compared to the PLS family.

The paper is organized as follows. Section 2 presents our models, starting with our representation of the data in Section 2.1 which takes into account its functional nature. The representation, based on a functional version of the Lasso [26], simultaneously achieves denoising and compression. Our main regression model, discussed in Section 2.2, builds on the functional representation and takes into account the tensor aspect of the spatial spectroscopy data. In Section 3, we apply our methods to three spectroscopy datasets. The first is a fracture healing experiment, in which spatial in vivo Raman data were obtained from a collection of rats, with the aim of predicting progression of healing, as measured by the bone mineral density (BMD). The second example is ex vivo microscopy data from the same fracture healing experiment. The ex vivo data have higher signal-to-noise ratio than the in vivo data and are expected to provide much more accurate estimates of the BMD. Our final example is an NMR spectroscopy experiment with the aim of predicting lipoprotein concentrations. The paper concludes, in Section 4 with a discussion of present shortcomings and possible extensions of this work.

2. Models and methods

We consider data that have one continuous (functional) dimension and several discrete (spatial) dimensions. In order to describe the model in concrete terms, we focus on the case of two discrete dimensions and a single continuous one, although the results can be easily generalized. To be specific, the data is a collection

$$\big\{X_{ij}^k(\cdot) : (i,j,k) \in [d_1] \times [d_2] \times [n]\big\} \tag{2}$$

of functions, recorded at spatial positions (i,j). Here, we are using the notation $[d_1] := \{1,\dots,d_1\}$ and similarly for $[d_2]$ and $[n]$. Index k is used to enumerate the samples which we assume to be i.i.d. across k. We will use t to denote the continuous index, and assume that each function $t \mapsto X_{ij}^k(t)$ is observed in some interval $T \subset \mathbb{R}$, typically at discrete points $\{t_1, \dots, t_N\} \in T$. Since our main application here is spectroscopy, we may refer to the continuous domain T as the spectral domain and to its elements as wavenumbers, though the model is general.

We consider two analysis goals: finding a compact and informative representation of the data, and using the high-dimensional dataset $\{X_{ij}^k(t)\}$ as covariates in a regression problem to predict a response vector $\{y^k\}$. These goals are related, since obtaining a compact representation of the data greatly facilitates the regression analysis.

2.1. Functional representation

In order to obtain a compact representation of the data, we assume that each function $X_{ij}^k(\cdot)$ lies in a reproducing kernel Hilbert space (RKHS) [17], generated by some kernel function $\mathbb{K} : T \times T \to \mathbb{R}_+$. Usually, the functions $\{X_{ij}^k(\cdot)\}$ are only observed at a discrete set of points $\mathcal{T} := \{t_1, t_2, \dots, t_N\}$. We additionally assume that each $X_{ij}^k(\cdot)$ can be well approximated by a finite linear combination of the kernel functions anchored at points of $\mathcal{T}$. That is,

$$X_{ij}^k(t) \approx \widehat{X}_{ij}^k(t) := \sum_{v=1}^{N} \widehat{x}_{ijv}^k \, \mathbb{K}(t, t_v), \quad t \in T. \tag{3}$$

This assumption simplifies the subsequent derivations and is motivated and to some extent justified by the representer theorem [27]. The kernel function $\mathbb{K}$ can be taken to be any valid kernel (positive semidefinite, symmetric bivariate function), though our main focus will be on the Lorentzian

$$\mathbb{K}(t, s) = \frac{1}{1 + (\frac{t-s}{W})^2} =: \mathbb{L}(t - s; W), \tag{4}$$

where W is a bandwidth parameter. In spectroscopy, it has been found empirically that this kernel provides a good model for spectra obtained from a chemically pure sample, and is also justified by physical considerations [28]. Another restriction that we impose is for the coefficients $\{x_{ijv}^k\}$ to be nonnegative. This is also in accordance with the physics of how spectra are formed as a weighted linear combination of spectra of pure chemical components, without any cancellations.

Let us fix (i, j, k) for the rest of this section. Based on (3), the idea is to turn the collection $\{X_{ij}^k(t)\}_{t \in T}$ into the vector of coefficients

$$\widehat{x}_{ij\bullet}^k := (\widehat{x}_{ijv}^k, \, v \in [N]) \in \mathbb{R}^N$$

which is easier to work with. To achieve a compact representation, we impose a sparsity constraint on the vector $x_{ij\bullet}^k$, seeking a representation of the form (3) with as few nonzero coefficients as possible. Let

$$\widehat{X}_{ij}^k(t) = \varphi(t; \widehat{x}_{ij\boldsymbol{\cdot}}^k) \quad \text{where} \quad \varphi(t;z) := \sum_{v=1}^{n} z_v \mathbb{K}(t, t_v),$$

and let $\|f\|_{2,N} := [\sum_{v=1}^N f^2(t_v)]^{1/2}$ be the empirical L^2 norm. The sparse representation $\widehat{x}_{ij\boldsymbol{\cdot}}^k$ can be obtained by solving the following ℓ_1-regularized least-squares problem

$$\widehat{x}_{ij\boldsymbol{\cdot}}^k = \operatorname*{argmin}_{z \,\in\, \mathbb{R}_+^N} \Big\{ \frac{1}{2} \|X_{ij}^k(\boldsymbol{\cdot}) - \varphi(\,\boldsymbol{\cdot}\,; z)\|_{2,N}^2 + \lambda_{\mathcal{H}} \|\varphi(\,\boldsymbol{\cdot}\,; z)\|_{\mathcal{H}}^2 + \lambda_1 \sum_{v=1}^N |z_v| \Big\}, \tag{5}$$

where $\mathbb{R}_+^N$ is the set of N-vectors with nonnegative components, and $\|f\|_{\mathcal{H}}$ denotes the RKHS norm. When the RKHS norm measures roughness of the function, regularizing by this norm leads to smoother solutions. Problem (5) can be written in the expanded form

$$\operatorname*{argmin}_{z \,\in\, \mathbb{R}_+^N} \Big\{ \frac{1}{2} \sum_{u=1}^N \Big[X_{ij}^k(t_u) - \sum_{v=1}^N z_v \ \mathbb{K}(t_u, t_v) \Big]^2 + \lambda_{\mathcal{H}} \sum_{u,v=1}^N z_u z_v \mathbb{K}(t_u, t_v) + \lambda_1 \sum_{v=1}^N |z_v| \Big\}. \tag{6}$$

Let K be the $N \times N$ symmetric matrix with entries $\mathbb{K}(t_u, t_v)$, and let $\|z\|_p := \big(\sum_{v=1}^N |z_v|^p\big)^{1/p}$ denote the ℓ_p norm of $z = (z_1, z_2, \ldots, z_N)$. Moreover, let $\boldsymbol{x}_{ij}^k := \big(X_{ij}^k(t_u),\, u \in [N]\big)$ so that $\boldsymbol{x}_{ij}^k$ is an N-vector. Then, (6) can be rewritten in the compact form

$$\widehat{x}_{ij\boldsymbol{\cdot}}^k \in \operatorname*{argmin}_{z \,\in\, \mathbb{R}_+^N} \Big\{ \frac{1}{2} \|\boldsymbol{x}_{ij}^k - Kz\|_2^2 + \lambda_{\mathcal{H}}\, z^T K z + \lambda_1 \|z\|_1 \Big\}. \tag{7}$$

This is a standard convex problem which can be solved efficiently. Figure 2.1 shows examples of fitted Raman spectra.

2.2. Regression model

In this section we devise a model to predict a one-dimensional response $\{y^k\}$ based on the observed tensor covariates $\{X_{ij}^k(t)\}$. Since we expect the covariates to be noisy, we in fact base the model on the denoised covariates $\{\widehat{X}_{ij}^k(t)\}$, as derived in Section 2.1. Perhaps the simplest model is a rank-one multilinear map relating the covariates to the response, that is,

$$y^k = \sum_{ij} \int_T \alpha_i \beta_j w(t)\, \widehat{X}_{ij}^k(t)\, dt + \varepsilon^k \tag{8}$$

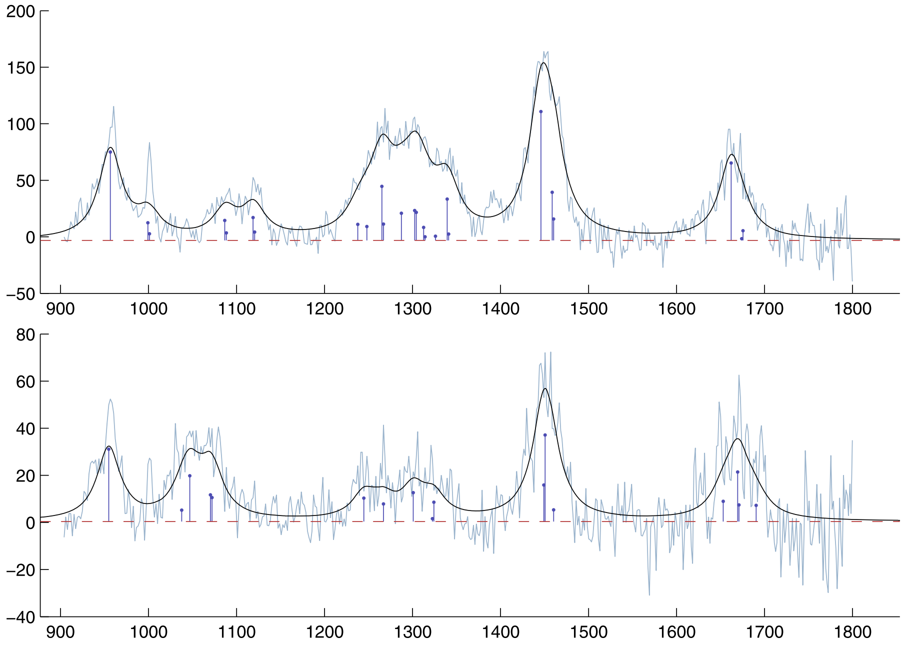


FIG 1. *Examples of fitting model* (3) *to Raman spectra. The fitted spectra are shown in black; the estimated non-zero coefficients* $\{\widehat{x}^k_{ijv}\}$ *are shown as vertical lines, with height representing their magnitude.*

