

DEPARTMENT OF STATISTICS

University of Wisconsin

1300 University Avenue

Madison, WI 53706

TECHNICAL REPORT NO. 1098

December 8, 2004

A Characterization of the log density smoothing spline ANOVA model

Yi Lin & Yongho Jeon

Department of Statistics, University of Wisconsin, Madison WI

This research was supported in part by National Science Foundation grant DMS 0134987.

A Characterization of the log density smoothing spline ANOVA model

BY YI LIN AND YONGHO JEON

Department of Statistics, University of Wisconsin–Madison

yilin@stat.wisc.edu yjeon@stat.wisc.edu

SUMMARY

In this paper we introduce a characterization of the log density smoothing spline ANOVA model. We show that in a log density ANOVA model of order r (consisting of the main effects and all the interactions of order up to r), the joint density function is uniquely determined by the collection of all r dimensional marginal densities. Furthermore, the order r model is the largest log density ANOVA model under which the joint density function is uniquely determined by the r dimensional marginals. Our results are valid for log density ANOVA model with other general structures. In general, in log density ANOVA models the joint density function is uniquely determined by the marginal densities corresponding to the terms present in the ANOVA model. For example, the ANOVA model consists of main effects, the two way interactions η_{12} , η_{23} , η_{13} , η_{14} , η_{34} , and the three way interaction η_{134} is the largest log density ANOVA model under which the joint density is uniquely determined by the 12, 23, 134 marginals.

1. INTRODUCTION

In a general nonparametric density estimation problem we are given a random sample of a d -dimensional random vector $X = (X_1, \dots, X_d)$, and we are interested in estimating the density function $p(\cdot)$ of X . The density function $p(\cdot)$ is not assumed to lie in a prespecified parametric family of distributions, but is only assumed to satisfy certain smoothness conditions.

Much progress has been made in low (one or two) dimensional nonparametric density estimation, and a number of practical algorithms have been produced. See, for example, Tapia & Thompson (1978), Silverman (1986), and Wand & Jones (1995). In particular, Silverman (1982) proposed to estimate log density instead of density, and studied the penalized maximum likelihood estimator for log density. When it comes to high dimensional nonparametric density estimation, however, extra difficulties are encountered due to the so called curse of dimensionality. A general high dimensional density function is hard to estimate, both in terms of accuracy and computational cost. Even when an accurate estimate is available, a complicated high dimensional density function can be very hard to interpret.

The log density functional ANOVA model provides a powerful framework for the estimation and interpretation of high dimensional density functions. In such a model the log density function is decomposed as a sum of a constant term, one dimensional functions (main effects), two dimensional

functions (two way interactions), and so on:

$$\eta(\mathbf{x}) = \text{constant} + \sum_{j=1}^d \eta_j(x_j) + \sum_{j < k} \eta_{jk}(x_j, x_k) + \cdots, \quad (1.1)$$

where the components satisfy side conditions that guarantee uniqueness, and the series is usually truncated in some manner to enhance interpretability. The log density functional ANOVA model has been studied extensively in the literature. See, for example, Gu (2002) and Gu & Wang (2003) for a review. Notice that the additive log density model (with no interaction terms) actually assumes independence among the variables. It is also known that conditional independence assumptions can be incorporated in the model by excluding particular terms in the decomposition. For example, in a three dimensional problem, the absence of the terms η_{23} and η_{123} in the log density ANOVA decomposition (1.1) indicates that the random variables X_2 and X_3 are conditionally independent given X_1 . This underlies the close connection between the log density ANOVA model and the popular graphical models, which use graphs to intuitively represent the dependence structure among the variables (see, for example, Edwards (2000)).

