Recursive identification of smoothing spline ANOVA models

Marco Ratto, Andrea Pagano *
European Commission, Joint Research Centre, Ispra, ITALY

May 31, 2009

Keywords: Smoothing spline ANOVA models, recursive algorithms, backfitting, sensitivity analysis.

1 Introduction

In this paper we present a unified discussion of different approaches to identification of smoothing spline ANOVA models. The 'classical' approach to smoothing spline ANOVA models can be referred to in the line of Wahba (1990) and Gu (2002). Recently, Storlie et al. presented 'a new regularization method for simultaneous model fitting and variable selection in nonparametric regression models in the framework of smoothing spline ANOVA'. This method is an improvement on the COSSO (Lin and Zhang, 2006), penalizing the sum of component norms, instead of the squared norm employed in the traditional smoothing spline method. Storlie et al. introduce an adaptive weight to be used in the COSSO penalty which allows for more flexibility to estimate important functional components while giving heavier penalty to unimportant functional components.

In a 'parallel' stream of research, using the the so-called State-Dependent Parameter Regression (SDR) approach of Young (2001), Ratto et al. (2007) have developed a non-parametric approach which is very similar to smoothing splines and kernel regression approaches, but which is based on recursive filtering and smoothing estimation (the Kalman Filter combined with Fixed Interval Smoothing). Such a recursive least-squares implementation has some key characteristics: (a) it is couched with optimal Maximum Likelihood estimation, thus allowing for an objective estimation of the smoothing hyper-parameters , and (b) it allows for greater flexibility in adapting to local discontinuities, heavy non-linearity and heteroscedastic error terms.

The purposes of this paper are:

1. develop a formal comparison and demonstrate equivalences between the 'classical' tensor product cubic spline approach with reproducing Kernel Hilbert space algebra (RKHS) and the SDR approach;

* *Address for correspondence:* Marco Ratto, Joint Research Centre, European Commission, TP361 IPSC, via Fermi 1, 21027 Ispra (VA), Italy . E-mail: marco.ratto@jrc.it

2. discuss advantages and disadvantages of these approaches;

3. propose a unified approach to smoothing spline ANOVA models that combines the best of the discussed methods, in particular the use of the recursive algorithms can be very effective in detecting the important functional components, adding valuable information in the ACOSSO framework.

2 State Dependent Regressions and smoothing splines

2.1 Additive models

Denote the generic mapping as $z(\mathbf{X})$ and assume without loss of generality that $\mathbf{X} \in [0,1]^p$, where p is the number of parameters. The simplest example of smoothing spline mapping estimation of z is the additive model:

$$f(\mathbf{X}) = f_0 + \sum_{j=1}^{p} f_j(X_j) \tag{1}$$

To estimate f we can use a multivariate smoothing spline minimization problem, that is, given λ, find the minimizer $f(X_k)$ of:

$$\frac{1}{N}\sum_{k=1}^{N}(z_k - f(\mathbf{X}_k))^2 + \sum_{j=1}^{p}\lambda_j \int_0^1 [f_j''(X_j)]^2 dX_j \tag{2}$$

where a Monte Carlo sample of dimension N is assumed.

This minimization problem requires the estimation of the p hyper-parameters λ_j (also denoted as smoothing parameters). Various ways of doing that are available in the literature, by applying generalized cross-validation (GCV), generalized Maximum Likelihood procedures (GML) and so on (see e.g. Wahba, 1990; Gu, 2002). Here we discuss the estimation approach, where the additive model is put into a *State-Dependent Parameter Regression* (SDR) form of Young (2001) and as applied to the estimation of ANOVA models by Ratto et al. (2007). We synthesize here the key features of the recursive algorithms of SDR, by considering the case of $p = 1$ and $z(X) = g(X) + e$, with $e \sim N(0, \sigma^2)$, i.e. we rewrite the smoothing problem as $z_k = s_k + e_k$, where $k = 1, \ldots, N$ and s_k is the estimate of $g(X_k)$.