for $k = 1, \ldots, n$, where $\{\varepsilon^k\}$ are i.i.d. noise variables. In accordance with (3), we simplify the model further by assuming the following representation for w,

$$w(t) = \sum_{u=1}^{N} \gamma_u \mathbb{K}(t, t_u), \quad t \in T. \tag{9}$$

Combining with representation (3) for $\{\widehat{X}^k_{ij}(t)\}$ and ignoring its approximation error, we arrive at the model

$$y^k = \sum_{ijuv} \alpha_i \beta_j \gamma_u \, G_{uv} \, \widehat{x}^k_{ijv} + \varepsilon^k$$

where $G_{uv} := \int_T \mathbb{K}(t, t_u)\mathbb{K}(t, t_v)\, dt$. Note that this is the $L^2(T)$ inner product of the functions $\mathbb{K}(\cdot, t_u)$ and $\mathbb{K}(\cdot, t_v)$. Let $G := (G_{uv}) \in \mathbb{R}^{N \times N}$, and note that this is a Gram matrix. The model can be written more compactly as

$$y^k = \sum_{iju} \alpha_i \beta_j \gamma_u \, \widetilde{x}^k_{iju} + \varepsilon^k, \quad \text{where} \quad \widetilde{x}^k_{iju} = \sum_{v=1}^{N} G_{uv} \, \widehat{x}^k_{ijv}, \tag{10}$$

and the summation is over $(i, j, u) \in [d_1] \times [d_2] \times [N]$. Note that the advantage of this rank-one model is that it contains $d_1 + d_2 + N$ variables, which in high dimensions is far less than that of a full linear model with $d_1 d_2 N$ variables.

In order to fit model (10), we solve a regularized least-squares problem. Since (i,j) represents spatial dimensions, we do not expect sparsity in $\alpha = (\alpha_i)$ and $\beta = (\beta_j)$. We regularize α and β by imposing constraints $\sum_i \alpha_i = \sum_i \beta_i = 1$. This gives α and β the natural interpretation of being probability vectors. On the other hand, we expect considerable sparsity in the functional (spectral) domain variable $\gamma = (\gamma_u)$, thus we regularize by its ℓ_1 norm.

The final element of our proposed regularizer is a penalty which tends to bring the coefficients assigned to nearby points in T closer together. This is justified if proximity in the functional domain signifies similar impact on the response variable, as is the case for many functional data including spectra. This regularization also provides a practical advantage which is discussed in Section 2.3. As a measure of similarity, we can use the Gram matrix G. We consider two possibilities:

(a) Weighted fused Lasso [29, 30]: A penalty of the form

$$f_G(\gamma) = \rho_G \sum_{uv} G_{uv} |\gamma_u - \gamma_v|, \quad \text{for some } \rho_G > 0. \tag{11}$$

(b) Exact fused Lasso: A constrained version of the above where we force equality among (γ_u); more explicitly:

$$f_G(\gamma) = \delta_C(\gamma), \quad \text{for} \quad C = \{\gamma : \ \gamma_u = \gamma_v \ \text{ if } \ G_{uv} > \tau\} \tag{12}$$

where δ_C is the indicator of set C in the sense of convex analysis (i.e., $\delta_C(x) = 0$ if $x \in C$, otherwise $= \infty$), and $\tau \in [0,1]$ is a threshold controlling the degree of regularization.

Unless otherwise stated, in the numerical experiments, the weighted form (a) is used. The exact form (b) is discussed in Remark 3 below and in the context of synthetic data model of Section 3.1.

Let $\mathbb{D}^d := \{v \in \mathbb{R}^d : \ v \geq 0, \ \sum_i v_i = 1\}$ be the probability simplex in $\mathbb{R}^d$. Putting the pieces together, we solve the following problem:

$$(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}) = \underset{\substack{\alpha \in \mathbb{D}^{d_1}, \\ \beta \in \mathbb{D}^{d_2}, \\ \gamma \in \mathbb{R}^N}}{\operatorname{argmin}} \ \frac{1}{2n} \sum_{k=1}^{n} \Big(y^k - \sum_{iju} \alpha_i \beta_j \gamma_u \, \widetilde{x}^k_{iju} \Big)^2 + \rho_\gamma \, \|\gamma\|_1 + f_G(\gamma) \tag{13}$$

where ρ_γ and either of ρ_G or τ (in f_G) are tuning parameters chosen by cross-validation. Note that there is a scale ambiguity in model (10) since (α, β, γ) and $(c_1\alpha, c_2\beta, c_2\gamma)$ give the same map as long as $c_1 c_2 c_3 = 1$, but not in (13) due to the presence of the regularizers. (Also, note that there is no sign ambiguity.)

Remark 1. *Enforcing an exact rank-one constraint on the regression function has been recently proposed in [31] in the context of regression with matrix covariates. Their model is similar to* (8)*, without the functional dimension, i.e.,* $y^k = \sum_{ij} \alpha_i \beta_j X^k_{ij} + \varepsilon^k$. *They enforce sparsity on both sets of coefficients* (α)

and (β) using a multiplicative ℓ_1 penalty. Our model is a more general variant for the tensor case with mixed functional and discrete dimensions. In our application, the coefficients (α) and (β) have physical (spatial) interpretations and are not expected to be sparse.

Remark 2. *Enforcing a low-rank assumption on the regression function has been studied extensively in recent years. The preferred approach to the problem is via nuclear norm penalization. We refer to [18, 19] for more details. Empirically, we found that imposing a rank-one assumption directly enhances the interpretation by considerably reducing the dimensionality.*

Although we have considered a rank-one model in (8), the idea can be easily extended to a higher-rank version. The number of parameters for the rank r model is about $r(d_1 + d_2 + N)$, a rough measure of the complexity of the model. One has to consider the sample size relative to this complexity in choosing r. For most applications in spectroscopy, the sample size is usually quite small, making the choice of $r = 1$ almost inevitable. For example, in the in vivo Raman case, after reduction of the wavenumbers, we have $N \approx 40$, $d_1 = 5$ and $d_2 = 10$, giving a total parameter count of ≈ 55 for the rank-one model relative to the sample size $n = 37$. The rank-two model has $\approx 110 \gg 37$ which requires severe penalization (hence severe shrinkage effect on nonzero parameters) to combat over-fitting.

Remark 3. *An advantage of the exact fused Lasso formulation (12) is the potential for dramatic reduction in the number of parameters. Note that this penalty is equivalent to a set of homogeneous linear constraints on γ of the form $A\gamma = 0$, for a matrix A. An equivalent representation is $\gamma = V\theta$ where the columns of $V \in \mathbb{R}^{N\times q}$ form a basis for the null space of A. Here, q is the dimension of this null space and is related to how many pairs (u, v) satisfy $G_{uv} > \tau$. Thus, in the case of the exact fused Lasso, (13) can be written as*

$$(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}) = \underset{\substack{\alpha \in \mathbb{D}^{d_1}, \\ \beta \in \mathbb{D}^{d_2}, \\ \theta \in \mathbb{R}^q}}{\operatorname{argmin}} \; \frac{1}{2n}\sum_{k=1}^{n}\Big(y^k - \sum_{iju} \alpha_i \beta_j (V\theta)_u \, \widetilde{x}^k_{iju}\Big)^2 + \rho_\gamma \, \|V\theta\|_1 \tag{14}$$

For sufficiently large τ, the "effective dimension" q of the spectral parameter θ is much smaller than N, providing a better guard against overfitting and leading to faster computation. For small q, one might even achieve enough regularization in the spectral domain that ℓ_1 penalty is not needed (i.e., $\rho_\gamma = 0$).

Convex tensor regularizers. Very recently, and while this paper was under review, in an interesting work, [20] studied some convex regularizations in the context of high-dimensional tensor regression, and provided theoretical analysis of these approaches. Since their work is very relevant to our setup, we briefly review two of their approaches to third-order tensor regression. The model is

$$y^k = \langle B, \widetilde{X}^k \rangle + \varepsilon^k, \quad k = 1, \dots, n \tag{15}$$

where $\widetilde{X}^k = (\widetilde{x}^k_{iju}) \in \mathbb{R}^{d_1 \times d_2 \times N}$ and $B \in \mathbb{R}^{d_1 \times d_2 \times N}$ are third-order tensors, the latter being a parameter to be estimated. Here, $\langle B, \widetilde{X}^k \rangle = \sum_{iju} B_{iju} \widetilde{x}_{iju}$ is the natural inner product on $\mathbb{R}^{d_1 \times d_2 \times N} \simeq \mathbb{R}^{d_1 d_2 N}$, hence (15) is the natural linear regression model. In [20, Section 5.1], the following M-estimators are considered:

$$\widehat{B} \in \underset{A \,\in\, \mathbb{R}^{d_1 \times d_2 \times N}}{\operatorname{argmin}} \; \frac{1}{2n} \sum_{k=1}^{n} \| y^k - \langle A, \widetilde{X}^k \rangle \|_F^2 + \lambda \sum_{u=1}^{N} \| B_{**u} \|_a \tag{16}$$

where $B_{**u} \in \mathbb{R}^{d_1 \times d_2}$ is the slice of the tensor at third index u, and we either take $\| \cdot \|_a := \| \cdot \|_F$, the Frobenius norm, or $\| \cdot \|_a := \| \cdot \|_*$, the nuclear norm (i.e., the sum of singular values). We will refer to these two approaches as `grp_lasso_fro` and `grp_lasso_nuc`, respectively.

