Multidimensional Single-Index Signal Regression

Brian D. Marx[1], Paul H. C. Eilers[2], Bin Li[3]

[1] Dept of Experimental Statistics, Louisiana State University, Baton Rouge, USA
[2] Dept of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands
[3] Dept of Experimental Statistics, Louisiana State University, Baton Rouge, USA

Abstract:
We take a novel approach the signal regression (multivariate calibration) problem, in particular where the signal (spectra) regressors have two dimensional structure. In general linearity is assumed to hold, but this may not be true. 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, and (2) an unknown, possibly nonlinear, *link function*. Using (tensor product) P-splines for each component, we will see that their combination can lead to a systematic and tractable statistical modeling approach, while having improved external prediction performance when compared to standard signal regression approaches and partial least squares. Optimal tuning will be discussed.

Keywords: Multivariate calibration; P-splines; spectra.

1 Introduction

Our application considers rich and indexed two-dimensional regressor information of UV-VIS spectra taken over several temperatures that are used to predict scalar components of a ternary mixture. We will see that the basic appeal of our particular modelling approach is its explicit estimation of meaningful components: (1) a *smooth regression coefficient surface* associated with the two-dimensional signal (Marx and Eilers, 2005), and (2) an unknown, possibly nonlinear, *link function*. Although the first is linear, the second component explicitly models the nonlinearity, while enhancing insight into the measurement process. Linking the response to the linear predictor is in the spirit of single-index models (Eilers, Li and Marx, 2009).

1.1 First modeling component MPSR

The multidimensional signal regression's (MPSR) goal is to provide a 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. Regularization is needed, and we choose a P-spline approach. Specifically, we take two steps towards 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. Given the ith regressor matrix $X_i = [x_{ijk}]$ of dimension $p \times \breve{p}$, signal regressor support on $(v, \breve{v})$, and coefficient surface $\alpha(v, \breve{v})$, express the mean

$$\mu_i = \sum_{j=1}^{p} \sum_{k=1}^{\breve{p}} x_{ijk}\alpha_{jk} = \sum_{j=1}^{p} \sum_{k=1}^{\breve{p}} x_{ijk} \sum_{r=1}^{n} \sum_{s=1}^{\breve{n}} B_{rj}\breve{B}_{sk}\gamma_{rs} = \mathbf{x}_i'\mathbf{T}^{\star}\gamma, \qquad (1)$$

where $i = 1, \ldots, m$; $j = 1, \ldots, p$; $k = 1, \ldots, \breve{p}$, with tensor product B-splines $\mathbf{T}^{\star}$, where $\mathbf{x}_i' = \text{vec}(X_i)$. We can further express (1) in matrix form as $\mu = \mathbf{X}\mathbf{T}^{\star}\gamma = \mathbf{M}\gamma$, where $\mathbf{X}$ is the $m \times p\breve{p}$ matrix of vectorized signals, $\mathbf{M} = \mathbf{X}\mathbf{T}^{\star}$.

In the P-spline spirit, the objective function is to minimize

$$\begin{aligned} Q_P(\gamma) &= \sum_{i=1}^{m}(y_i - \mathbf{x}_i'\mathbf{T}^{\star}\gamma)^2 + \lambda \sum_{r=1}^{n} \gamma_{r\bullet} D_d' D_d \gamma_{r\bullet}' + \breve{\lambda} \sum_{s=1}^{\breve{n}} \gamma_{\bullet s}' D_{\breve{d}}' D_{\breve{d}} \gamma_{\bullet s} \\ &= ||y - \mathbf{M}\gamma||^2 + \lambda||P\gamma||^2 + \breve{\lambda}||\breve{P}\gamma||^2, \end{aligned}$$

where $\gamma_{r\bullet}$ ($\gamma_{\bullet s}$) denotes the rth row (the sth column) of Γ. The explicit P-spline solution is

$$\hat{\gamma} = (\mathbf{M}'\mathbf{M} + \lambda P'P + \breve{\lambda}\breve{P}'\breve{P})^{-1}\mathbf{M}'y.$$

Two tuning parameters, associated with the row and column penalties, respectively, allowing continuous control over the surface. The predicted values are $\hat{y} = \mathbf{M}\hat{\gamma}$.

1.2 Second modeling component SISR

The second modeling component is single-index signal regression (SISR), which was presented in Eilers, Li, and Marx (2009) for one-dimensional signals, and is a method that can provide additional insight through the explicit modeling of any nonlinear behavior that may exist with the response. In fact, one could view the standard multivariate calibration problem as using an identity link function, which in actuality may be (slightly) misspecified.In effect, there may exist a true, but "missing link" function (that is nonlinear and monotone) (Cox, 1984), and this approach serves the purpose of estimating this link while improving external prediction. SISR introduces a modification: $\mu_i = f(\sum_{jk} x_{ijk}\alpha_{jk})$. The function $f(\cdot)$

