Model diagnostics for smoothing spline ANOVA models

Chong GU

Key words and phrases: ANOVA decomposition; diagnostic; Kullback-Leibler projection; penalized likelihood estimate.


Abstract: The author proposes some simple diagnostics for the assessment of the necessity of selected model terms in smoothing spline ANOVA models; the elimination of practically insignificant terms generally enhances the interpretability of the estimates, and sometimes may also have inferential implications. The diagnostics are derived from Kullback-Leibler geometry, and are illustrated in the settings of regression, probability density estimation, and hazard rate estimation.

Model diagnostics for smoothing spline ANOVA models

Résumé : The author proposes some simple diagnostics for the assessment of the necessity of selected model terms in smoothing spline ANOVA models; the elimination of practically insignificant terms generally enhances the interpretability of the estimates, and sometimes may also have inferential implications. The diagnostics are derived from Kullback-Leibler geometry, and are illustrated in the settings of regression, probability density estimation, and hazard rate estimation.

1. INTRODUCTION

Penalized likelihood method estimates a function of interest, say \( \eta \), by the minimizer of a functional of the form

\[
L(\eta|\text{data}) + \lambda J(\eta)
\]

where \( L(\eta|\text{data}) \), usually taken as the minus log likelihood, measures the goodness-of-fit of \( \eta \) to the data, and \( J(\eta) \), usually taken as a quadratic roughness functional, measures the smoothness (regularity) of \( \eta \); the smoothing parameter \( \lambda \) controls the tradeoff between the two conflicting goals. Comprehensive treatments of penalized likelihood estimation can be found in, e.g., Wahba (1990) and Gu (2002).

On a product domain, say \( \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \), one may decompose the function as

\[
\eta(x) = \eta(x_1, x_2) = \eta_0 + \eta_1(x_1) + \eta_2(x_2) + \eta_{1,2}(x_1, x_2),
\]

where \( x = (x_1, x_2) \in \mathcal{X} \) with \( x_1 \in \mathcal{X}_1 \) and \( x_2 \in \mathcal{X}_2 \) and \( \eta_1, \eta_2, \) and \( \eta_{1,2} \) satisfy certain side conditions to ensure identifiability; extensions to higher dimensions are straightforward. Such decomposition extends the classical ANOVA decomposition on discrete domains to generic domains, and is simply referred to as (functional) ANOVA decomposition. ANOVA decompositions can be built into penalized likelihood estimation using the tensor product spline technique; see the references given above and Section 2.2. As with classical ANOVA models, regression models
consisting of main effects and selected lower order interactions are the most useful in practical applications.

For the estimation of a probability density \( f(x, y) \) of random variables \((X, Y)\) on domain \(X \times Y\), one may employ the logistic density transform \( f = e^{\eta} / \int_{X \times Y} e^{\eta} \) and estimate \( \eta(x, y) = \eta_x(x) + \eta_y(y) + \eta_{x,y}(x, y) \), where the constant is eliminated for a one-to-one transform. The absence of the interaction \( \eta_{x,y} \) characterizes the independence of the random variables \(X\) and \(Y\). More generally, the exclusions of selected interactions in higher dimensions may characterize various (conditional) independence structures among random variables; see, e.g., Gu (2002, Section 1.3.3) for a few more examples.

For the estimation of hazard rate based on censored lifetime data, one may employ the log transform and estimate the log hazard \( \eta(t, u) = \theta(t) + \lambda u + \eta_{t,u} \) as a function of time \(t\) and covariate \(u\); the absence of \(\eta_{t,u}\) characterizes a proportional hazard model.

The purpose of this article is to devise and illustrate some simple diagnostics for practically insignificant terms in an ANOVA decomposition estimated through the penalized likelihood method. The diagnostics are based on Kullback-Leibler geometry, and are implemented in open-source software and illustrated in the settings of regression, density estimation, and hazard estimation.

The rest of the article is organized as follows. In Section 2, detailed formulation of penalized likelihood estimation is briefly reviewed. The diagnostics are derived in Section 3 and illustrated in Section 4; Section 5 reports some supplementary simulation results. Brief discussion in Section 6 concludes the article. For the first reading, one may choose to skip to Section 3.1 for the basic idea of the diagnostics, then to Section 4 for application examples.

2. PENALIZED LIKELIHOOD ESTIMATION

The likelihood term \( L(\eta|\text{data}) \) in (1) reflects the sampling structure of the observed data, and the penalty term \( J(\eta) \) largely characterizes the model to be fitted to the data. Some details concerning the background formulations are given below.