With the choice of the Frobenius norm, the regularizer is a group lasso penalty on B, imposing sparsity on entire slices of the form $B_{**u}, u = 1, \dots, N$; that is, the penalty is useful when one expects, for any given u, that either B_{**u} has mostly nonzero entries, or almost all its entries are zero. This is very much what we expect with the spectra: entire measurements for a given wavenumber u are either relevant or irrelevant.

The choice of nuclear norm penalty imposes a slice-wise low-rank restriction, i.e., one tends to get slices B_{**u} that have low rank. (This is well-known and can been seen by noting that the nuclear norm is the ℓ_1 norm of the spectrum of the matrix.) This type of restriction is exactly the assumption that we made in this work, with one major difference: In the extreme case of large λ, one hopes that `grp_lasso_nuc` forces the rank of most slices $B_{**u}, u \in [N]$ to be the minimal possible value, i.e. 1. However, there is no guarantee that one can simultaneously achieve a low rank for all the slices (e.g., it is very unlikely that one gets all B_{**u} to be of rank r, for a small r, at the same time).

Even when one has such a solution, one obtains different low rank solutions for each slice B_{**u}. This is in contrast to our model where a single rank-one (or rank r) model is enforced for all the slices. One might consider the possible variation in low-rank solutions an advantage of `grp_lasso_nuc`, and in some applications this might indeed be the case. However, in cases with an extremely impoverished sample size as ours, a simple parameter counting favors our approach: In the extreme case, assuming all slices can be forced to have the same rank r, `grp_lasso_nuc` will have at most $r(d_1 + d_2)N$ free parameters, whereas our approach (generalized to rank r) will have at most $r(d_1 + d_2 + N)$, which is much smaller. In Section 3.1, we compare these approaches on synthetic data and empirically confirm these observations in very low sample size settings.

2.3. Optimization and computational considerations

Let us consider some practical issues regarding the models and methods discussed earlier. The cost function in (13) is not (jointly) convex in (α, β, γ), but it is separately convex in each of these variables. A standard way to optimize

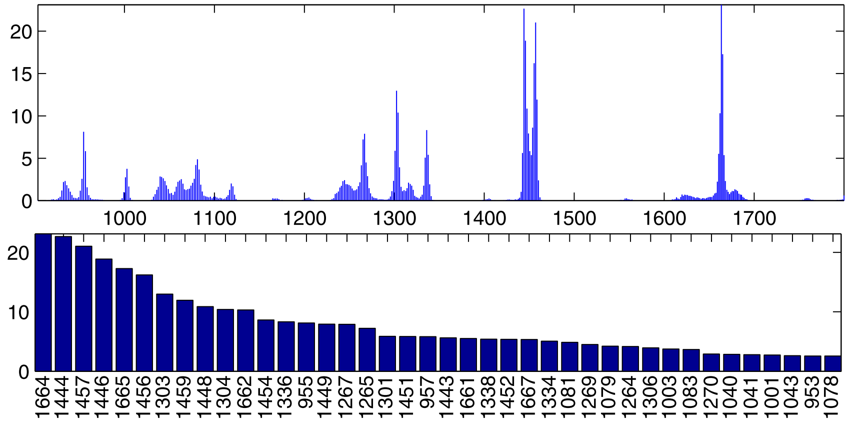


FIG 2. *Scores obtained from* (17) *for a sample of Raman data. The bottom plot shows the* 40 *wavenmubers with the highest scores.*

such functions is by alternating minimization, fixing two variables at a time and minimizing over the other.

In practice, we might observe $\{X^k_{ij}(\cdot)\}$ at more points than we want to keep in the model, observing the functions at a set of points $\mathcal{T}' \supset \mathcal{T}$ and only using the observations from the set $\mathcal{T}$ in fitting the model. This is primarily done to avoid over-fitting by reducing model complexity, though it also helps reduce the overall computational cost. One approach to choose which points to keep is to assign a score to each point in $\mathcal{T}' = \{t'_1, t'_2 \dots, t'_{N'}\}$, based on their weighted frequency of appearing in the dataset, or simply the frequency, $\frac{1}{pdn}\sum_{ijk} 1\{x^k_{ijv} \neq 0\}$. For example, we can assign the following score to t'_v,

$$s_v := \frac{1}{d_1 d_2 n} \sum_{ijk} \widehat{x}^k_{ijv}, \quad v \in [N'] \tag{17}$$

where $\{\widehat{x}^k_{ijv}\}$ are the coefficients in expansion (3) of Section 2.2, and the sum runs for $(i,j,k) \in [d_1] \times [d_2] \times [n]$. We can then keep wavenumbers corresponding to the N largest scores. Figure 2.3 shows the scores obtained for the example Raman data and the wavenumbers corresponding to $N = 40$ largest scores, out of $N' = 544$. An advantage of the weighted fused Lasso penalty used in (13) is that we do not need to worry about the order of the wavenumbers kept in $\mathcal{T}$. The Gram matrix G automatically tries to match the coefficients of nearby wavenumbers regardless of how they are ordered in $\mathcal{T}$.

Although theoretically not necessary, in practice some crude normalization of covariates is useful prior to fitting the model. For example, one can normalize so that the maximum amplitude for a particular spatial combination is 1, that is, work with the normalized sequence $\{\widetilde{x}^k_{iju}/(\max_{k,u} \widetilde{x}^k_{iju})\}$. Other variations, such as ℓ_2 normalization, are possible and empirically lead to comparable results.

The results reported in the following are based on the maximum amplitude normalization, unless stated otherwise.

3. Applications to spectroscopy

3.1. Synthetic data

Data generating model. Let us first describe a simplified data generating model that captures the spectroscopy applications we have in mind. We will then evaluate the performance of the proposed framework on simulated data from this model. Assume for simplicity that the spatial domain is the unit square $[0,1]^2$ and that we partition this space into blocks, or cells, of equal size, each with side length $1/d$. We assume that the spatial features are constant over the cells. Letting u be the spatial index, we can set $\xi = (i,j)$ where $i,j = 1,\dots,d$. We assume that the spatio-spectral signal is generated as follows:

$$f(\xi,t) = (\mu_\xi + w_{0\xi})_+ h_0(t) + (w_{1\xi})_+ h_1(t), \quad t \in T. \tag{18}$$

Here $h_0(\cdot)$ is the spectrum of the compound of interest, for example, the bone in Raman experiments described below and $h_1(\cdot)$ is the spectrum of background material, e.g., collagen in the Raman case. We refer to these as the signal and the clutter, respectively. The coefficient $(\mu_\xi + w_{0\xi})_+$ models the spatial variation in the signal, with μ_ξ modeling the mean and $w_{0\xi}$ the variation about the mean. Here $(\cdot)_+$ is the positive part of a number. We assume that $\mu_\xi = 1\{\xi \in S\}$ (indicator of S) where $S \subset [d]^2$ is a randomly generated subset with cardinality $|S| = s \ll d^2$, hence the signal is spatially sparse. We model $w_{k\xi}, k = 0,1$ as spatially correlated zero-mean Gaussian process with:

$$w_{k\xi} \sim N(0,\sigma_k^2), \quad \mathbb{E}[w_{k\xi} w_{k\xi'}] = \sigma_k^2 \exp\Big[-\frac{\delta(\xi,\xi')^2}{\tau_k^2}\Big], \quad k = 0,1$$

where $\delta(\xi,\xi') = \sqrt{|i-i'|^2 + |j-j'|^2}$, is the ℓ_2 distance between $\xi = (i,j)$ and $\xi' = (i',j')$. We assume that the two processes $(w_{0\xi})$ and $(w_{0\xi'})$ are independent.

