Chemometrics and Intelligent Laboratory Systems 109 (2011) 120–130

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Multidimensional single-index signal regression

Brian D. Marx [a,*], Paul H.C. Eilers [b], Bin Li [a]

[a] Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803, United States
[b] Department of Biostatistics, Erasmus Medical Centre, 3015 GE, Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 14 December 2010
Received in revised form 29 July 2011
Accepted 3 August 2011
Available online 25 August 2011

Keywords:
Multivariate calibration
P-splines
Signal regression
Single-index
Spectra
Tensor product
Ternary mixtures

ABSTRACT

In general, linearity is assumed to hold in multivariate calibration (MVC), but this may not be true. We approach the MVC problem using multidimensional penalized signal regression, which can be extended with an explicit link function between linear prediction and response and in the spirit of single-index models. As the two-dimensional surface of calibration coefficients is smoothly and generally estimated with tensor product P-splines, the unknown link function is estimated using univariate Psplines. The methods presented are grounded in penalized regression, where difference penalties are placed on the rows and columns of the tensor product coefficients, as well as on the link function coefficients, each having its own tuning parameter. An application to ternary mixture data shows that a non-linearity is present. Performance comparisons are made to standard penalized signal regression, not only demonstrating the nonlinear effect, but also improvements in external prediction.

1. Introduction

In this paper, we take a novel approach in the multivariate calibration problem, in particular where the signal (spectra) regressors have two-dimensional structures. Our application considers UV–VIS spectra taken over several temperatures. Through simultaneous estimation, we parse out and estimate two separate modeling components: (1) a single *smooth regression coefficient surface* associated with the two-dimensional signal [12], and (2) an unknown, possibly nonlinear, *link function* [3]. Although the first component is linear, the second component explicitly models the nonlinearity, allowing us to learn something about its features, while enhancing insight into the measurement process. We will see that the combination of these components can lead to a systematic and tractable modeling approach, that is statistical in nature, while having improved external prediction performance when compared to standard signal regression approaches and partial least squares.

1.1. Multivariate calibration with two-dimensional signals

At the heart of a multivariate calibration problem is rich regressor data, often compactly given as a digitized signal, curve, or spectra. Such regressor information can also be in two or more dimensions, of such digitized images. An often ironic consequence of such data is that as more and more precisely regressor information are obtained, the more and more ill-conditioned estimation becomes. Since classical least squares modeling approaches usually fail, there have been numerous competing methods developed to provide tractable and reliable prediction; see Eilers et al. [3] and Eriksson et al. [7] for partial lists. We will see that, unlike most of the other approaches, our proposed method additionally takes advantage of the ordered or array structure among the regressors.

To motivate the problem, Fig. 1 displays signal regressors (at two different temperatures) for each of $m = 34$ observations, coming from a ternary mixture experiment using spectroscopy. Each "signal" actually consists of numerous digitizations ($p = 401$) along the wavelength axis (700 to 1100, by 1 nm). The top (bottom) panels present the raw (first differenced) spectra. If such optical regressors are to be related, e.g. to a chemometric response, then some regularization is needed. Generally, not only is $p \gg m$, but the regressors are highly correlated.

Notice that the left and right panels of Fig. 1 presents "signals" at *temperature* levels of 30° and 70 °C, respectively, and one could imagine even more, forming a sequence of several "extremely narrow images". Thus a natural question to ask is: What if the signal regressors become fully two-dimensional, and we wish to take into account spatial information in both directions? One could view this problem as multivariate calibration with multidimensional spectra, where, e.g., the second dimension is temperature. Fig. 2 presents such a two-dimensional spectra structure with 4800 regressors, summarized in a 12×400 matrix (using first differences), for the center mixture unit, with corresponding scalar responses (water, 1,2-ethanediol, 3-amino-1-propanol, each at 0.33).

* Corresponding author.
E-mail addresses: bmarx@lsu.edu (B.D. Marx), p.eilers@erasmusmc.nl (P.H.C. Eilers), bli@lsu.edu (B. Li).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.08.006

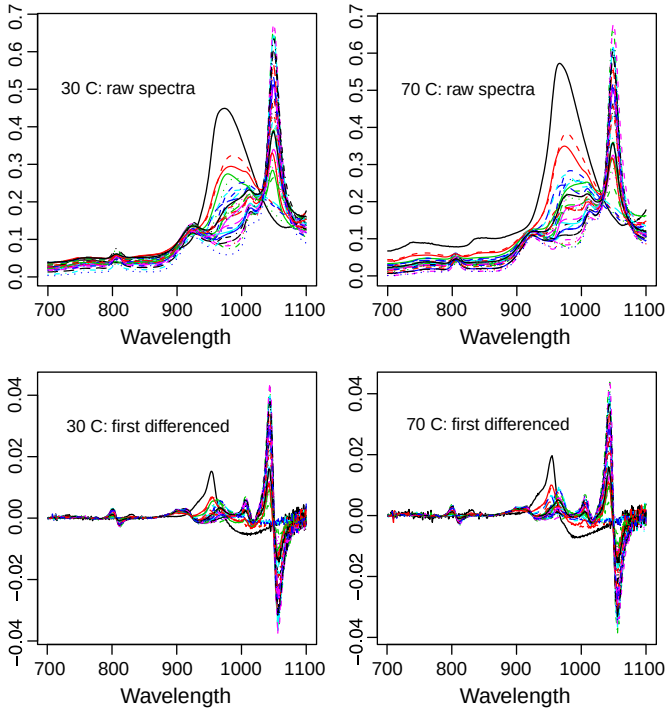

Fig. 1. Signal regressors (raw and first differenced) for mixture experiment, at two different temperatures.