When there is no prior information on the dependence structure, the common practice is to truncate the decomposition in (1.1) by excluding all interactions of order higher than a certain chosen number r , resulting in a log density ANOVA model of order r . The most common choice of r is two. Such truncation enhances interpretability since each component function is then of low dimension and can be visualized. The truncation also helps compute the estimate by including fewer components in the model. However, any truncation in the log density functional ANOVA decomposition puts a constraint on the multivariate density that may or may not be appropriate for the application at hand, and it is of practical importance to have an intuitive understanding of the consequence of assuming a particular (truncated) functional ANOVA model. In this paper we give an intuitive characterization of the log density (truncated) functional ANOVA model. Our results can be motivated by the following consideration. In the log density ANOVA model of order one (the additive model), it is clear that the joint density is the product of the one dimensional marginal densities, and therefore the one dimensional marginals uniquely determine the joint density. It is then of interest whether the r dimensional marginals uniquely determine the joint density in the log density ANOVA model of order $r \geq 2$. We show that the answer to this question is positive, and that the order r model is the largest log density ANOVA model under which the joint density function is uniquely determined by the r dimensional marginals. Our results hold more generally: in the log density ANOVA model the joint density function is uniquely determined by the marginal densities corresponding to the terms present in the ANOVA model.

For simplicity of presentation, we mainly concentrate on the popular smoothing spline log density ANOVA model, but our results apply to other types of functional ANOVA models, as discussed in Section 4. In Section 2 we briefly introduce the log density smoothing spline ANOVA model. In Section 3, we give our main results. In Section 4 we give some discussions.

2. SMOOTHING SPLINE ANOVA MODELS

Let H_j be the Sobolev Hilbert space of univariate functions of $x_j \in [0, 1]$: $H = \{g : g, g', \dots, g^{(m-1)}$ are absolutely continuous, $g^{(m)} \in \mathcal{L}_2[0, 1]\}$. Here $m \geq 1$ is the order of the Sobolev Hilbert space. Following Wahba (1990), we define the norm in H by

$$\|g\|^2 = \sum_{\nu=0}^{m-1} \left\{ \int_0^1 g^{(\nu)}(t) dt \right\}^2 + \int_0^1 \{g^{(m)}(t)\}^2 dt. \quad (2.1)$$

With this norm, H can be decomposed as the direct sum of the constant space and its orthogonal complement: $H = \{1\} \oplus \bar{H}$. The tensor product space $\otimes_{j=1}^d H_j$ then has the following decomposition:

$$\otimes_{j=1}^d H_j = \{1\} \oplus \sum_{j=1}^d \bar{H}_j \oplus \sum_{j < k} [\bar{H}_j \otimes \bar{H}_k] \oplus \dots. \quad (2.2)$$

In the commonly used smoothing spline ANOVA model (SS-ANOVA), each functional component in the functional ANOVA decomposition (1.1) is assumed to lie in the corresponding component space in the orthogonal decomposition (2.2). That is, the main effect f_j lies in \bar{H}_j , and the interactions reside in the tensor product spaces of univariate function spaces. See Wahba (1990), Gu (2002), and Ke & Wang (2004).

There are other types of functional ANOVA models. They differ from the SS-ANOVA in that the assumed function spaces are different. For example, it is possible to assume that $f_j \in H_j$, f_{ij} lies in the m -th order Sobolev Hilbert space of bivariate functions on $[0, 1]^2$, and so on. In this case it is required that $m > r/2$, where r is the highest order of interactions in the functional ANOVA model.

3. A CHARACTERIZATION OF THE LOG DENSITY SS-ANOVA MODEL

Let \mathcal{F}_r be the function space corresponding to the SS-ANOVA model of order r . This is the subspace of (2.2) obtained by truncating the decomposition in (2.2) according to the truncation in the SS-ANOVA (1.1) of order r . For example,

$$\mathcal{F}_2 = \{1\} \oplus \sum_{j=1}^d \bar{H}_j \oplus \sum_{j < k} [\bar{H}_j \otimes \bar{H}_k].$$

For any $\eta_0 \in \mathcal{F}_r$, we define a functional $L_{\eta_0}(\cdot): \mathcal{F}_r \rightarrow R$ by

$$L_{\eta_0}(\eta) = \int \exp(\eta(\mathbf{x})) d\mathbf{x} - \int \eta(\mathbf{x}) \exp(\eta_0(\mathbf{x})) d\mathbf{x}, \quad \forall \eta \in \mathcal{F}_r.$$