We assume that the signal+clutter spectra are observed indirectly via the following measurement process: An excitation device moves along the i-axis and a detection device along the j-axis. The excitation device at position i_0 generates a field (in the case of Raman, a laser producing light) and the detector at position j_0 collects the response. Consider a general model of attenuation where the field attenuates along i-axis with profile $\alpha_{ii_0} = A_{i_0} q_1(i - i_0)$ and the response along the j-axis with profile $\beta_{jj_0} = B_{j_0} q_2(j - j_0)$ for some function q_1 and q_2. Thus, we can model the observed signal at joint position $\xi_0 = (i_0, j_0)$ for excitation and detection devices as

$$X_{\xi_0}(t) = \sum_{ij} \alpha_{ii_0} \beta_{jj_0} \big[f(\xi,t) + e(\xi,t)\big], \quad \xi = (i,j),\ \xi_0 = (i_0,j_0). \tag{19}$$

where $e(\xi,t) \equiv e(i,j,t)$ is the noise modeled as i.i.d. samples (across i,j and t) from $N(0,\sigma_X^2)$. (19) describes the a general convolution model for the observed

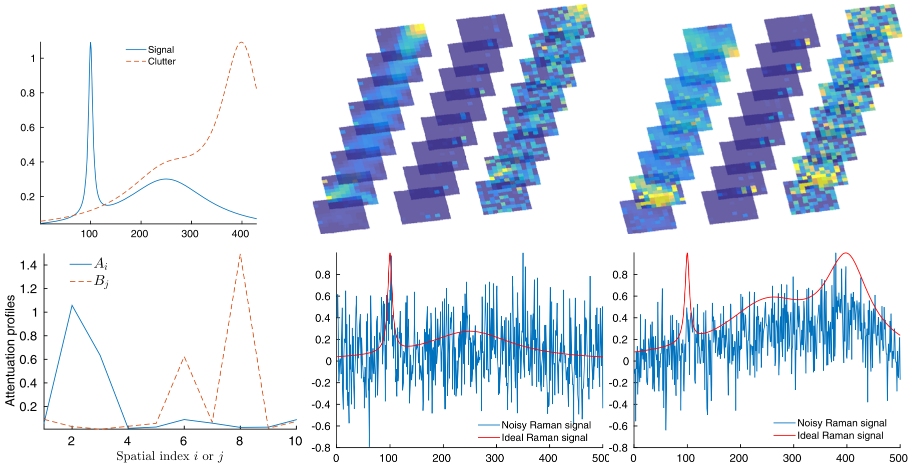


FIG 3. *Data generated from the synthetic model of Section 3.1 with $d = 10$, $N = 500$ and $n = 50$. . (Top-left) The spectral profiles of the signal $h_0(t)$ and clutter $h_1(t)$. (Bottom-left) The random attenuation factors A_i and B_j along the two spatial dimensions, each with $d = 10$ cells. (Top-middle and right) The 3D nature of spatio-spectral signals illustrated by showing $d \times d$ slices at $t = 50, 100, \ldots, 400$. This is done for two different samples $k = 13$ (Top-middle) and $k = 23$ (Top-right). For each sample, three different versions are shown, the noiseless version $f^k(\xi, t)$, the observed version $X^k_\xi(t)$ and the optimal amplitude-corrected version $(A_iB_j)^{-1}X^k_\xi(t)$ where $\xi = (i, j)$. Note the variation in spatial patterns of the signal and clutter across samples. (Bottom-middle and right) Individual spectra at a particular spatial location $\xi = (6, 3)$, both the ideal version $f^k((6, 3), t)$ and the noisy amplitude-corrupted version $X^k_{(6,3)}(t)$ for the two samples $k = 13$ (Bottom-middle) and $k = 23$ (Bottom-right).*

spectra. In the sequel, we focus on the ideal case where the attenuation profiles are $q_1(i - i_0) = 1\{i - i_0 = 0\}$ and $q_2(j - j_0) = 1\{j - j_0 = 0\}$. In this case, there is no mixing from neighboring cells and we arrive at the model

$$X_{\xi_0}(t) = A_{i_0} B_{j_0} \big[f(\xi_0, t) + e(\xi_0, t)\big], \quad \xi_0 = (i_0, j_0). \tag{20}$$

We will assume that A_{i_0} and B_{j_0} are randomly generated unknown quantities. This simplified model captures the difficulty of the model to the first-order, through unknown amplitudes A_{i_0} and B_{j_0}, and is a reasonable approximation in cases where q_1 and q_2 are sharply concentrated around 0. Let $\boldsymbol{X} = \{X_{\xi_0}(\cdot) \equiv X_{i_0,j_0}(\cdot) : \ i_0, j_0 \in [d]\}$ be the collection of spatio-spectral observations.

In order to have a regression model, we need a response variable. Here, we assume it to be the aggregate (across space) of the underlying signal amplitude:

$$y = \sum_{\xi} (\mu_\xi + w_{0\xi})_+ \equiv \sum_{ij} (\mu_{ij} + w_{0ij})_+. \tag{21}$$

Equations (18), (20) and (21) fully describe our generating model. We assume that we have n i.i.d. samples $(y^k, \boldsymbol{X}^k) \sim (y, \boldsymbol{X}), k = 1, \ldots, n$ from this model. Note that $\{\boldsymbol{X}^k : k \in [n]\} \equiv \{X^k_{i_0,j_0}(\cdot) : (i_0, j_0, k) \in [d] \times [d] \times [n]\}$ corresponds

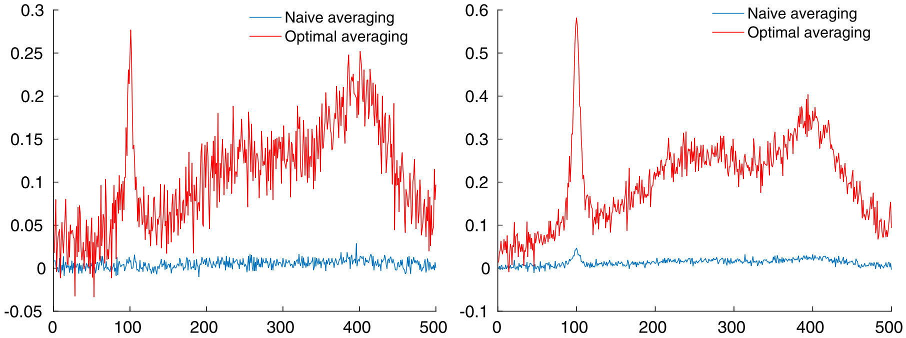


FIG 4. *The optimal versus naive spatial averaging of spectra, namely,* $d^{-2}\sum_\xi (A_iB_j)^{-1}X_\xi^k(t)$ *versus* $d^{-2}\sum_\xi X_\xi^k(t)$ *respectively, for the two samples* $k = 13$ *(Left) and* $k = 23$ *(Right).*

to (2) with $d_1 = d_2 = d$, which has been assumed here for simplicity. We note that only h_0, h_1 and $\{A_i, B_j\}$ remain the same across samples; all other parameters are generated independently for each sample, including $\mu_\xi^k, w_{0\xi}^k, w_{1\xi}^k$, changing the spatial patterns of signal and clutter in each sample (cf. Figure 3).

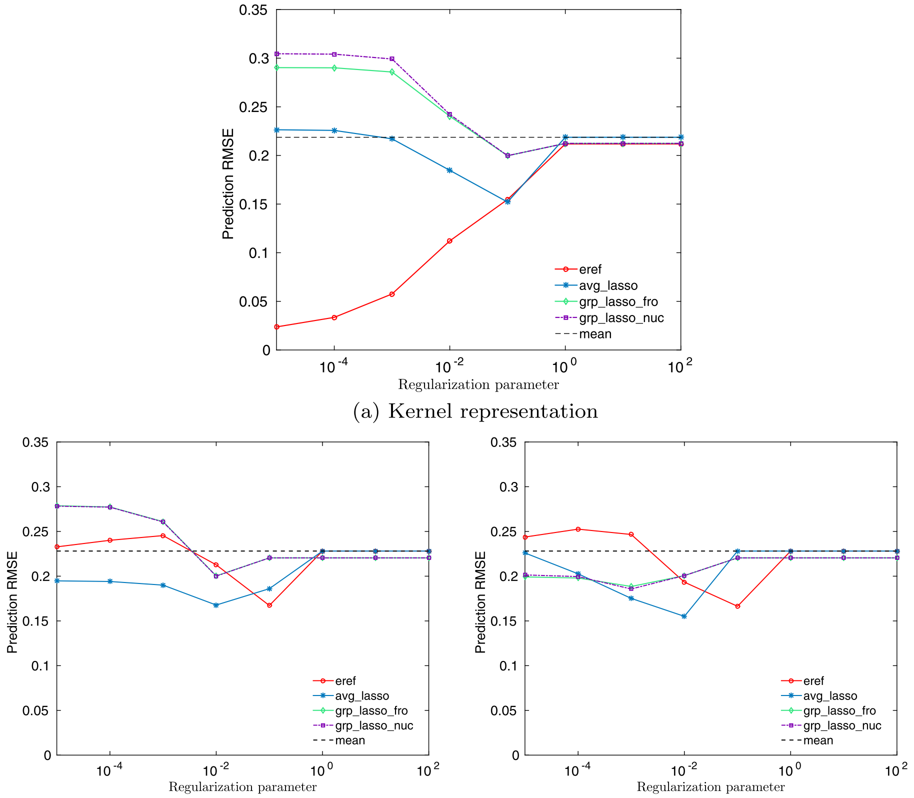
What synthetic data looks like. Figure 3 shows an example of the data generated from this synthetic model. We have taken $N = |T| = 500$ (the number of spectral indices), $d = 10$ and $n = 50$ (the number of samples). The signal and clutter are generated using combinations of Lorentz kernels, cf. (4), with one distinct component and one shared component, namely, $h_0(t) = \mathbb{L}(t-100; 5) + 0.3\mathbb{L}(t-250; 100)$ and $h_1(t) = \mathbb{L}(t-400; 50) + 0.3\mathbb{L}(t-250; 100)$. The signal is mostly sharply concentrated around $t = 100$, while the clutter is vaguely concentrated around $t = 400$ and the two have an overlapping component at $t = 250$. The randomly generated amplitudes $\{A_i\}$ and $\{B_j\}$ are also shown. The 3D nature of the spectra is illustrated by showing $d \times d$ slices at $t = 50, 100, \dots, 400$. This is done for two different samples $k = 13, 23$. For each sample, three different versions are shown, the noiseless version $f^k(\xi, t)$, the observed version $X_\xi^k(t)$ and the optimal amplitude-corrected version $(A_iB_j)^{-1}X_\xi^k(t)$ where $\xi = (i, j)$. Note that the observed spectra are quite weak, due primarily to two factors: (1) Most amplitudes are small; (2) it is very likely that large amplitudes do not match the high signal locations. For example, where A_iB_j is high, $f^k(\xi, 100)$ could be small, the latter carrying the unique signal component. The optimal amplitude correction is unavailable in practice, since A_iB_j is unknown, and even then the result will be highly noisy. The noise level is taken to be $\sigma_X = 0.25$. Individual spectra depicted in the bottom right corner of Figure 3 illustrate the highly noisy nature of the observations.