2.1 Likelihood

First consider regression problems with data from exponential families,

\[ Y | x \sim \exp\{ (y \theta(\eta(x)) - b(\theta(\eta(x)))) / a(\phi) + c(y, \phi) \}, \]

where \( \eta \) is a monotone transform of the canonical parameter \( \theta \), taking values in \((-\infty, \infty)\), and \( a(\phi) \) is a common dispersion parameter; a unrestricted range of \( \eta \) prevents the complication of constrained minimization of (1). Observing \((x_i, Y_i), i = 1, \ldots, n\), one may take

\[ L(\eta|\text{data}) = -\frac{1}{n} \sum_{i=1}^{n} \{ Y_i \theta(\eta(x_i)) - b(\theta(\eta(x_i))) \}, \]

(2)

where the term \( c(y, \phi) \), independent of \( \eta \), is dropped and the dispersion \( a(\phi) \) is to be absorbed into the smoothing parameter \( \lambda \). Examples of this setting include least squares regression with \( \eta \) the normal mean, logistic regression with \( \eta \) the logit, Poisson regression with \( \eta \) the log intensity, etc.; see, e.g., Gu (2002, Section 5.4) for further details. The general formulation of penalized likelihood regression with exponential family responses was due to O’Sullivan, Yandell, & Raynor (1986). See, e.g., Gu (2002, Chapter 5) for a comprehensive treatment and further references. See also Gu & Kim (2002).

For probability density estimation on a bounded domain \(X\) with independent samples \(X_i, i = 1, \ldots, n\), the minus log likelihood is simply

\[ L(\eta|\text{data}) = -\frac{1}{n} \sum_{i=1}^{n} \left\{ \eta(X_i) - \log \int_X e^{\eta(x)} dx \right\}, \]

(3)
where the logistic density transform \( f = e^\eta / \int e^\eta \) is employed; a bounded domain permits a proper uniform distribution with \( \eta = 0 \). Penalized likelihood density estimation was pioneered by Good & Gaskins (1971); see also Silverman (1982) and Gu and Qiu (1993). A comprehensive treatment and further references can be found in Gu (2002, Chapter 6).

Let \( T \) be the lifetime of an item with survival function \( S(t|u) = P(T > t|u) \), possibly dependent on a covariate \( U \). The hazard function is given by \( e^{\eta(t,u)} = -\partial \log S(t|u)/\partial t \). Let \( Z \) be the left-truncation time and \( C \) be the right-censoring time, independent of \( T \) and of each other. Observing \((U_i, Z_i, X_i, \delta_i), \ i = 1, \ldots, n\), where \( X = \min(T,C) \), \( \delta = I_{[T \leq C]} \), and \( Z < X \), the minus log likelihood is seen to be

\[
L(\eta|\text{data}) = - \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i \eta(X_i, U_i) - \int_{Z_i}^{X_i} e^{\eta(t,U_i)} dt \right\}.
\]

The formulation through (4) was found in Gu (1996), which covers various earlier models as special cases; see, e.g., Gu (2002, Chapter 7) for a comprehensive treatment and further references.

When parametric models are assumed on the time axis, say in the case of accelerated life models with \( \tilde{T} = \log T \) following a location-scale family distribution with density \( f((\tilde{t} - \mu)/\sigma) \), the task is to estimate a parameter of the parametric model as a function of the covariate \( u \); it is essentially a regression problem. For example, for \( T \) Weibull, the hazard is given by \((\alpha/\tilde{t})(\tilde{t}/\beta)^{\alpha - 1}/\sigma \tilde{t}\) where \( \alpha = 1/\sigma \) is the shape parameter and \( \beta = e^{\mu} \) is the scale parameter, and the minus log likelihood, upon setting \( \eta = \mu \), is seen to be

\[
L(\eta|\text{data}) = - \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i (\alpha (\log X_i - \eta(U_i)) + \log \alpha) - (X_i^\alpha - Z_i^\alpha) e^{-\alpha(U_i)} \right\}.
\]

Other examples of accelerated life models include the log normal and the log logistic distributions; see, e.g., Gu (2002, Section 7.5) for further details.

2.2 Penalty

We now discuss the construction of \( J(\eta) \). To keep the notation simple, the domain of \( \eta \) is to be denoted by \( \mathcal{X} \) regardless of the stochastic setting.

