Penalized regression on principal manifolds with application to combustion modelling

Jochen Einbeck[1], Benjamin J. Isaac[2,3], Ludger Evers[4], Alessandro Parente[2]

[1] Durham University, Department of Mathematical Sciences, England
[2] Université Libre de Bruxelles, Service d'Aéro-Thermo-Mécanique, Brussels, Belgium
[3] University of Utah, Department of Chemical Engineering, United States
[4] University of Glasgow, School of Mathematics and Statistics, Scotland

E-mail for correspondence: jochen.einbeck@durham.ac.uk

Abstract: For multivariate regression problems featuring strong and non–linear dependency patterns between the involved predictors, it is attractive to reduce the dimension of the estimation problem by approximating the predictor space through a principal surface (or manifold). In this work, a new approach for non-parametric regression onto the fitted manifold is provided. The proposed penalized regression technique is applied onto data from a simulated combustion system, and is shown, in this application, to compare well with competing regression routines.

Keywords: Smoothing; principal component analysis; local principal manifolds; combustion model; numerical simulation.

1 Chemical Background

Combustion systems constitute a particular challenge for numerical modelling due to their high-dimensional and non–linear character. Typically, such systems involve a set of variables $\mathbf{\Phi} = [T, Z_1, \ldots, Z_{n_s-1}]$ where T is the temperature, and $Z_j, j = 1, \ldots, n_s - 1$ are the chemical species mass fractions of n_s chemical species. For instance, for simple fuels such as methane, the transport equations form a system of more than 50 highly coupled PDEs, of type

$$\rho \frac{D\mathbf{\Phi}}{Dt} = -\nabla \cdot (j_{\mathbf{\Phi}}) + s_{\mathbf{\Phi}} \tag{1}$$

where $\frac{D}{Dt}$ is the material–derivative operator, $j_{\mathbf{\Phi}}$ is the mass-diffusive flux of $\mathbf{\Phi}$, and $s_{\mathbf{\Phi}}$ is the "source term", that is the volumetric rate of production of

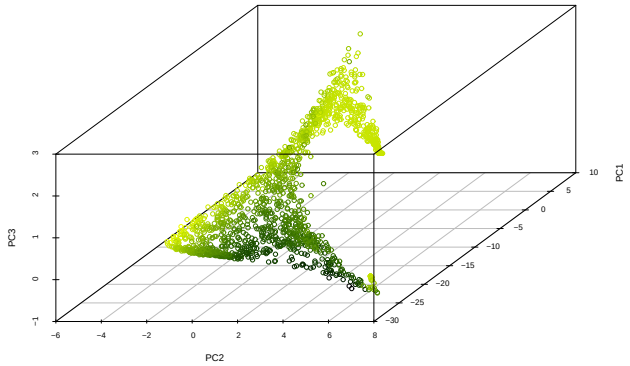

FIGURE 1. PC scores for a combustion system. Dark (green) color corresponds to small values of the first PC source term.

$\mathbf{\Phi}$. Increasingly complex fuels lead to an increase in the number of chemical species and reactions, and, hence, in the number of coupled PDEs as well as the computational costs. Moreover, large chemical mechanisms are usually stiff, i.e. a broad range of chemical time-scales exist, thus complicating the numerical simulations including detailed chemistry. Recognizing that the thermodynamic state of a reacting system relaxes onto a low-dimensional, strongly attracting manifold, Sutherland and Parente (2009) suggested the substitution of $\mathbf{\Phi}$ in (1) by a subset of its principal components, say $\boldsymbol{\eta}$, leaving a more tractable system of 2 or 3 transport equations,

$$\rho \frac{D\boldsymbol{\eta}}{Dt} = -\nabla \cdot (j_{\boldsymbol{\eta}}) + s_{\boldsymbol{\eta}}. \tag{2}$$

However, now the PC source terms $s_{\boldsymbol{\eta}}$ are unknown, and have to be found by regression onto the principal component scores. This tends to lead to unsatisfactory results, due to the nonlinear shape of the manifold. The ability to obtain precise regressions of the source terms is crucial to correctly solve the convection diffusion equation (2) that would describe the variation of the principal component during a numerical simulation. This paper addresses this problem by modelling the state space structure explicitly through local principal manifolds. A novel approach for penalized regression on the manifold surface is provided, and is shown to compare favourably with competing multivariate regression techniques. Of course, the applicability of the proposed method is not restricted to the chemical context considered in here.

2 Data and initial analysis

The data that we had available for this study was a "high-fidelity" data set (i.e., with tabulated source terms) provided by the University of Utah.

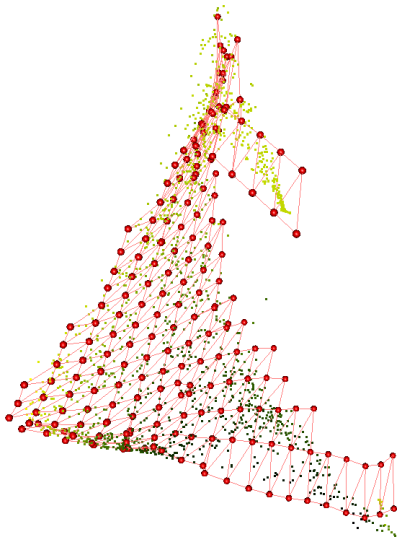

FIGURE 2. Principal surface (red vertices), with (training) data colored according to fitted response (see Sec. 4).