In order to recursively estimate the s_k in SDR, it is necessary to characterize it in some stochastic manner, borrowing from non-stationary time series processes. In general this is accomplished by assuming that the evolution of s_k follows one member of the Generalized Random Walk (GRW) class on non-stationary random sequences (see e.g. Young and Ng, 1989; Ng and Young, 1990). In the present context, the integrated random walk (IRW) process provides the same smoothing properties of a cubic spline, in the overall State-Space (SS) formulation:

$$\begin{array}{lrcl} \text{Observation Equation:} & z_k & = & s_k + e_k \\ \text{State Equations:} & s_k & = & s_{k-1} + d_{k-1} \\ & d_k & = & d_{k-1} + \eta_k \end{array} \tag{3}$$

where d_k is the 'slope' of s_k, $\eta_k \sim N(0, \sigma_\eta^2)$ and η_k ('system disturbance' in systems terminology) is assumed to be independent of the observation noise e_k.

In order to estimate s_k using the recursive Kalman Filter (KF) and the associated recursive Fixed Interval Smoothing (FIS) algorithm (see e.g. Kalman, 1960; Young, 1999, for details), the Monte Carlo sample has to be sorted in ascending order of X, i.e. the k and $k-1$ subscripts in (3) *denote adjacent elements under such ordering.*

First, it is necessary to optimize the hyper-parameters associated with the state space model (3), namely the white noise variances σ^2 and σ_η^2. In fact, by a simple reformulation of the KF and FIS algorithms, the IRW model can be entirely characterized by one Noise Variance Ratio (NVR) hyper-parameter, where NVR= σ_η^2/σ^2. These NVR values are, of course, unknown *a priori* and need to be optimized: for example, in the above references, this is accomplished by maximum likelihood optimization using prediction error decomposition (Schweppe, 1965). The NVR plays the inverse role of a smoothing parameter: the smaller the NVR, the smoother the estimate of s_k (and in the limit NVR=0, s_k will be a straight line). Given the NVR, the FIS algorithm then yields an estimate $\hat{s}_{k|N}$ of s_k at each data sample and it can be seen that the $\hat{s}_{k|N}$ from the IRW process is the equivalent of $f(X_k)$ in the smoothing spline model. At the same time, the recursive procedures provide, in a natural way, standard errors of the estimated $\hat{s}_{k|N}$, that allow for testing their relative significance.

We need to clarify here the meaning of the ML optimization in this recursive context. In the 'classical' smoothing spline estimates, a 'penalty' is always plugged in the objective function (GCV, GML, etc.) used to optimize the λ's, in order to limit the 'degrees of freedom' of the spline model. For example, in generalized cross-validation, we have to find λ that minimizes

$$GCV_\lambda = 1/N \cdot \frac{\sum_k (z_k - f_\lambda(X_k))^2}{(1 - df(\lambda)/N)^2}, \tag{4}$$

where $df \in [0, N]$ denotes the 'degrees of freedom' of the spline and where we have explicitly indicated the dependency on λ in the GCV formula. In the recursive notation just introduced, we could equivalently write

$$GCV_{NVR} = 1/N \cdot \frac{\sum_k (z_k - \hat{s}_{k|N})^2}{(1 - df(NVR)/N)^2}. \tag{5}$$

Without the penalty term, the optimum would always be attained at $\lambda = 0$ (or $NVR \to \infty$), i.e. perfect fit. In the SDR recursive context, however, the penalty concept is intrinsically plugged in by the fact that the prediction error decomposition (ML) estimate is based on the *filtered* estimate $\hat{s}_{k|k-1} = s_{k-1} + d_{k-1}$ and not on the smoothed estimate $\hat{s}_{k|N}$, namely we find NVR that minimizes:

$$\begin{aligned} -2 \cdot \log(L) &= const + \sum_{k=3}^{N} \log(1 + P_{k|k-1}) + (N-2) \cdot \log(\hat{\sigma^2}) \\ \hat{\sigma^2} &= \frac{1}{N-2} \sum_{k=3}^{N} \frac{(z_k - \hat{s}_{k|k-1})^2}{(1 + P_{k|k-1})} \end{aligned} \tag{6}$$