Naive versus optimal averaging. Below, we will consider simulation results for some of the methods discussed in Section 2.2, in addition to what we call averaged lasso, `avg_lasso` for short, which is the usual lasso applied to averaged spectra. This averaging is done naively without taking into account the random amplitudes A_iB_j; namely, naive averaging computes $X^k_{\text{avg}}(t) = d^{-2}\sum_\xi X^k_\xi(t)$ and avg_lasso fits the regression model $y^k \sim \gamma_0 + \sum_t \gamma_t X^k_{\text{avg}}(t)$ using ℓ_1 penalty on (γ_t). Figure 4 compares $X^k_{\text{avg}}(t)$ with the result of optimal averaging, $d^{-2}\sum_\xi (A_iB_j)^{-1}X^k_\xi(t)$, where $\{A_i, B_j\}$ are the true attenuation factors that are not available. We note that there is a strong signal in optimally-averaged spectra, whereas in naive averaging much of the signal is lost. The tensor regression captures a stronger signal by adaptively estimating the weights and in effect fitting a regression $y^k \sim \gamma_0 + d^{-2}\sum_\xi(\hat{A}_i\hat{B}_j)^{-1}X^k_\xi(t)$, for estimated factors $\{\hat{A}_i, \hat{B}_j\}$.

Simulation results. Figure 5(a) and the first row of the table there shows the result of applying the methods discussed in Section 2.2 to the kernel representation of the spectra (as discussed in Section 2.1). For the kernel representation, we have used the scoring scheme of Section 2.3 and kept the top $N = 100$ spectral features (a 5-fold reduction from 500 wavenumbers). For our tensor regression model, we consider the variant given in (14), and we denote it as eref for "exact rank, exact fusion", since it enforce the rank, and the fusion of the γ parameter explicitly. More precisely, we set (γ_u) coefficients corresponding to neighboring spectral indices to be equal; indices u and v are neighbors if $G_{uv} \geq 0.95$. (The Gram matrix G is built based on $\mathbb{L}(t-s;8)$, matching what is used in the real-data applications in the sequel.) The effective dimension of (γ_u) turned out to be 7 for this exact fusion scheme (cf. Remark 3).

In addition, we consider the two tensor regularization approaches given in (16), which we have called grp_lasso_fro and grp_lasso_nuc. Also included in the comparison is the avg_lasso described earlier. To solve the convex optimization problems involved in these methods, we have relied on CVX, a package for specifying and solving convex programs [32, 33]. All these four methods have a single regularization parameter which we vary over a range. The plots in Figure 5 show the prediction Root-MSE (RMSE) as a function of this regularization parameter, where the error is computed by cross-validation over 20 batches each containing 10 prediction samples out of $n = 50$. As a baseline we have also included the RMSE of the mean model, which can be considered a regression with just the intercept present.

Figure 5(a), and the first row of the table, clearly show the advantage of eref over the other approaches, when deployed with the kernel representation features. Notably, because of the low effective dimension in eref, the error is not as sensitive to regularization parameter once it is sufficiently small, in contrast to the other approaches. In other words, the bulk of improvement comes from better rescaling of the spectra done by the tensor model (and not regularization of the wavenumber parameter γ). The relatively poor performance of grp_lasso_fro and grp_lasso_nuc can be attributed to the very low sample size we have here. This can be easily seen by the blow-up of their error as the regularization parameter goes to zero, showing that their effective parameter dimension is quite large compared to the sample size (At $\lambda = 0$, they are dealing



(a) Kernel representation

(b) $W = 2$

(c) $W = 8$

Denoising approach	`eref`	`avg_lasso`	`grp_lasso_fro`	`grp_lasso_nuc`
Kernel rep.	**0.0237**	0.1521	0.1997	0.1998
Naive smoothing (8)	0.1663	**0.1551**	0.1885	0.1857
Naive smoothing (2)	**0.1672**	0.1677	0.2003	0.2001

Fig 5. *Results on the synthetic data. Plots show the prediction Root-MSE (RMSE) for our proposed "exact rank, exact fusion" tensor lasso* (14)*, abbreviated* `eref`*, the averaged lasso* (`avg_lasso`)*, i.e., lasso on the naively-averaged spectra, and the two tensor regularizations of* (16)*, abbreviated* `grp_lasso_fro` *and* `grp_lasso_nuc`*., for the Frobenius and nuclear norm versions, respectively. The plots show RMSE as the single regularization parameter of each method is varied. The prediction errors are computed by cross-validation. Also shown is the mean model, i.e., regression with only the intercept present, which has RMSE of 0.2281. (Top-left) With kernel representation. (Top-right) With naive local smoothing of zeroth order, with window lengths = 2,8.*

with a parameter to sample-size ratio of $d_1 d_2 N/n = 10^2 \cdot 100/50 = 200$.) Nevertheless, we believe that they are very interesting approaches and should perform quite well for moderate sample sizes. It is also interesting to note that the two convex tensor regularizers behave almost identical, with a slight advantage for the Frobenius norm version at lower regularization values.

Effect of kernel representation. In order to study the effectiveness of the kernel representation of Section 2.1, as a combined denoising and dimension reduction step, we have also run the aforementioned methods on the spectra denoised by a more naive approach. Namely, we partition each waveform into blocks, or windows, of length W, and average the signal within each window. In effect, this approach is equivalent to running a zeroth order local (polynomial) smoothing, with window length W, followed by subsampling at rate $1/W$. Thus, the approach provides both denoising and dimension reduction, albeit somewhat naively. We consider two window sizes, the very short width $W = 2$ as a surrogate for no denoising, and $W = 8$ for moderate denoising and dimension reduction. (We have avoided $W = 1$ due to the large covariates it produces and the computational challenges of running on the resulting large parameter spaces.) We have used these denoised waveforms as input to the 4 approaches considered earlier.

Figure 3(b) and (c), as well the second and the third row of the table there, summarize the results. It is seen that in both cases ($W = 2, 8$) all methods perform no better than `avg_lasso`. The results are more or less similar for the two window lengths: `eref` does as well as `avg_lasso` (and in fact slightly better for $W = 2$), while `grp_lasso_fro` and `grp_lasso_nuc` perform slightly worse. Overall, however, no method achieves a substantial gain over the mean model with RMSE 0.2281, as opposed to the `eref` when applied to the kernel representation. Note also that `grp_lasso_fro` and `grp_lasso_nuc` behave almost identically over the whole regularization path, similar to what observed earlier; they also behave somewhat worse for $W = 2$ relative to $W = 8$, which is expected since the (ambient) dimension of the tensor parameter in this case is four times bigger. The conclusion is that the kernel representation paired with an appropriate (very) low-dimensional approach, such as `eref`, is indeed effective in driving down the prediction error, and both components are in fact necessary for this success.

3.2. In vivo Raman data

The models in this paper were motivated by Raman data collected from a bone fracture healing experiment. The study was conducted on 30 rats. Each rat underwent a surgical procedure to induce a small defect in one of its tibias, removing a thin slice of bone and fixing the bone to a metal plate so that it can heal back. The rats were then monitored for an eight-week period, at two-week intervals, starting from week 2. Six rats were sampled at all the four time points, while the rest were only sampled at a single time point (either at week 2, 4, 6 or 8) and then sacrificed to collect ex vivo Raman and micro-CT data. We have discarded week 2 data, due to equipment calibration issues that were resolved later. In total, $n = 37$ usable rat-week samples are available for the analysis.

In order to collect the Raman spectra, a ring-shaped apparatus was devised. The ring has $d_2 = 10$ holes around its circumference, where an illumination or a detection fiber can be inserted. The illumination (source) fiber emits laser light, and the detection fibers capture the resulting scattered light. At any given time, only one illumination fiber is used, and the remaining holes contain detection fibers. To obtain each Raman measurement, the ring was placed around the leg of the rat, aligned with the defect, and the light source was placed in five different positions resulting in measurements around the perimeter of the ring. The spectra were recorded for wavenumbers approximately in the range 954 cm^{-1} to 1700 cm^{-1}, for a total number $N = 544$ of wavenumbers.