is assumed to be smooth and is estimated from the data using univariate P-splines, having its own additional tuning parameter. This model is generally related to *projection pursuit* (Friedman and Stuetzle, 1981), with additional smoothness demands on α.

```
Algorithm MSISR
1. Initializations:
    • Choose the tuning parameter values $(\lambda, \breve{\lambda}, \lambda_f)$ for Steps 1 and 2
    • Choose number of knots $(n, \breve{n}, n_f)$
    • Choose penalty order $(d, \breve{d}, d_f)$
    • Set all tuning parameters to $\lambda_0$ for the initial Step 1 (default $10^6$)
    • Create $\mathbf{M} = X\mathbf{T}^\star$
    • Calculate $\hat{\gamma} = \text{MPSR}(\mathbf{M}, y, (\lambda_0, \lambda_0), (d, \breve{d}), (n, \breve{d}))$
2. Cycle until convergence of $\hat{\gamma}$'s
    • Estimate $\hat{f}$ and the estimate of the derivative $\dot{f}$ from $S(\mathbf{M}\hat{\gamma}, y, \lambda_f, d_f, n_f)$
    • Obtain $y^\star$ and $\mathbf{M}^\star$
    • Update $\hat{\gamma} = \text{MPSR}(\mathbf{M}^\star, y^\star, (\lambda, \breve{\lambda}), (d, \breve{d}), (n, \breve{n}))$
    • Constrain $\hat{\gamma}/||\hat{\gamma}||$
3. Prediction: $\hat{y}^{new} = \hat{f}(x^{new}\mathbf{T}^\star\hat{\gamma})$
end algorithm
```

1.3 The combined MSISR Methodology

The MSISR model has the form $\mu = f(\mathbf{M}\gamma)$, where the function f and the smooth coefficient surface are unspecified and approximated with P-spline coefficients α and γ. Consequently, the modified MPSR objective can be rewritten as

$$Q_P^\star = ||y - f(\mathbf{M}\gamma)||^2 + \lambda||P\gamma||^2 + \breve{\lambda}||\breve{P}\gamma||^2 + \lambda_f||D_d\alpha||^2. \qquad (2)$$

Given the tensor B-spline coefficient vector γ, the estimation of function f becomes a one-dimensional smoothing problem, and we can apply any scatter-plot smoother to obtain its estimate, which driven by the basis coefficient estimates $\hat{\alpha}$. We estimate f using a (cubic) P-spline scatter smoother (Eilers and Marx, 1996). The penalty on α ensures a smooth f; recall that α is the vector of B-spline coefficients with equally-spaced knots placed along η. Due to the virtue of using B-splines, the first derivative of f (denoted as $\dot{f}$), which is needed in our algorithm, can be easily computed (using a basis with one degree less and first differenced basis coefficients).

Once given an estimate of f, the coefficient vector γ can be estimated using a (first-order) Taylor series approximation of the function f (about the current estimate, γ_0). Specifically, if γ_0 is the current estimate for γ, then the current estimate of $\mu = f(\mathbf{M}\gamma)$ can be approximated by

$$f(\mathbf{M}\gamma) \approx f(\mathbf{M}\gamma_0) + \dot{f}(\mathbf{M}\gamma_0)\mathbf{M}(\gamma - \gamma_0). \tag{3}$$

Using (3), with fixed f, we have an approximation of $Q_P^\star$

$$\begin{aligned} Q_P^\star &\approx ||y - f(\mathbf{M}\gamma_0) - \dot{f}(\mathbf{M}\gamma_0)\mathbf{M}(\gamma - \gamma_0)||^2 + \lambda||P\gamma||^2 + \check{\lambda}||\check{P}\gamma||^2 \\ &= ||y^\star - \mathbf{M}^\star\gamma||^2 + \lambda||P\gamma||^2 + \check{\lambda}||\check{P}\gamma||^2, \end{aligned} \tag{4}$$

where $y^\star = y - f(\mathbf{M}\gamma_0) + \dot{f}(\mathbf{M}\gamma_0)\mathbf{M}\gamma_0$ and $\mathbf{M}^\star = \text{diag}\{\dot{f}(\mathbf{M}\gamma_0)\}\mathbf{M}$. Note that (4) implies that given f, the optimal α that minimizes the right-hand side of (4) can be obtained through a MPSR$(\mathbf{M}^\star, y^\star, (\lambda, \check{\lambda}), (D_d, D_{\check{d}}), (n, \check{n}))$. Hence, in our algorithm, we first carry out a MPSR with the response y on $\mathbf{M}$ (Step 1). Then, given γ, an estimate of f is obtained (Step 2). The two steps, estimation of f and γ, are iterated until convergence of $\hat{\gamma}$.

1.4 Aims and benefits of the combined MSISR approach

The estimation between f and α is iterative and tractable, essentially boiling down to repeated alternate applications of MPSR and P-spline smoothing on “working” responses and regressors. Some additional features of MSISR that are worthy of note include: (a) Although smooth, f can be assumed to be very general, an explicit function can be estimated. (b) Heavy penalization associated with f typically produces low degree polynomial estimates for f. (c) The entire signal can be used as regressors. (d) The number of highly spatially correlated regressors can far exceed the number of observations. (e) The parameterization yields a very manageable system of equations. (f) The candidate coefficient surface can be non-additive. (g) Since the two-dimensional signals and single estimated coefficient surface have a common indexing plane, potentially important regions can be visually identified.

