Chemometrics and Intelligent Laboratory Systems 143 (2015) 111–121

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Varying-coefficient single-index signal regression

Brian D. Marx

Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803, United States

ARTICLE INFO

Article history:
Received 26 September 2014
Accepted 6 February 2015
Available online 28 February 2015

Keywords:
Multivariate calibration
Tensor P-spline
Signal regression
Single-index
Varying-coefficient models
Unknown link function

ABSTRACT

The penalized signal regression (PSR) approach to multivariate calibration (MVC) assumes a smooth vector of coefficients for weighting a signal or spectrum to predict the unknown concentration of a chemical component. P-splines (i.e. B-splines and roughness penalties, based on differences) are used to estimate the coefficients. In this paper we allow the PSR coefficient vector to vary smoothly along a covariate (e.g. temperature), which results in a smooth surface on the wavelength–temperature domain. Estimation is performed using two-dimensional tensor product P-splines. As such, a slice of this surface effectively estimates the vector of coefficients at any arbitrary temperature. As an added generalization, we further relax the implicit assumption of an identity link function by allowing an unknown, but explicit, link function between the linear predictor and the response. Again, we allow the signal's link function to vary smoothly along a covariate, which produces a two-dimensional *link surface*. The unknown link surface is also estimated using two-dimensional P-splines, which is sliced at the same arbitrary temperature to bend prediction. Typically we use a common covariate (e.g. temperature) to vary the associated link function, as with the signal coefficients, but nothing prohibits the use of two different ones. We term our method: varying single-index signal regression (VSISR). The methods presented are grounded in penalized regression, where difference penalties are placed on the rows and columns of the tensor product coefficients. Each row and column of each surface has its own tuning parameter. An application to ternary mixture data illustrates that both the varying-coefficient and varying-nonlinearity (due to the link) are present. External prediction performance comparisons are made for both the identity link varying-coefficient penalized signal regression (VPSR) and partial least squares (PLS).

1. Introduction

In this paper, we take yet another approach to the multivariate calibration problem, in particular where the signal (spectra) regressors appear to have two-dimensional structure. Although we generally use the term *signal* throughout the paper, our application considers NIR spectra (taken over several temperatures). Through simultaneous estimation, we identify and estimate two separate modeling components, both of which are surfaces: (a) a single *smooth regression coefficient vector*, which effectively ensembles a smooth surface while varying along the temperature covariate [1], and (b) an unknown and nonlinear link function, which also varies along the temperature covariate, yielding a *link surface* and thus extends the work of Eilers, Li and Marx [2] and [3]. Although the first component is linear, the second component explicitly models the nonlinearity, allowing us to learn something about features of the transformed mean, which in some cases enhances insight into the process. We choose to use a common covariate (e.g. temperature) to vary the associated link function, as with the signal coefficients, but the interacting covariate could differ. We will see that the combination of these components can lead to a systematic and tractable modeling approach, that is statistical in nature, while in some cases having improved external prediction performance when compared to identity link model variants and partial least squares.

2. Motivating example

We revisit data used in [3], with permission from Zhenyu Wang and Age Smilde, where 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). These data are an expanded version of the data used in [4,5], and [6]. The ternary plot for the $m = 34$ mixtures is provided in Fig. 1. 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. The components are modeled one at a time, and not jointly.

Corresponding to each ternary mixture, there exists an extremely rich spectroscopy regressor information, taken under $\breve{p} = 12$ temperature conditions: (30, 35, 37.5, 40, 45, 47.5, 50, 55, 60, 62.5, 65, 70 °C). Fig. 2 displays signal regressors (at only two different temperatures) for each of $m = 34$ observations. Each "signal" actually consists of

E-mail address: bmarx@lsu.edu.

http://dx.doi.org/10.1016/j.chemolab.2015.02.005
0169-7439/

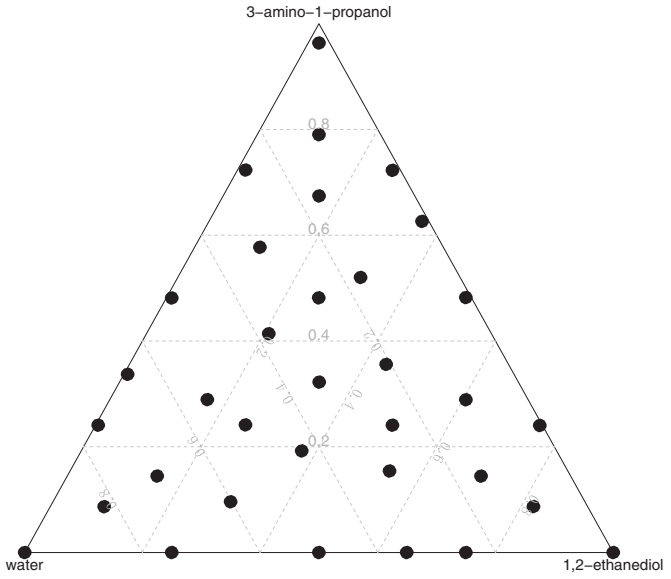

Fig. 1. Ternary plot for mixtures, with $m = 34$: 3 pure, 12 edge, 19 interior.

numerous digitizations ($p = 401$) along the wavelength axis v (700 to 1100, equally-spaced by 1 nm). The top (bottom) panels present the raw (first differenced) spectra. The latter will be our choice, which is in part attractive since constant shifts across spectra are removed.