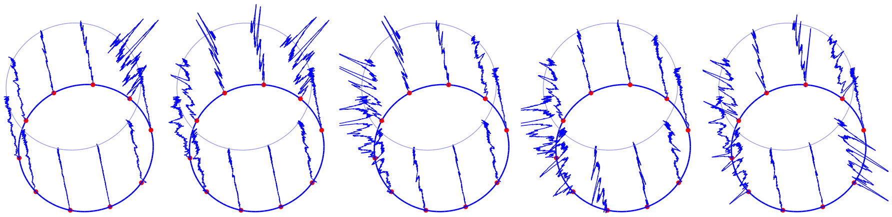


Fig 6. *Ring view of in vivo Raman tensor for each of the $p = 5$ source positions. Each cylinder represents all the spectra collected for a particular source. The dimension perpendicular to the page represents the wavenumber. The dots around the ring correspond roughly to detector positions (one of which is also a source position in each case.)*

Thus the data for each rat obtained on a single measurement occasion is a $544 \times 10 \times 5$ array, with the first dimension corresponding to wavenumbers, the second to the detector position, and the third to the source position. Figure 6 visualizes these data by placing each detector waveform at the corresponding position around the ring. Note that as the source rotates, the location of the highest amplitude detector rotates too – the closer the detector is to the source, the larger is the amplitude of its waveform. Figure 7 is another visualization of a single measurement, where the source and detector dimensions are stacked to obtain a 544×50 matrix. The four different plots in Figure 7 show the same measurement at different scales. In each plot, the spectra whose amplitude exceeds the scale are omitted from the plot.

Figure 7 clearly shows the need for normalization. Different source-detector combinations produce spectra of highly different amplitudes, which cannot be explained by the relative source/detector positions. After simple pre-processing normalization discussed above, the model learns a “proper” normalization from data by assigning a coefficient to each source and each detector. Figure 7 also shows the noisy nature of Raman data, which is especially evident at lower scales, suggesting that some form of denoising might be helpful. This will be handled by our functional representation, as discussed in Section 2.1

Our main goal for the Raman data from this experiment is to test its ability to predict well-established biomarkers of fracture healing which can be obtained by the more costly micro-CT approach. The mirco-CT data is obtained ex vivo, and it can measure various quantities more accurately, the primary measure of interest being bone mineral density (BMD), which is a good indicator of how the fracture is healing. We had access to an average BMD value within the region of interest, resulting in $n = 37$ scalar BMD measurements for the available rat-weeks.

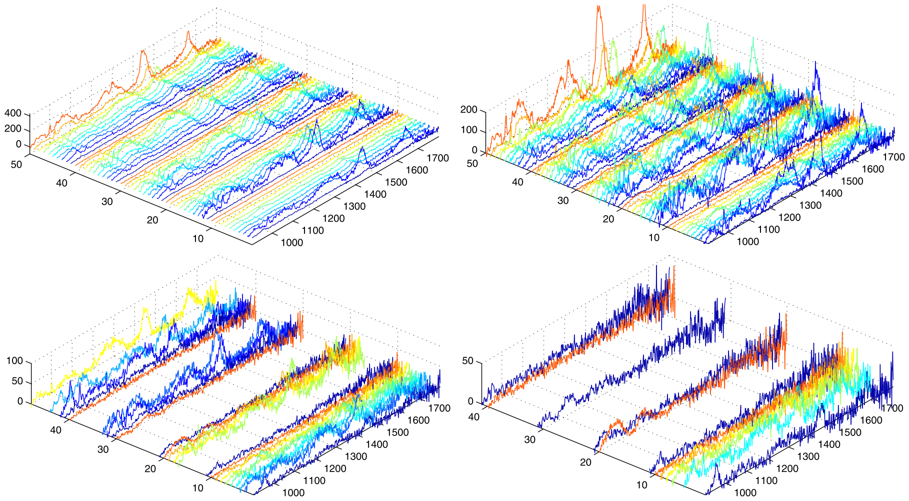


FIG 7. *Flattened view of the in vivo Raman tensor. The same data is shown on four different scales of y-axis. In each plot, the spectra with peak amplitudes exceeding the scale are not shown.*

Here, the models of Section 2 are applicable verbatim. We centered and normalized the response $\{y^k\}$ so that $\frac{1}{n}\sum_k y_k = 0$ and $\max_k |y_k| = 1$. The modified Raman tensor $\{\widetilde{x}^k_{iju}\}$ was normalized as discussed in Section 2.3, so as to have the maximum amplitude of 1 for every source-detector pair. Figure 8 shows some examples of predictive performance of the regression model (10). In each case, the sample is split randomly into a training set and a test set, the latter containing 2 rats from each of weeks 4, 6 and 8, corresponding to a "CV batch" with 26 training and 6 test rats. We have discarded 5 rats from the sample as outliers, based on their average prediction error across all the partitions, which were significantly higher than the rest. A total of 50 CV batches were considered. We employed two approaches in setting the regularization parameters: (1) *Adaptive regularization* where for each batch, we chose the regularization parameters to minimize the prediction error on the corresponding test set (the oracle choice). This gives the best performance we could hope that the model achieves in each case. (2) *Fixed regularization*, the more common approach, where a fixed set of parameters is used for all batches, and they are chosen so that the average error over all batches (i.e., the CV error) is minimized.

Figure 9 shows prediction errors measured by the median absolute deviation (MAD). As baseline, the x-axis on this plot shows the error of the *prediction by the mean*, that is, the error of the model with no covariates. The y-axis shows the training and test errors of our model, together with the test error of a simple approach traditionally used by spectroscopists, which we call *ratio*

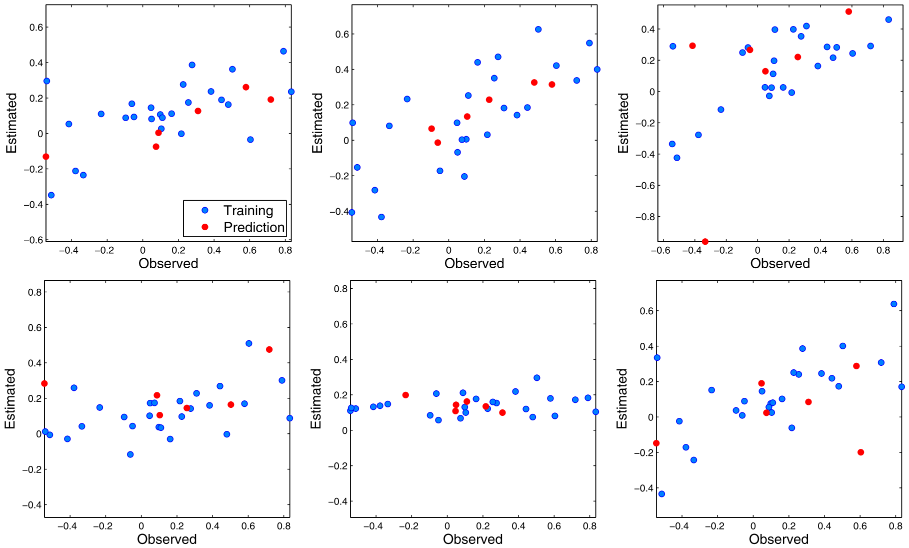


Fig 8. *Normalized BMD estimation. The plots show examples of estimated versus observed normalized BMD for the training (blue) and prediction (red) sets. The six plots correspond to six random splittings of the data into training and prediction sets.*

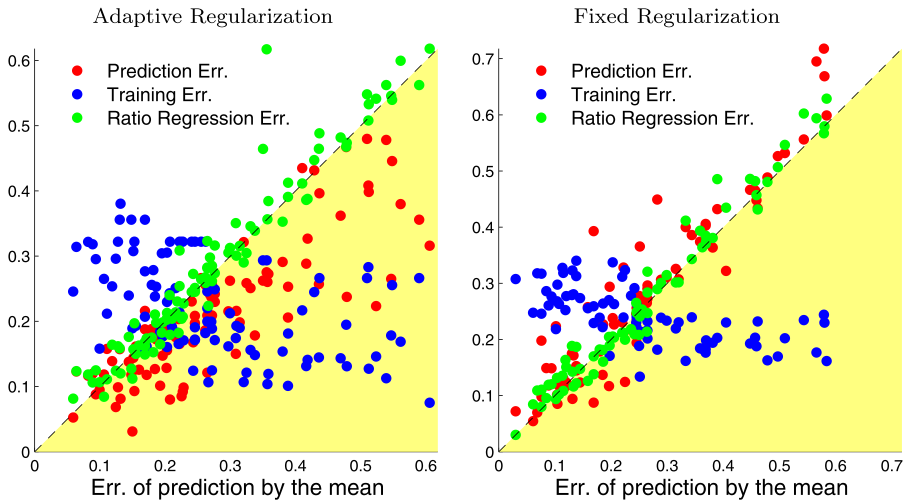


Fig 9. *Relative prediction performance. The plots show the mean absolute deviation error, in normalized BMD estimation, for the training (blue) and prediction (red) sets of our proposed regression model (10), versus the error of prediction by the mean. Also shown (green) is the error of the simpler ratio regression approach. The two panels correspond to the adaptive versus fixed choice of regularization parameters in regression model (10).*

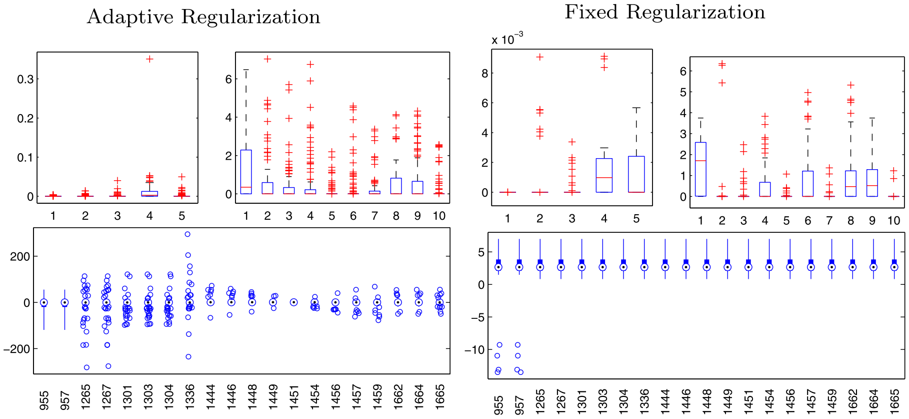
regression. The ratio regression approach predicts the BMD by regressing it on the ratio of spectra at two specific wavenumbers $954\,\mathrm{cm}^{-1}$ and $1450\,\mathrm{cm}^{-1}$, after averaging over all source-detector positions which are known to correspond to bone density. More precisely, adopting the notation in (8), ratio regression assumes the following model

$$y^k = \beta_0 + \beta_1 \frac{\widehat{X}^k_{\mathrm{avg}}(t_1)}{\widehat{X}^k_{\mathrm{avg}}(t_0)} + \varepsilon^k \tag{22}$$

where $\widehat{X}^k_{\mathrm{avg}}(t) = \frac{1}{d_1 d_2}\sum_{ij}\widehat{X}^k_{ij}(t)$, $t_1 = 954$ and $t_0 = 1450$. Note that prediction by the mean corresponds to assuming $\beta_1 = 0$. (In fact, since we also standardize $\{y^k\}$, $\beta_0 = 0$ is the optimal choice in prediction by the mean.)