where $P_{k|k-1}$ is the one step ahead forecast error of the state $\hat{s}_{k|k-1}$ provided by the Kalman Filter. It seems useful to underline here that $\hat{s}_{k|k-1}$ is based only on the information contained in the sample values $[1, \ldots, k-1]$ while smoothed estimates use the entire information set $[1, \ldots, N]$. Therefore, it can be easily seen that, for the recursive case, the limit $NVR \to \infty$ ($\lambda \to 0$) is no longer a 'perfect fit' situation, since a zero variance for e_k implies $\hat{s}_{k|k-1} = s_{k-1} + d_{k-1} = z_{k-1} + d_{k-1}$, i.e. the one step ahead prediction of z_k is given by the linear extrapolation of the adjacent value z_{k-1}, so implying a non-zero prediction error in this limit case. So, it is clear the 'perfect fit' will not be, in general, the optimal solution.

To complete the equivalence between the SDR and cubic spline formulations, we need to link the NVR estimated by the ML procedure to the smoothing parameters λ. This is easily accomplished by setting $\lambda = 1/(\text{NVR} \cdot N^4)$.

In the general additive case (1), the recursive procedure just described needs to be applied, in turn, for each term $f_j(X_{j,k}) = \hat{s}_{j,k|N}$, requiring a different sorting strategy for each $\hat{s}_{j,k|N}$. Hence the 'backfitting' procedure, as described in Young (2000, 2001), is exploited. Finally, the estimated NVR_j's can be converted into λ_j values and the additive model put into the standard cubic spline form.

2.2 ANOVA models with interaction functions

The additive model concept (1) can be generalized to include 2-way (and higher) interaction functions via the functional ANOVA decomposition. For example, we can let

$$f(\mathbf{X}) = f_0 + \sum_{j=1}^{p} f_j(X_j) + \sum_{j<i}^{p} f_{j,i}(X_j, X_i) \tag{7}$$

In the ANOVA smoothing spline context, corresponding optimization problems with interaction functions and their solutions can be obtained conveniently with the reproducing kernel Hilbert space (RKHS) approach (see Wahba 1990). In the SDR context, an interaction function is formalized as the product of two states $s_1 \cdot s_2$, each of them characterized by an IRW stochastic process. Hence the estimation of a single interaction term $z(\mathbf{X}_k) = f(X_{1,k}, X_{2,k}) + e_k$ is formalized as:

$$\begin{array}{lrcl} \text{Observation Equation:} & z_k & = & s^I_{1,k} \cdot s^I_{2,k} + e_k \\ \text{State Equations: } (j = 1, 2) & s^I_{j,k} & = & s^I_{j,k-1} + d^I_{j,k-1} \\ & d^I_{j,k} & = & d^I_{j,k-1} + \eta^I_{j,k} \end{array} \tag{8}$$

where $I = 1, 2$ is a multi-index denoting the interaction term under estimation and $\eta^I_{j,k} \sim N(0, \sigma^2_{\eta^I_j})$. The two terms $s^I_{j,k}$ are estimated iteratively by running the recursive procedure in turn, i.e.

- take an initial estimate of $s^I_{1,k}$ and $s^I_{2,k}$ by regressing z with the product of simple linear or quadratic polynomials $p_1(X_1) \cdot p_2(X_2)$ and set $s^{I,0}_{j,k} = p_j(X_{j,k})$;
- iterate $i = 1, 2$:
  - fix $s^{I,i-1}_{2,k}$ and estimate NVR^I_1 and $s^{I,i}_{1,k}$ using the recursive procedure;

– fix $s_{1,k}^{I,i}$ and estimate NVR_2^I and $s_{2,k}^{I,i}$ using the recursive procedure;

- the product $s_{1,k}^{I,2} \cdot s_{2,k}^{I,2}$ obtained after the second iteration provides the recursive SDR estimate of the interaction function.

Unfortunately, in the case of interaction functions we cannot derive an explicit and full equivalence between SDR and cubic splines of the type mentioned for first order ANOVA terms. Therefore, in order to be able to exploit the estimation results in the context of a smoothing spline ANOVA model, we take a different approach, similarly to the ACOSSO case.