The data were obtained through an ODT (one-dimensional turbulence) model (Punati et al., 2011) and comprised a total of 11 variables, that is, 10 species mass fractions plus temperature.
Initially, PCA was applied onto (a scaled version of) a training data set of size $n = 4000$. Fig. 1 provides a plot of the first three principal component scores ($\boldsymbol{\eta}$). Dark (green) colors correspond to low first PC source terms. The manifold structure is evident here, as is the relevance of the position on the manifold for the first principal component source term. We proceeded with approximating the structure by a 'local principal surface' (Fig. 2), which effectively approximates the data by a mesh of tiny connected triangles (Einbeck and Evers, 2010).
The next, and most challenging step, is the regression of the PC source terms on the manifold. We explain the necessary methodology in the following section, and return to the combustion problem in Section 4.

3 Penalized regression on principal manifolds

For regression on principal surfaces (i.e., 2D manifolds), Einbeck and Evers (2010) suggested an algorithm which, in step 1, computes average responses within triangles, and in step 2, provides the fitted response in each triangle as the kernel-weighted average over the responses obtained in step 1. Here, we improve this (rather crude) method considerably by fitting piecewise linear functions on each triangle, "glued" together by a second order penalty

penalizing differences in the fitted responses at the triangle edges.
More precisely, the regression algorithm starts with projecting the data onto the manifold. Each data point $\mathbf{x}_i$ is projected onto the closest simplex of the principal manifold (which in our case is a triangle). Denote this simplex by s_i. The projection of $\mathbf{x}_i$ onto this triangle can then be expressed using the sides of the simplex as basis functions. Denote this coordinate vector of the projection of $\mathbf{x}_i$ onto the j-th simplex by $\mathbf{c}^{(j)}(\mathbf{x}_i)$.
The method now assumes different regression models for each simplex, i.e. for simplex j

$$y_i = \mathbf{c}^{(j)}(\mathbf{x}_i)'\boldsymbol{\beta}^{(j)} + \epsilon_i \qquad \text{for all } i \text{ such that the closest simplex } s_i = j.$$

Clearly, without additional penalty this model would be too parsimonious: neighbouring simplices would be allowed to have completely different regression functions. Thus a quadratic penalty is introduced which penalises the differences between predictions of neighbouring simplices at shared vertices. Denote by K the set of vertices (with coordinates $\mathbf{v}_k$) and by S_k the set of all simplices which contain the vertex k. Then the first quadratic penalty is

$$\sum_{k\in K}\sum_{j\in S_k}\left(\hat{y}_k^{(j)} - \bar{\hat{y}}_k\right)^2,$$

where $\bar{\hat{y}}_k = \frac{1}{|S_k|}\sum_{j\in S_k}\hat{y}_k^{(j)}$ and $\hat{y}_k^{(j)} = \mathbf{c}^{(j)}(\mathbf{v}_k)'\boldsymbol{\beta}^{(j)}$. This penalty however only shrinks the solution towards a continuous regression function. In order to obtain shrinkage towards a smooth regression function a second penalty is required. This penalty is based on the differences between the regression functions of neighbouring simplices. Define the opposite simplex $o(j,k)$ of simplex j w.r.t. vertex k as the simplex which shares all vertices with j, except k and one further vertex. Then the smoothness penalty can be written as

$$\sum_{k\in K}\sum_{j\in S_k}\left(\hat{y}_k^{(j)} - \hat{y}_k^{(o(j,k))}\right)^2$$

The problem can now be solved using one large penalised regression fit using $\mathbf{Z} = \mathbf{J}\square\mathbf{C}$ as design matrix, with $\mathbf{J}_{ij} = 1$ if $s_i = j$ and $\mathbf{J}_{ij} = 0$ otherwise, and $\mathbf{C} = \left(\mathbf{c}^{(s_1)}(\mathbf{x}_1)', \ldots, \mathbf{c}^{(s_n)}(\mathbf{x}_n)'\right)'$. The symbol $\square$ denotes the row-wise Kronecker product ("box product"), i.e. the i-th row of $\mathbf{Z}$ is the Kronecker product of the i-th row of $\mathbf{J}$ and the i-th row of $\mathbf{C}$. Using $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \\ \vdots \end{pmatrix}$ and rewriting the quadratic penalties from above as $\boldsymbol{\beta}'\mathbf{D}'\mathbf{D}\boldsymbol{\beta}$ and $\boldsymbol{\beta}'\mathbf{E}'\mathbf{E}\boldsymbol{\beta}$, the corresponding optimisation problem can be written as

$$\|\mathbf{Z}\boldsymbol{\beta} - \mathbf{y}\|^2 + \lambda\|\mathbf{D}\boldsymbol{\beta}\|^2 + \mu\|\mathbf{E}\boldsymbol{\beta}\|^2.$$

Though the matrices $\mathbf{Z}$, $\mathbf{D}$ and $\mathbf{E}$ can be very large, they are also very sparse, which allows for quick computations.