Each point, on the plots in Figure 9, corresponds to a CV batch, with points below the diagonal corresponding to partitions of the data where the model has predictive value. With the adaptive regularization (left panel), this holds for almost all the batches, whereas the ratio regression result for this dataset is very similar to the baseline of predicting by the mean without using the spectral data at all. Note also the dependence on the CV batch, with some partitions of the data allowing for a good prediction and some not. With fixed regularization (right panel), all the models are similar to the baseline.

Figure 10 shows the estimated coefficients of the model. The three sets of coefficients are the 5-vector of source weights (α), the 10-vector of detector weights (β) and the 40-vector of wavenumber weights (γ). The high number of outliers in source/detector weights and the general tendency of the wavenumber weights to fluctuate around zero are in alignment with poor prediction performance.


Adaptive Regularization

Fixed Regularization

FIG 10. *Estimated model coefficients for the in vivo dataset, for both adaptive and fixed regularization. Boxplots of source weights (5-vector), detector weights (10-vector) and wavenumber weights (40-vector) are illustrated.*

We draw the following conclusions from these results: Predicting BMD (or similar measures) from in vivo Raman is inherently difficult (and to the best of our knowledge, this experiment was the first to attempt this to monitor fracture healing). This may be due to any number of reasons – the mixing of bone and tissue signals, variability in the rats and the small sample size, and the quality of the Raman data from this experiment. The success of adaptive regularization in predicting most batches (to varying degrees) and the failure of fixed regularization suggest that the samples are not homogeneous; if this is due to inherent population variability, a larger sample size would improve prediction. This may also be due to data collection problems for selected experiments, due to the sometimes inaccurate placement of the complicated ring setup, and our discussion with the chemists who conducted the experiment suggests this is the likely explanation.

3.3. Ex vivo Raman data

The ex vivo Raman dataset was collected from the same experiment described in Section 3.2, obtained directly from the bone without tissue interference after the animals were sacrificed. The data are Raman spectral maps of cross-sections of bone. The specific locations at which the Raman spectra were collected differed by rat, based on bone morphology, and the number of measurement locations varied from 3 to 7, with up to 3 sub-locations for each. Thus we averaged the spectra within rats to obtain a single average spectrum per rat. Since the BMD is also an average value, this can be viewed as a prediction of spatially averaged BMD from spatially averaged Raman spectra. In total, there were 23 rats measured at 806 wavenumbers.

For ex vivo data, there are no spatial dynamics to be modeled. That is, $d_1 = d_2 = 1$, and the model becomes a functional regression with Lasso and fused Lasso penalties. In this case, we only have one set of parameters to estimate, namely, $\{\gamma_u\}$ or equivalently, the functional weights $w(t)$ of (9). More precisely, the model in (8) is reduced to $y^k = \int_T \, w(t) \, \widehat{X}^k_{11}(t) \, dt + \varepsilon^k$ and the optimization problem in (13) to

$$\widehat{\gamma} = \underset{\gamma \in \mathbb{R}^N}{\operatorname{argmin}} \; \frac{1}{2n} \sum_{k=1}^{n} \Big(y^k - \sum_u \gamma_u \, \widetilde{x}^k_{11u} \Big)^2 + \rho_\gamma \, \|\gamma\|_1 + f_G(\gamma),$$

where f_G is the weighted fused Lasso penalty in (11).

Figure 11 (left panel) shows cross-validated errors for our model (prediction and training) and prediction errors for the baseline (prediction by the mean) and ratio regression, described in Section 3.2. Here, a fixed set of the tuning parameters of our model was chosen by CV in the usual fashion, so that average CV error is minimized. The partitioning of data into prediction and training sets, and the number of batches are as in Section 3.2. Our model provides a noticeable improvement over the simple ratio regression which is only slightly better than the baseline.

Also shown in Figure 11 (right panel) are the boxplots of the estimated coefficients $\{\gamma_u\}$. In contrast to the in vivo case, one can clearly identify regions

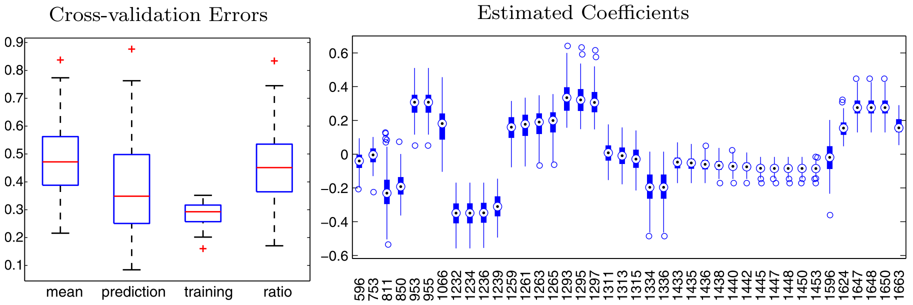


FIG 11. *The results for ex vivo dataset. Left panel shows, from left to right: prediction error for the mean (baseline), prediction error for our model, training error for our model, and prediction error for ratio regression. All errors are computed from cross-validation. Right panel shows the estimated model coefficients.*

of wavenumbers that consistently exhibit large coefficients and hence are helpful in predicting the BMD. For example, the mineral band at ≈ 954, which is known to be excited by calcium in the bone, is prominent in the plot. Overall, these results suggest that the main challenges in using Raman spectra to predict BMD come from the attempt to do it in vivo, rather than some inherent problem with the bone spectra themselves.

3.4. NMR data

To further explore the effectiveness of our approach, we also considered a dataset of 2D diffusion-edited H NMR spectra. This dataset is extensively studied in [34]. It is also used by [31] to illustrate their matrix regression approach, discussed in Remark 1 in Section 2.2. The dataset contains NMR spectra and lipoprotein concentrations for 25 human subjects. The concentrations of cholesterol and triglyceride were obtained by ultracentrifugation, for various fractions and subfractions in terms of lipoprotein density. The primary fractions of interest are very low, low, intermediate, and high density lipoproteins, abbreviated as VLDL, LDL, IDL and HDL. A total of 32 concentration levels are reported, from which we have used the variable `'CH_V2'`, following [31]. For this variable, the concentrations were missing for 5 subjects; hence, we take the response vector to be the $n = 20$ (scalar) recorded concentration levels, and discard the spectra corresponding to missing responses.

The NMR spectra are measured as intensity for each chemical shift (in the range 2.5–0.6ppm); see [34] for details. For each subject, a 2-D spectrum of dimensions 24×1600 is produced, where the second dimension is the spectral one. The first dimension corresponds to 24 steps of gradient pulse strength. Figure 12 illustrates a typical example. Note that the spectral range is indexed sequentially from 1, as the dataset does not include the exact chemical shift value. This dataset has both a spatial and a spectral dimension, making it a

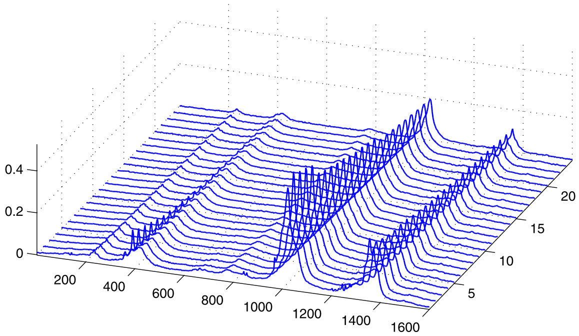


Fig 12. *A typical example of a NMR spectrum. The spectral dimension runs to* 1600.

good fit for our modeling approach. Since the spectra are smooth along the spectral dimension, we have sub-sampled them at a rate 1/3 as a preprocessing step to reduce the dimension of each observation to 534 × 24.