LEMMA 3.1. *Fix any $\eta_0 \in \mathcal{F}_r$. The problem of minimizing $L_{\eta_0}(\eta)$ over $\eta \in \mathcal{F}_r$ has the unique solution η_0 .*

Proof. Consider the first and second order Fréchet derivatives of $L_{\eta_0}(\eta)$:

$$\begin{aligned} DL_{\eta_0}(\eta)g &= \int e^{\eta(\mathbf{x})}g(\mathbf{x})d\mathbf{x} - \int g(\mathbf{x})e^{\eta_0(\mathbf{x})}d\mathbf{x} = \int g(\mathbf{x})(e^{\eta(\mathbf{x})} - e^{\eta_0(\mathbf{x})})d\mathbf{x}, \\ D^2L_{\eta_0}(\eta)gh &= \int e^{\eta(\mathbf{x})}g(\mathbf{x})h(\mathbf{x})d\mathbf{x}, \end{aligned}$$

where D denotes the Fréchet derivative operator. Since $D^2L_{\eta_0}(\eta)gg = \int \exp(\eta(\mathbf{x}))g^2(\mathbf{x})d\mathbf{x} > 0$ for any $g \neq 0$, we have that $L_{\eta_0}(\cdot)$ is strictly convex over \mathcal{F}_r . Setting $DL_{\eta_0}(\eta)g = 0, \forall g$, we get $\eta = \eta_0$. Therefore η_0 is the unique minimizer of $L_{\eta_0}(\eta)$ over η . \square

We remark that in Lemma 3.1 we only require $\eta_0 \in \mathcal{F}_r$ and $\eta \in \mathcal{F}_r$. Neither of them needs to be a log density, since $\exp(\eta_0)$ and $\exp(\eta)$ may not integrate to unity.

THEOREM 3.1. *Let $p^{(1)}, p^{(2)}$ be two densities over $[0, 1]^d$ with $\log p^{(1)}, \log p^{(2)} \in \mathcal{F}_r$. If all the r dimensional marginals of $p^{(1)}$ and $p^{(2)}$ are the same, then $p^{(1)} = p^{(2)}$.*

Proof. For notational simplicity we write in the case of $r = 2$. The proof for general r is similar.

Since the two dimensional marginal densities of $p^{(1)}$ and $p^{(2)}$ are the same, we can denote them by $p_{jk}, 1 \leq j < k \leq d$.

For any $\eta \in \mathcal{F}_2$, we can write $\eta = \sum_{j < k} \eta_{jk}$, therefore

$$\begin{aligned} L_{\log p^{(1)}}(\eta) &= \int e^\eta - \int \eta p^{(1)} \\ &= \int e^\eta - \int \left(\sum_{j < k} \eta_{jk} \right) p^{(1)} \\ &= \int e^\eta - \sum_{j < k} \int \eta_{jk} p_{jk} \end{aligned}$$

By the same derivation we have $L_{\log p^{(2)}}(\eta) = \int e^\eta - \sum_{j < k} \int \eta_{jk} p_{jk}$. Therefore $L_{\log p^{(1)}}(\eta) = L_{\log p^{(2)}}(\eta)$ for any $\eta \in \mathcal{F}_2$. On the other hand, by Lemma 3.1 we know that $\log p^{(1)}$ is the unique minimizer of $L_{\log p^{(1)}}(\eta)$ over $\eta \in \mathcal{F}_2$, and that $\log p^{(2)}$ is the unique minimizer of $L_{\log p^{(2)}}(\eta)$ over $\eta \in \mathcal{F}_2$. Therefore $\log p^{(1)} = \log p^{(2)}$. The proof of the theorem is complete. \square

From Theorem 3.1 we see that the log density SS-ANOVA model of order r has the property that the r dimensional marginals uniquely determine the joint density function. The next theorem implies that the order r model is the largest log density SS-ANOVA model with such a property.