1.2. Notation and data structure

The data structure is as follows, each observation consists of the data pair: (y_i, X_i), where $i = 1,\ldots,m$. The response y_i is scalar. We assume independence among the responses, with common variance $\text{var}(y) = \sigma^2$.

The two-dimensional signal consists of (often thousands of) digitized regressors, X_i, arranged in a $p \times \breve{p}$ array. The indexing axes, i.e. v and $\breve{v}$, that define the support coordinates of X_i are usually on a regular grid, but the only requirement for our method is that the scatter of digitizations are common for all i. As suggested by Fig. 2, the number of regressors are rich, over one hundred times greater in number than observations. The regressor support is specified as $v^\star$ (wavelength) with $p = 400$ channels (701 to 1100 nm, by 1 nm) and $\breve{v}^\star$ with $\breve{p} = 12$ temperature channels (30, 35, 37.5, 40, 45, 47.5, 50, 55, 60, 62.5, 65, 70 °C).

The response y comes from the composition (mole fraction) of a mixture, here consisting of three components (water, 1,2-ethanediol, 3-amino-1-propanol). The ternary plot for the $m = 34$ mixtures is provided in Fig. 3. The center data point in the triangle represents equal concentrations of the three components, the edge points are mixtures containing only two components, and the corners are pure. Note that there are 3 pure, 12 edge, and 19 interior (1 center) mixtures.

1.3. First modeling component: MPSR

The multidimensional signal regression (MPSR) model was first presented in Marx and Eilers [12], initially motivated by both Marx and Eilers [10] and Eilers and Marx [6]. The model's goal is to provide an extremely practical solution for functional linear models using the entire two-dimensional signal as regressors. Associated with the regressors is a single overarching coefficient surface which serves to smoothly weigh each two-dimensional signal digitization over its support. Regularization is needed, and we choose to impose some sensible constraints: ones that take into account the spatial structure of the regressors, while ensuring smoothness in the coefficient surface. As with any P-spline approach [4], we take two steps toward smoothness: (a) The coefficient surface (not the signal) is intentionally overfit using two-dimensional tensor product B-splines, making the surface more flexible than needed. (b) Tensor product coefficient estimates are penalized using difference penalties on each of the rows and columns.

The first step provides an initial reduction in parameter estimation through smoothness, as the higher dimensional two-dimensional signal coefficient surface is projected onto a lower dimensional tensor product basis, where the knots are “richly” chosen on a regular grid, thus circumventing knot selection schemes. The second step ensures further smoothness, as well as regularization, while allowing general surface candidates. Two tuning parameters, associated with the row and column penalties, respectively, are needed to allow for continuous control over the surface. Fig. 4 displays a variety examples of (coefficient) surfaces using tensor products B-splines. The upper, left

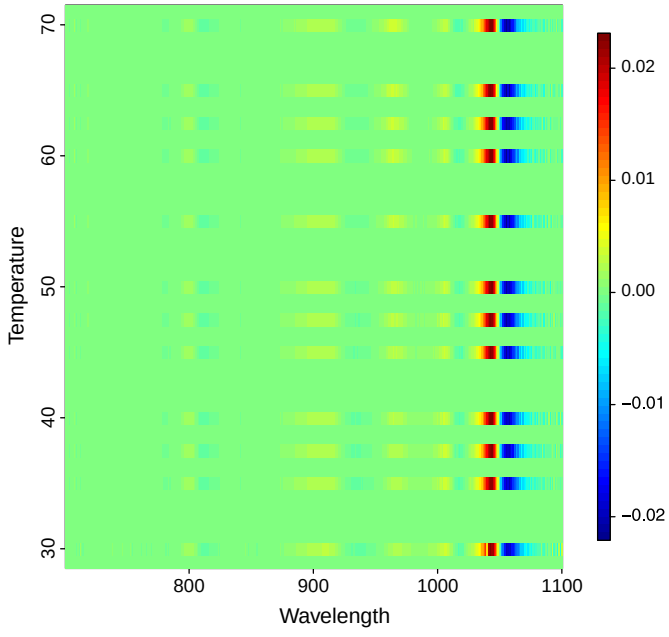

Fig. 2. Two-dimensional (first differenced) signal regressor image for center mixture.

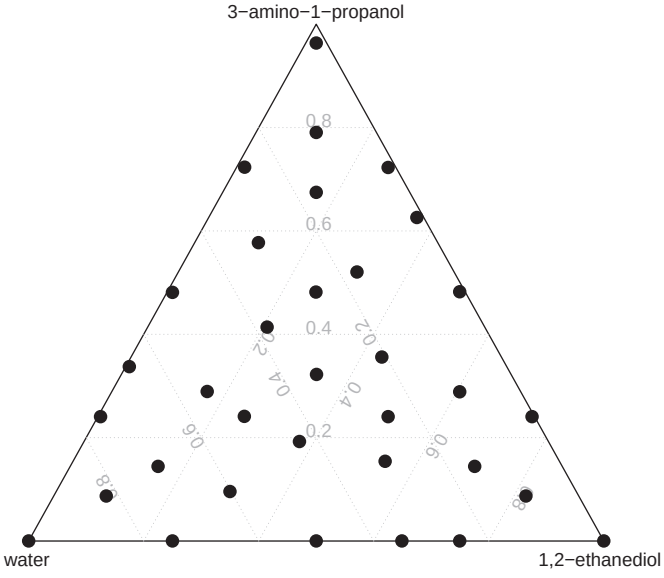

Fig. 3. Ternary plot for mixtures, with $m = 34$: 3 pure, 12 edge, 19 interior.