Notice that the left and right panels of Fig. 2 present signals at the extreme temperature levels of 30° and 70 °C, respectively. One could imagine many more given (or interpolated) temperatures, resulting in a sequence of several "extremely narrow images" to build out a two-dimensional regressor surface.

2.1. Motivation for this paper

Thus a natural question to ask is: what is the true, or more importantly, the most useful regressor structure to predict y?

The primary goal is reliable future (external) prediction. The data set brings some unique structure and several challenges: (a) for all practical purposes, the response is measured *exactly* at the molar level, and only at several dozen concentrations. (b) The rich covariate information has dimension far greater (at least an order of magnitude greater) than the number of observations. (c) Internal prediction is *not* of interest, as it could be perfectly done, if desired, in infinitely many ways. (d) Oddly, it is the signal regressors themselves, and *not* the responses, that change with changes in the covariate t.

The data structure considered by Marx and Eilers [7] and Marx, Eilers, and Li [3] is rethought, where in the latter each of the $m = 34$ mixtures had one image regressor (400×12). As such, the composite of signal regressors was then viewed as fully two-dimensional, where spatial information was taken into account in both (the wavelength and temperature) directions, and this information was related to the response (component concentration). The problem was viewed in the light of a multivariate calibration with multi-dimensional spectra, where, e.g., the second dimension was temperature. Fig. 3 illustrates such a two-dimensional spectra structure with 4800 regressors, summarized in a 400×12 matrix (using first differences), for the center mixture unit, with corresponding scalar responses (water, 1,2-

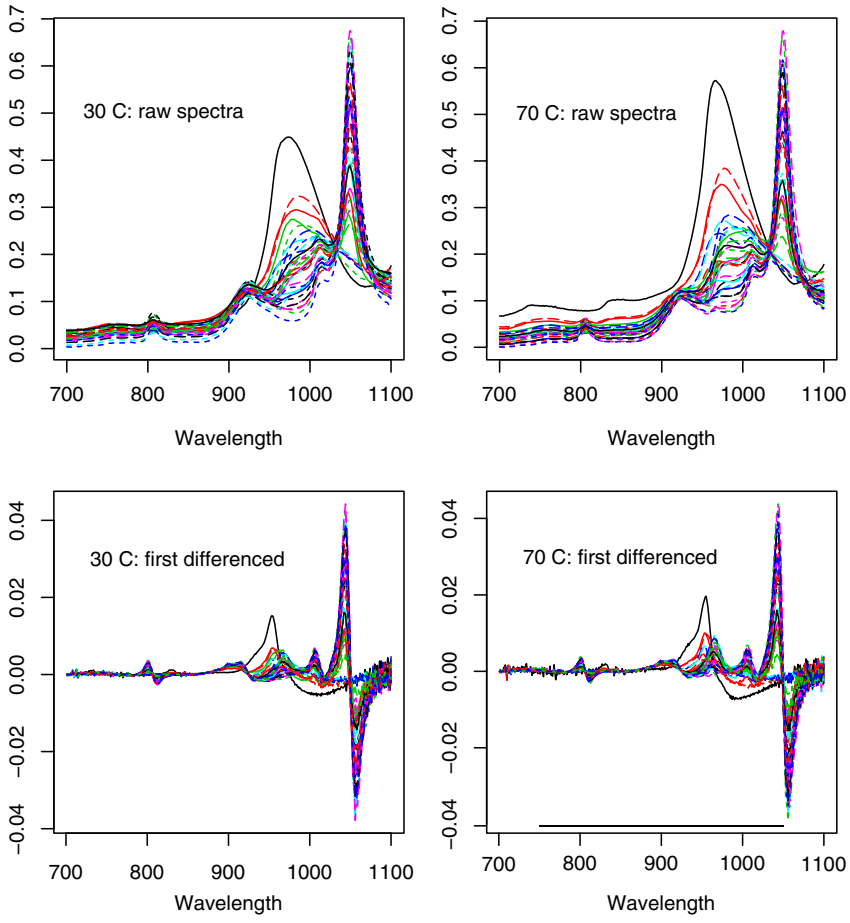

Fig. 2. Signal regressors (raw and first differenced) for mixture experiment, at two different temperatures.