THEOREM 3.2. *Let $f \in \otimes_{j=1}^d H_j$ be any function not belonging to \mathcal{F}_r . Then for any density p satisfying $\log p \in \mathcal{F}_r$, there exists a density $\bar{p} \neq p$ of the form $\bar{p} = \exp(\bar{\eta} + \bar{\lambda}f)$ such that \bar{p} and p has the same r dimensional marginals. Here $\bar{\eta} \in \mathcal{F}_r$ and $\bar{\lambda} > 0$ is a scalar.*

For technical purpose, we cite below a lemma from Lin (2000) (Lemma 2.1), which will be used in the proof of Theorem 3.2:

LEMMA 3.2. Let H_j be the m -th order Sobolev Hilbert space of univariate functions of $x_j \in [0, 1]$ with $m \geq 1$. Let $\|\cdot\|$ be the norm in $\otimes_{j=1}^d H_j$. Then there exists a constant C depending only on m and d , such that $\sup_{\mathbf{x} \in [0,1]^d} |f(\mathbf{x})| \leq C\|f\|$, $\forall f \in \otimes_{j=1}^d H_j$.

Note that the norm in \mathcal{F}_r is the norm in $\otimes_{j=1}^d H_j$ restricted on the subspace \mathcal{F}_r , and will also be denoted as $\|\cdot\|$.

Proof of Theorem 3.2. Without loss of generality we write in the case of $r = 2$. Let p be a density function satisfying $\log p \in \mathcal{F}_r$.

For any fixed $\lambda \geq 0$, we can define a functional of $\eta \in \mathcal{F}_r$: $A_\lambda(\eta) = \int \exp(\eta + \lambda f) - \int \eta p$. Consider the problem of minimizing $A_\lambda(\eta)$ over \mathcal{F}_r . The first and second order derivatives of $A_\lambda(\eta)$ are $DA_\lambda(\eta)g = \int \exp(\eta + \lambda f)g - \int gp$ and $D^2A_\lambda(\eta)gh = \int \exp(\eta + \lambda f)gh$. Therefore $D^2A_\lambda(\eta)gg > 0$, $\forall g \neq 0$, and the minimization problem is strictly convex.

We will first show that (i) for some $\bar{\lambda} > 0$, the functional $A_{\bar{\lambda}}(\eta)$ has a unique minimizer $\bar{\eta} \in \mathcal{F}_r$. Then we show that (ii) $\bar{p} = \exp(\bar{\eta} + \bar{\lambda}f)$ has the same r dimensional marginals as those of p . That $\bar{p} \neq p$ is obvious since $\log \bar{p} = \bar{\eta} + \bar{\lambda}f \notin \mathcal{F}_r$ while $\log p \in \mathcal{F}_r$.

To prove (i), denote $a = \min_{\{\eta \in \mathcal{F}_r: \|\eta - \log p\| = 1\}} A_0(\eta)$ and $\delta = a - A_0(\log p)$. From the definition of the functional A_λ we see that $A_0(\eta) = L_{\log p}(\eta)$. Therefore from Lemma 3.1 $\log p$ is the unique minimizer of A_0 over \mathcal{F}_r , and hence $\delta > 0$.

From $f \in \otimes_{j=1}^d H_j$, we see that the function f is bounded. Let b be an upper bound of $|f|$. Then $A_\lambda(\log p) - A_0(\log p) = \int \exp(\log p + \lambda f) - \int \exp(\log p) \leq \exp(\lambda b) - 1 \rightarrow 0$ as $\lambda \rightarrow 0$. On the other hand, for any $\eta^1 \in \mathcal{F}_r$ such that $\|\eta^1 - \log p\| = 1$, we have $A_\lambda(\eta^1) - A_0(\eta^1) = \int \exp(\eta^1 + \lambda f) - \int \exp(\eta^1) \geq \int \exp(\eta^1)\{\exp(-\lambda b) - 1\} \geq \int \exp(\log p + C)\{\exp(-\lambda b) - 1\} = \exp(C)\{\exp(-\lambda b) - 1\} \rightarrow 0$ as $\lambda \rightarrow 0$. The last inequality follows from Lemma 3.2.