2 Illustration and Optimization

We apply our MSISR to ternary mixture data. The responses are the mole fraction of a mixture, consisting of three components: water, 1,2-ethanediol, and 3-amino-1-propanol. There are 3 pure, 12 edge, and 19 interior (1 center) mixtures. The two-dimensional signal is constructed using the $p \times \check{p} = 4800$ digitized regressors, X_i, arranged using the (first) differenced UV-spectra, across the temperature levels. The indexing axes that define the support coordinates of X_i are specified as wavelength with $p = 400$ wavelength channels (701 to 1100nm, by 1 nm) and with $\check{p} = 12$ temperature levels (30, 35, 37.5, 40, 45, 47.5, 50, 55, 60, 62.5, 65, 70^o C). The data were not preprocessed in any other way.

We divided the $m = 34$ observation into three subsets as follows. The training set consisted of $m^{train} = 16$ observations using the 3 pure, 12 edge, and 1 center mixtures. The remaining 18 interior observations were divided into a validation set (to optimize tuning parameters) and a test set (to quantify quality of external prediction). Optimal tuning parameters were determined by minimizing RMSEV in the trained model. Given these optimal tuning parameters, external prediction was evaluated on the test data using RMSEP using the newly trained model that combined both the training and validation data. Table 1 presents the root mean square error of prediction (RMSEP) for the external prediction set, using optimal MSISR, MPSR, and PLS models. For responses water and 1,2-ethanediol, we find an improvement in external prediction for MSISR over both MPSR and PLS, leading to RMSEP reductions that range from 30% to 55%. For MSISR, the external RMSEP values are between 0.0214 and 0.0241, which when multiplied by 100 gives units of percent mixture. Figure 1 displays the optimal MSISR model using the response mixture component 1,2-ethanediol.

Table 1. MSISR, MPSR, PLS external prediction RMSEP, optimal models.

Response	MSISR	MPSR	PLS
Water	0.0214	0.0365	0.0465
1,2-ethanediol	0.0241	0.0338	0.0382
3-amino-1-propanol	0.0306	0.0251	0.0359

3 Discussion

We have shown how to estimate nonlinear relationships in multivariate calibration, by combining the single index model with multidimensional penalized signal regression. We found that the explicit estimation of the nonlinearity can provide some insights into the physical and chemical process underlying the measurements, which we view as a contribution over some of the other more "black box" approaches, while modestly improving external prediction. In the present case the response is assumed to have a normal distribution. Our other current research generalizes SISR, e.g., for binary classification, e.g. a Bernoulli response with probability π_i could be modeled with $\log(\pi/(1-\pi)) = f(X\beta)$. Additionally we are investigating two-dimensional surfaces for f, over another indexing variable, that allows for f to interact with, e.g., temperature.

References

Cox, C. (1984). Generalized linear models- the missing link. *Journal of the Royal Statistical Society, Series C*, **33**, 18-24.

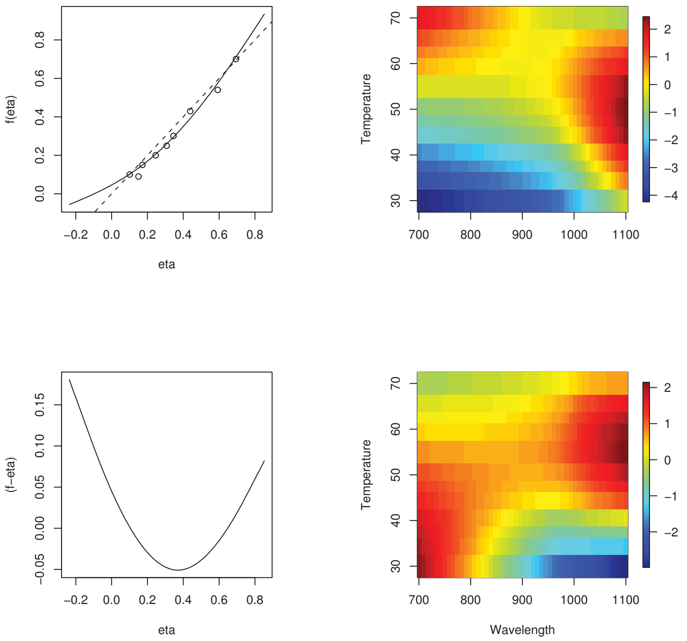

FIGURE 1. 1,2-ethanediol: The estimated $\hat{f}$ function is given (upper, left), along with $(\hat{f} - \hat{\eta})$ (lower, left). The plotted points represent the nine observations in the external test data set. The right panels provide the "optimal" image plots for the estimated coefficient surface (upper) and the coefficient surface difference, MSISR−MPSR (lower).

Eilers, P.H.C., Li, B. and Marx, B.D. (2009). Multivariate calibration with single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, **96**, 196-202.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science*, **11**, 89-121.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817-823.

Marx, B.D. and Eilers, P.H.C. (2005). Multidimensional penalized signal regression, *Technometrics* **47**, 13-22.