2.3 Very short summary of ACOSSO

We make the usual assumption that $f \in \mathcal{F}$, where $\mathcal{F}$ is a RKHS. The space $\mathcal{F}$ can be written as an orthogonal decomposition $\mathcal{F} = \{1\} \oplus \{\bigoplus_{j=1}^{q} \mathcal{F}_j\}$, where each $\mathcal{F}_j$ is itself a RKHS and $j = 1, \ldots, q$ spans over ANOVA terms of various order. We re-formulate (2) for the general case with interactions as the function f that minimizes:

$$\frac{1}{N}\sum_{k=1}^{N}(z_k - f(\mathbf{X}_k))^2 + \lambda_0 \sum_{j=1}^{q} \frac{1}{\theta_j}\|P^j f\|_{\mathcal{F}}^2 \tag{9}$$

where $P^j f$ is the orthogonal projection of f onto $\mathcal{F}_j$ and the q-dimensional vector of θ_j smoothing parameters needs to be optimized somehow. This is typically a formidable problem and in the simplest case θ_j is set to one, with the single λ_0 estimated by GCV of GML. The COSSO (Lin and Zhang, 2006) penalizes the sum of norms, which allows to identify the informative predictor terms f_j with an estimate of f that minimizes

$$\frac{1}{N}\sum_{k=1}^{N}(z_k - f(\mathbf{X}_k))^2 + \lambda \sum_{j=1}^{q} \|P^j f\|_{\mathcal{F}} \tag{10}$$

using a single smoothing parameter λ. COSSO improves considerably the problem (9) with $\theta_j = 1$ and is much more computationally efficient than the full problem (9) with optimized θ_j's.

In the adaptive COSSO (ACOSSO) of Storlie et al., $f \in \mathcal{F}$ minimizes

$$\frac{1}{N}\sum_{k=1}^{N}(z_k - f(\mathbf{X}_k))^2 + \lambda \sum_{j=1}^{q} w_j \|P^j f\|_{\mathcal{F}} \tag{11}$$

where $0 < w_j \leq \infty$ are weights that depend on an initial estimate of $\tilde{f}$, either using (9) with $\theta_j = 1$ or the COSSO estimate (10). The adaptive weights are obtained as $w_j = \|P^j \tilde{f}\|_{L_2}^{-\gamma}$, with $\gamma = 2$ typically and the L_2 norm $\|P^j \tilde{f}\|_{L_2} = (\int (P^j \tilde{f}(\mathbf{X}))^2 d\mathbf{X})^{1/2}$.

2.4 Combining SDR and ACOSSO for interaction functions

There is an obvious way of exploiting the SDR identification and estimation steps in the ACOSSO framework: namely, the SDR estimates of additive and interaction function terms can be taken as the initial $\tilde{f}$ used to compute the weights in the ACOSSO. However, this would be a minimal approach while the SDR identification and estimation provides a more detailed information about f_j terms that is worth exploiting. We define $\mathcal{K}_{\langle j \rangle}$ the reproducing kernel of an additive term $\mathcal{F}_j$ of the ANOVA decomposition of the space $\mathcal{F}$. In the cubic spline case, this is constructed as the sum of two terms $\mathcal{K}_{\langle j \rangle} = \mathcal{K}_{01\langle j \rangle} \oplus \mathcal{K}_{1\langle j \rangle}$ where $\mathcal{K}_{01\langle j \rangle}$ is the r.k. of the parametric (linear) part and $\mathcal{K}_{1\langle j \rangle}$ is the r.k. of the purely non-parametric part. The second order interaction terms are constructed as the tensor product of the first order terms, for a total of four elements, i.e.

$$\begin{aligned} \mathcal{K}_{\langle i,j \rangle} &= (\mathcal{K}_{01\langle i \rangle} \oplus \mathcal{K}_{1\langle i \rangle}) \otimes (\mathcal{K}_{01\langle j \rangle} \oplus \mathcal{K}_{1\langle j \rangle}) \qquad (12) \\ &= (\mathcal{K}_{01\langle i \rangle} \otimes \mathcal{K}_{01\langle j \rangle}) \oplus (\mathcal{K}_{01\langle i \rangle} \otimes \mathcal{K}_{1\langle j \rangle}) \oplus (\mathcal{K}_{1\langle i \rangle} \otimes \mathcal{K}_{01\langle j \rangle}) \oplus (\mathcal{K}_{1\langle i \rangle} \otimes \mathcal{K}_{1\langle j \rangle}) \end{aligned}$$