Therefore there exists a sufficiently small $\bar{\lambda} > 0$ not depending on η^1 , such that $A_{\bar{\lambda}}(\log p) - A_0(\log p) \leq \delta/2$ and $A_{\bar{\lambda}}(\eta^1) - A_0(\eta^1) \geq -\delta/2$. Thus, $A_{\bar{\lambda}}(\eta^1) \geq A_0(\eta^1) - \delta/2 \geq a - \delta/2 = A_0(\log p) + \delta/2 \geq A_{\bar{\lambda}}(\log p)$ for any $\eta^1 \in \mathcal{F}_r$ such that $\|\eta^1 - \log p\| = 1$. It then follows from the strict convexity of $A_{\bar{\lambda}}(\eta)$ that the minimizer of $A_{\bar{\lambda}}(\eta)$ exists and (i) is proved.

For (ii), denote the minimizer of $A_{\bar{\lambda}}(\eta)$ by $\bar{\eta} \in \mathcal{F}_r$. We show that $\bar{p} = \exp(\bar{\eta} + \bar{\lambda}f)$ is a density function and its $r = 2$ dimensional marginals \bar{p}_{jk} , $1 \leq j < k \leq d$ are the same as p_{jk} , $1 \leq j < k \leq d$, the r dimensional marginals of p .

The minimizer $\bar{\eta}$ satisfies $DA_{\bar{\lambda}}(\bar{\eta})g = 0, \forall g \in \mathcal{F}_r$, that is

$$\int g \bar{p} = \int gp, \quad \forall g \in \mathcal{F}_r. \quad (3.1)$$

Setting $g = 1 \in \mathcal{F}_r$ in (3.1), we get $\int \bar{p} = 1$, and therefore \bar{p} is a density function. Now for any $1 \leq j < k \leq d$, it follows from (3.1) that for any $g_{jk} \in H_j \otimes H_k \subset \mathcal{F}_r$, we have $\int_{[0,1]^2} g_{jk}(\bar{p}_{jk} - p_{jk}) = 0$. From this it is easy to see that $\bar{p}_{jk} = p_{jk}$ since the Fourier basis functions in $L^2([0,1]^2)$ all belong to $H_j \otimes H_k$. The theorem is proved. \square

4. DISCUSSIONS

The results in Section 3 can be generalized beyond order r models to general ANOVA structures. For example, the ANOVA model consists of main effects, the two way interactions η_{12} , η_{23} , η_{13} , η_{14} , η_{34} , and the three way interaction η_{134} is the largest log density ANOVA model under which the joint density is uniquely determined by the 12, 23, 134 marginals. Similar results for general log density SS-ANOVA models can be obtained with proofs that are virtually identical to those to the theorems in Section 3. We have chosen to present our results in the special case of order r ANOVA models only because the theorems and the proofs are simpler to state.

Results parallel to those in Section 3 also hold for log density functional ANOVA models other than the log density SS-ANOVA models. For example, in models discussed in the last paragraph of Section 2, these results can be proved with the proofs that are very similar to the proofs in Section 3.

REFERENCES

- EDWARDS, D. (2000). *Introduction to Graphical Modelling*. Springer-Verlag Inc.
- GU, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag Inc.
- GU, C. & WANG, J. (2003). Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica* **13**, 811–826.
- KE, C. & WANG, Y. (2004). Smoothing spline nonlinear nonparametric regression models. *Journal of the American Statistical Association* **99**, 1166–1175.
- LIN, Y. (2000). Tensor product space ANOVA models. *The Annals of Statistics* **28**, 734–755.
- SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics* **10**, 795–810.
- SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.
- TAPIA, R. A. & THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation*. Johns Hopkins University Press.
- WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.
- WAND, M. P. & JONES, M. C. (1995). *Kernel Smoothing*. Chapman & Hall.