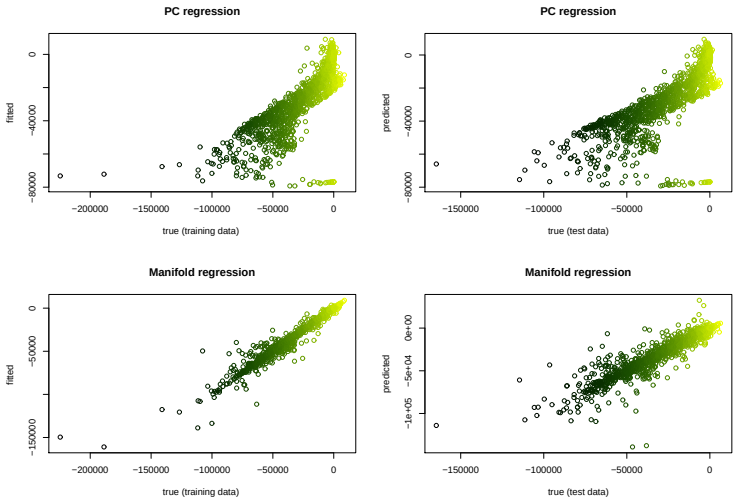

FIGURE 3. True versus fitted (predicted) responses, for training (left) and test data (right); each using linear PC regression (top), and manifold-based regression (bottom). The color scheme is the same as used in Fig. 1.

4 Results

The above technique, using $\lambda = \mu = 10^{-3}$, is now applied onto the system $\boldsymbol{\eta}$, with the first component of $s_{\boldsymbol{\eta}}$ serving as response, y. The fitted regression output is visualized by color in Fig. 2. A test data set of size 4000 was used to benchmark the performance of this regression technique against competing methods. Fig. 3 compares plots of true versus fitted (predicted) values for manifold-based and linear 'principal component' regression (PCR), which indicate that the manifold is able to produce good predictions for both training and test data.

In this study, we also consider the nonparametric additive model (AM), multivariate adaptive regression splines (MARS), and the support vector machine (SVM), each of them using the first three PC scores as predictors. Results are provided in Fig. 4, which also includes a comparison with the localized manifold regression technique proposed by Einbeck & Evers (2010), using smoothing parameter $\lambda = 0.1$.

The clear improvement, in particular of the median prediction error, compared to all other techniques is evident. The two manifold–based regression approaches perform similarly, but the penalized version appears superior since it enables regression *within* the triangle, enabling extremely precise predictions especially at parts of the flame where variability is low. It should be noted that the SVM did actually win the comparison in terms of *mean* (rather than median) prediction error, since it produces less 'very bad' predictions, but, in turn, performs (by construction) not very well where

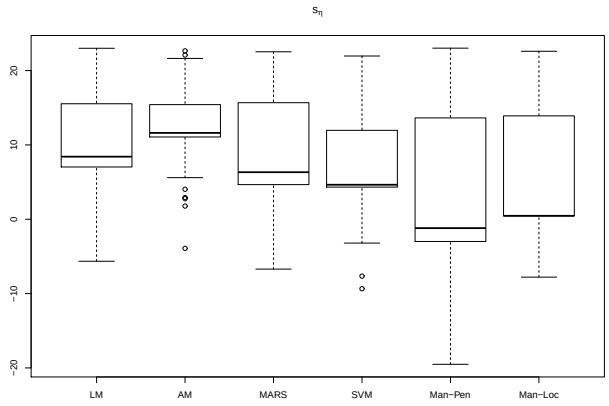

FIGURE 4. Log of squared prediction errors for a test data sample from the combustion data set, using, from left to right, PCR (LM), AM, MARS, SVM, penalized and localized regression on the manifold.

the information is very precise. We investigated this issue further and it appeared that those 'very bad' predictions for the penalized manifold regression relate to relatively 'unimportant' parts of the flame (burn-in process). Further improvement appears possible by refining the selection of smoothing parameters for the manifold estimation and regression, or by modifying the scaling used in the PCA step. Such issues are currently still under investigation.

Acknowledgments: This research was supported by scoping grant RF 060103 from the Durham Energy Institute.

References

Einbeck, J. and Evers, L. (2010). Localized regression on principal manifolds. In: *Proceedings of the 25th International Workshop on Statistical Modelling*, Glasgow, pages 179 – 184.

Punati, N., Sutherland, J.C., Kerstein, A.R., Hawkes, E.R., and Chen, J.H. (2011). An evaluation of the one-dimensional turbulence model: Comparison with direct numerical simulations of CO/H2 jets with extinction and reignition, *Proceedings of the Combustion Institute*, **33**.

Sutherland, J.C. and Parente, A. (2009). Combustion modeling using principal component analysis. *Proceedings of the Combustion Institute*, **32**, 1563 – 1570.