In general, considering the problem (9), one should attribute a specific coefficient $\theta_{\langle \cdot \rangle}$ to each single element of the r.k. of $\mathcal{F}_j$ (see e.g. Gu, 2002, Chapter 3), i.e. two θ's for each main effect, four θ's for each two-way interaction, and so on. In fact, each $\mathcal{F}_j$ would be optimally fitted by opportunely choosing weights in the sum of $\mathcal{K}_{\langle \cdot,\cdot \rangle}$ elements. This, however, makes the estimation problem rather complex, so, usually, the tensor product (12) is directly used, without tuning the weights of each element of the sum. This strategy is also applied in ACOSSO.

Now, considering the SDR estimate of the interaction (8), we can see that it is given by the product of two univariate cubic splines. So, one can easily decompose each estimated $\hat{s}^I_j$ into the sum of a linear ($\hat{s}^I_{01\langle j \rangle}$) and non-parametric term ($\hat{s}^I_{1\langle j \rangle}$). This provides a decomposition of the SDR interaction of the form

$$\hat{s}^I_i \cdot \hat{s}^I_j = \hat{s}^I_{01\langle i \rangle}\hat{s}^I_{01\langle j \rangle} + \hat{s}^I_{01\langle i \rangle}\hat{s}^I_{1\langle j \rangle} + \hat{s}^I_{1\langle i \rangle}\hat{s}^I_{01\langle j \rangle} + \hat{s}^I_{1\langle i \rangle}\hat{s}^I_{1\langle j \rangle}, \qquad (13)$$

that is a proxy of the four elements of the r.k. of the second order tensor product cubic spline.

This is implies that the optimal use of the SDR identification and estimation in the ACOSSO framework is to apply specific weights to each element of the r.k. $\mathcal{K}_{\langle \cdot,\cdot \rangle}$, using the L_2 norms of each of the four elements in (13).

3 Examples

We compared the combined SDR-ACOSSO approach with ACOSSO and DACE on several examples (full details including routines are freely available at request). First we checked the behavior of SDR in identifying single 2-way interaction functions, i.e. we took a number of surfaces $z(X_1, X_2) = g(X_1, X_2) + e$, with $e \sim N(0, \sigma)$, using different levels of signal to noise ratios $SNR = V(z)/V(e)$: very large ($SNR > 10$), middle ($SNR \sim 3$), very small ($SNR \sim 0.1$). We compared SDR results with standard GCV estimation and with DACE (extended to include observation noise) using a training MC sample $\mathbf{X}$ of 256 elements and

method	$p = 4$ 'simple'	$p = 4$ 'nasty'	$p = 8$ 'simple'	$p = 10$ 'nasty'
SDR-ACOSSO	0.9994	0.8633	0.9928	0.1922
ACOSSO	0.9986	0.7910	0.9163	0.1963
DACE	0.9932	0.8174	0.9715	-0.0247

Table 1: SDR-ACOSSO, ACOSSO and DACE: average R^2 (out of sample) computed on 100 replicas for different types of the Sobol' g-functions.

tested the out-of sample performance of each method in predicting the 'noise-free' signal $g(X_1, X_2)$ using a new validation sample $\mathbf{X}^*$ of dimension 256. We repeated this exercise on 100 random replicas for each function and each SNR. We considered 9 types of surfaces of increasing order of complexity (i.e. 27 different surface identification, each replicated 100 times). Only for one out of the nine surfaces, DACE outperformed SDR or GCV estimation. In the other cases, SDR and GCV gave similar results, when the four terms in (13) have similar weights, while SDR was extremely efficient in better identifying surfaces characterized by different weights. These results suggested that SDR identification step can provide significant added value in smoothing spline ANOVA modelling.