Setting $d_2 = 1$ in the regression model introduced in Section 2.2, since here we have only one spatial dimension, results in two sets of parameters to be estimated, $\{\alpha_i\}$ and $\{\gamma_u\}$. More precisely, the model in (8) is reduced to $y^k = \sum_i \int_T \alpha_i w(t)\, \widehat{X}^k_{i1}(t)\, dt + \varepsilon^k$ and the optimization problem in (13) to

$$(\widehat{\alpha}, \widehat{\gamma}) = \underset{\substack{\alpha \in \mathbb{D}^{d_1}, \\ \gamma \in \mathbb{R}^N}}{\operatorname{argmin}} \; \frac{1}{2n} \sum_{k=1}^{n} \Big(y^k - \sum_{iu} \alpha_i \gamma_u \, \widetilde{x}^k_{i1u} \Big)^2 + \rho_\gamma \, \|\gamma\|_1 + f_G(\gamma)$$

where f_G is the weighted fused Lasso penalty in (11). As with the Raman datasets, we retain the $N = 40$ highest scoring chemical shifts in the model. Figure 13 shows errors for the NMR datasets, computed over a total of 50 CV batches, where in each batch we left 4 samples out for prediction (out of the total of $n = 20$). The left panel in Figure 13 shows CV errors for the baseline (prediction by the mean) and for our model (both prediction and training errors). Here we observe a substantial improvement by our model over the baseline. The left panel shows the boxplots for estimated coefficients α (on the top) and γ (on the bottom). These clearly show some spatial and spectral positions to predict the response. It further illustrates that with a sufficiently high signal-to-noise ratio, our models can be used for prediction and variable selection.

4. Discussion

We presented a functional model for spectra which can be used for denoising and compression as well as in downstream prediction. Based on this representation of the data, we proposed a regression model to predict a scalar response (e.g., bone mineral density or lipoprotein concentration) based on multi-dimensional spectroscopy data. The data was modeled as a tensor with several spatial dimensions and one spectral dimension. A rank-one multilinear map was proposed

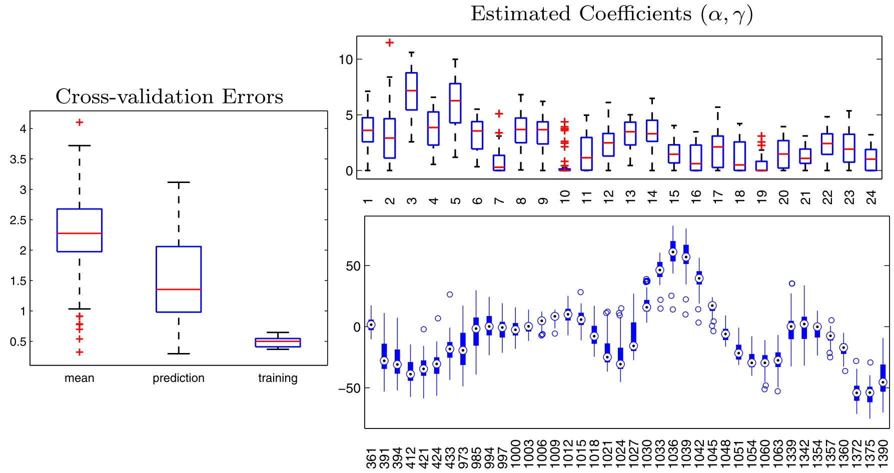


Fig 13. *The results for NMR dataset. Left panel shows, from left to right, CV error for prediction by the mean, prediction error for our model, and the training error for our model. Right panel shows the estimated model coefficients.*

to describe the relationship between the spectra and the response, with sparsity imposed on the spectral domain coefficients, and smoothing based on a similarity measure. The structure was enforced using regularization, with ℓ_1 penalty to induce sparsity, and a fused Lasso type penalty to enforce similarity.

We considered the effectiveness of the approach in three settings: in vivo and ex vivo Raman data from a fracture healing experiment, and using NMR data in a lipoprotein concentration study. For the in vivo Raman experiments, our results were mixed and led us to conclude that the data were not sufficiently homogenous to allow for good global prediction, although for some partitions of the data into training and test sets good prediction was possible. There are many possible explanations for that, from inhomogeneity among the rats to insufficient sample size to experimental errors in device placement, with the latter considered the most likely explanation by our chemistry collaborators. For the ex vivo Raman data and the NMR data, we showed that the model has good global predictive power.

Some of the lessons learned from a modeling perspective are as follows: One can use a kernel approach with nonnegative weights to achieve sparse representation of spectra. In addition to effective denoising and dimension reduction, the weights obtained can be directly mapped back to the spectral domain to facilitate interpretation. We have also observed while testing various models that whenever the choice is between simplicity and complexity, the simpler model is more effective when dealing with very noisy high-dimensional covariates. More precisely, simple parameter-counting and comparison with the sample size provides a good guide in deciding the complexity of the models we could hope to fit. This, for example, has been the main reason behind restricting to rank-one models in this paper, though our approach can be easily extended to accommodate rank $r > 1$ (which we did try during our experiments with in vivo Raman data). We believe it is better to try to pick the dominant mode of variation first ($r = 1$) and if successful try higher rank models. Finally, it is possible to use regression models as compelling evidence for the lack of a relation between the response and covariates, giving the model the alternate role of a diagnostic tool. In other words, the lack of predictive power is not always a defect of the model; rather, when applied with enough scrutiny, it could serve as revealing possible lack of the signal in the data. It is better to err on the side of caution by using simple models rather than fitting a complex model and picking up relations that do not necessarily exist.

In this paper we mainly focused on regression models. An interesting possibility for future work is to look at PCA type analyses, by which we mean models taking into account a decomposition of the spectra into relevant (and irrelevant) components. A promising direction is to use the response variable to guide the selection of components, along the lines of supervised PCA [35]. For the chemical interpretation to be meaningful, the components need to satisfy the physical constraints imposed on spectra, such as positivity. This makes the problem different and more challenging than classical PCA or PCA-based regression. A challenge in this case is to find alternatives to orthogonality constraints which are meaningful for positive components and still allow for efficient optimization.

Acknowledgements

This research was supported by NIH grant 5-R01-AR-056646-03. We are grateful to Prof. Michael Morris (Department of Chemistry, University of Michigan) and members of his lab, especially Dr. Paul Okagbare and Dr. Francis Esmonde-White, as well as Prof. Steven Goldstein (Orthopedic Surgery, University of Michigan) and members of his lab for collaborating on the grant and collecting and sharing the data, as well as for many useful discussions.

References

[1] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2005. MR2168993

[2] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis*. Springer Series in Statistics. Springer New York, 2006. MR2229687

[3] R.A. Moyeed and P.J. Diggle. Rates of convergence in semi-parametric modelling of longitudinal data. *Australian Journal of Statistics*, 36(1):75–93, mar 1994. MR1309507

[4] S. L. Zeger and P. J. Diggle. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50(3):689–699, 1994.

[5] K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. MR0836430

[6] J. J. Faraway. Regression analysis for a functional response. *Technometrics*, 39(3), 1997. MR1462586

[7] D. R. Hoover, J. A. Rice, C. O. Wu, and L. P. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, pages 809–822, 1998. MR1666699

[8] C. Wu, C. T. Chiang, and D. R. Hoover. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association*, 93(444):1388–1402, 1998. MR1666635

[9] J. Fan and J. T. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society. Series B*, 62(2):303–322, 2000. MR1749541

[10] D. Y. Lin and Z. Ying. Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 96(453):103–113, 2001. MR1952726

[11] G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society. Series B*, (2):1–18, 2001. MR1858401

[12] G. M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B*, (2), 2002. MR1924298

[13] F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1-2):161–173, oct 2003. MR2020144

[14] H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *The Annals of Statistics*, 33(2):774–805, apr 2005. MR2163159

[15] G. M. James, J. Wang, and J. Zhu. Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108, oct 2009. MR2543686

[16] Philip T. Reiss, Jeff Goldsmith, Han Lin Shang, and R. Todd Ogden. Methods for Scalar-on-Function Regression. *International Statistical Review*, pages 1–22, 2016.

[17] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, Boston, MA, 2004. MR2239907

[18] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, oct 2011. MR2906869

[19] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, apr 2011. MR2816348

[20] Garvesh Raskutti and Ming Yuan. Convex Regularization for High-Dimensional Tensor Regression. page 55, 2015.

[21] Peter D. Hoff. Multilinear tensor regression for longitudinal relational data. *Annals of Applied Statistics*, 9(3):1169–1193, 2015. MR3418719

[22] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor Regression with Applications in Neuroimaging Data Analysis. *Journal of the American Statistical Association*, 108(502):540–552, jun 2013. MR3174640

[23] Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal*

Statistical Society: Series B (Statistical Methodology), 76(2):463–483, mar 2014. MR3164874

[24] Yue Hu and Genevera I Allen. Local-aggregate modeling for big data via distributed optimization: Applications to neuroimaging. *Biometrics*, 71(4):905–917, dec 2015. MR3436716

[25] I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools using adaptive splines. Technical report, 1991.

[26] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996. MR1379242

[27] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. *Computational learning theory*, pages 416–426, 2001. MR2042050

[28] R. J. Meier. On art and science in curve-fitting vibrational spectra. *Vibrational Spectroscopy*, 39(2):266–269, oct 2005.

[29] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, feb 2005. MR2136641

[30] X. Chen, Q. Lin, and S. Kim. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, jun 2012. MR2976489

[31] Junlong Zhao and Chenlei Leng. Structured lasso for regression with matrix covariates. *Statistica Sinica*, To appear, 2014. MR3235399

[32] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[33] Michael Grant and Stephen Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`. MR2409077

[34] M. Dyrby, M. Petersen, A. K. Whittaker, L. Lambert, L. Nörgaard, R. Bro, and S. B. Engelsen. Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics. *Analytica Chimica Acta*, 531(2):209–216, feb 2005.

[35] E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, 101(473):119–137, mar 2006. MR2252436