We then performed full emulation exercises, considering the analytic Sobol' g-function (Saltelli et al., 2000) with different dimension p and degree of interaction (denoted as 'simple' and 'nasty' in Table 1). We considered a training sample of dimension 256 to estimate the emulators and used a new validation sample of the same dimension to check the out of sample performance. We repeated the analysis 100 times for each function and each method. The results in Table 1 indicate that SDR identification provides a very significant added value in smoothing spline ANOVA models, making it favorable also compared to DACE.

4 Conclusions

In general, it is not possible to identify a method (among ACOSSO, DACE, SDR-ACOSSO) which outperforms the others in all examples. SDR is extremely rapid, efficient and accurate in identifying *additive models*: the use of recursive algorithms, in fact, avoids the inversion of large matrices which is needed in the other methods (ACOSSO, DACE). In the case of ANOVA models with interaction components, ACOSSO confirms entirely its good performances in terms of efficiency and relatively low computational cost. When the model includes interactions, SDR combined with ACOSSO improves ACOSSO in many cases, although at the price of a significantly higher computational cost. So, while for additive models the advantage of SDR is in both low computational cost and of accuracy, when interactions are included the greater accuracy of SDR-ACOSSO has a cost. SDR-ACOSSO also compares very favorably with respect to DACE in many cases, even if there are cases where DACE outperforms SDR-ACOSSO in out-of-sample prediction. In terms of computational burden, we suggest that SDR (for additive models) and ACOSSO (for models with interactions) are taken as the first choice for a *rapid and reliable* emulation exercise. Should ACOSSO be unable to explain a large part of the mapping, SDR-ACOSSO or DACE should be taken into consideration. We also noted that DACE is not necessarily the best choice when the model is supposed to be very complex and

with significant interactions. DACE, as an any interpolation method, tries to exploit the 'zero-uncertainty' at observed samples of the mapping z. However, when the model is complex, it can wrongly identify spurious interaction terms involving unimportant X's, possibly explaining the number of cases of poorer performance in out-of-sample predictions with respect to smoothing methods. SDR-ACOSSO, on the other hand, can provide detailed information about the form of each additive and interaction term of a truncated the ANOVA decomposition, often allowing very good out-of-sample predictions.

References

C. Gu. *Smoothing Spline* ANOVA *Models*. Springer-Verlag, 2002.

R.E. Kalman. A new approach to linear filtering and prediction problems. *ASME Trans., Journal Basic Eng.*, 82D:35–45, 1960.

Y. Lin and H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics*, 34:2272–2297, 2006.

C.N. Ng and P. C. Young. Recursive estimation and forecasting of non-stationary time series. *Journal of Forecasting*, 9:173–204, 1990.

M. Ratto, A. Pagano, and P. C. Young. State dependent parameter metamodelling and sensitivity analysis. *Computer Physics Communications*, 177: 863–876, 2007.

A. Saltelli, K. Chan, and M. Scott, editors. *Sensitivity Analysis*. Wiley Series in Probability and Statistics. John Wiley and Sons, New York, 2000.

F. Schweppe. Evaluation of likelihood functions for Gaussian signals. *IEEE Trans. on Information Theory*, 11:61–70, 1965.

C.B. Storlie, H. Bondell, B. Reich, and H.H. Zhang. The adaptive cosso for nonparametric surface estimation and model selection. *The annals of statistics*. submitted.

G. Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics., 1990.

P. C. Young. Nonstationary time series analysis and forecasting. *Progress in Environmental Science*, 1:3–48, 1999.

P. C. Young. The identification and estimation of nonlinear stochastic systems. In A. I. et al. Mees, editor, *Nonlinear Dynamics and Statistics*. Birkhauser, Boston, 2001.

P. C. Young. Stochastic, dynamic modelling and signal processing: Time variable and state dependent parameter estimation. In W. J. Fitzgerald, A. Walden, R. Smith, and P. C. Young, editors, *Nonlinear and Nonstationary Signal Processing*, pages 74–114. Cambridge University Press, Cambridge, 2000.

P. C. Young and C. N. Ng. Variance intervention. *Journal of Forecasting*, 8: 399–416, 1989.