The minimization of (1) is in a space \( \mathcal{H} \subseteq \{ \eta : J(\eta) < \infty \} \) in which \( J(\eta) \) is a square seminorm. The evaluation functional \([x] = \eta(x)\) appears in the term \( L(\eta|\text{data}) \), and is assumed to be continuous in \( \mathcal{H} \). A space \( \mathcal{H} \) in which the evaluation is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a reproducing kernel (RK) \( k(\cdot, \cdot) \), a non-negative definite function satisfying \( R_k(\cdot) = R(x, \cdot) \in \mathcal{H}, \forall x \in \mathcal{X} \), and \( \langle R(x, \cdot), f(\cdot) \rangle = f(x), \forall f \in \mathcal{H} \), where \( \langle \cdot, \cdot \rangle \) is the inner product in \( \mathcal{H} \); the RK \( R(\cdot, \cdot) \) and the space \( \langle \cdot, \cdot \rangle \) determine each other uniquely. Typically, \( J(\cdot, \cdot) = J(\cdot) + J(\cdot) \) is the inner product in the null space \( \mathcal{N}_J = \{ \eta : J(\eta) = 0 \} \) of \( J(\eta) \) when restricted therein. There exists a tensor sum decomposition \( \mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J \), where the space \( \mathcal{H}_J \) has \( J(\eta) \) as its square norm and an RK \( R_J \) satisfying \( J(R_J(x, \cdot), f(\cdot)) = f(x), \forall f \in \mathcal{H}_J \). See, e.g., Gu (2002, Section 2.1).

An example on \( \mathcal{X} = [0,1] \) is \( J(\eta) = \int_0^1 \eta^2 dx \) with \( \mathcal{N}_J = \{ \eta : \eta(x) = \beta_0 + \beta_1 x \} \); this yields the popular cubic splines. A choice of \( J(f, g) \) is \( \int_0^1 f dx (\int_0^1 g dx) + (\int_0^1 f dx)(\int_0^1 g dx) \), yielding \( \mathcal{H}_J = \{ \eta : \int_0^1 \eta dx = \int_0^1 \eta^2 dx = 0, J(\eta) < \infty \} \) and the RK \( R_J(x, y) = k_2(x)k_2(y) - k_4(x - y), \) where \( k_4 = B_{\nu}/\nu! \) are scaled Bernoulli polynomials. See, e.g., Gu (2002, Section 2.3.3).

For \( \mathcal{X} \) a product domain, tensor product RKHS can be constructed from RKHS on marginal domains, and ANOVA decompositions can be induced from one-way decompositions on marginal domains; the construction of \( \mathcal{H}_J \) is indirect through that of the RK \( R_J \). The construction can be denoted by \( \mathcal{H} = \oplus_{\beta=0}^{\infty} \mathcal{H}_J \) and \( J(\eta) = \sum_{\beta=0}^{\infty} \theta_\beta J_\beta(\eta_\beta) \), where \( \eta_\beta \in \mathcal{H}_\beta \), \( 0 < \theta_\beta < \infty \), and \( J_\beta(\cdot) \) is the square norm in \( \mathcal{H}_\beta \), \( \beta > 0 \). One has \( \mathcal{N}_J = \mathcal{H}_0 \), \( \mathcal{H}_J = \oplus_{\beta=1}^{\infty} \mathcal{H}_\beta \), and \( R_J = \sum_{\beta=1}^{\infty} \theta_\beta R_\beta \), where
$R_\beta$ is the RK in $\mathcal{H}_\beta$. The $\theta_\beta$’s are an extra set of smoothing parameters, often suppressed in the notation, that adjust the relative penalties on the roughness of different components. See, e.g., Gu (2002, Section 2.4).

As an example, consider the tensor product cubic splines on $\mathcal{X} = [0,1]^2$. On the marginal domain $[0,1]$, one has

$$\{\eta : \int_0^1 \bar{\eta}^2 \, dx < \infty \} = \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1$$

$$= \text{span}\{1\} \oplus \text{span}\{k_1(x)\} \oplus \{\eta : \int_0^1 \eta \, dx = \int_0^1 \bar{\eta} \, dx = 0, \int_0^1 \bar{\eta}^2 \, dx < \infty \},$$

where $k_1(x) = x - 0.5$, with RKs $R_{00}(x,y) = 1$, $R_{01}(x,y) = k_1(x)k_1(y)$, and $R_1 = k_2(x)k_2(y) - k_4(x-y)$; an one-way ANOVA is built in with $f \in \mathcal{H}_{01} \oplus \mathcal{H}_1$ satisfying the side condition $\int_0^1 f \, dx = 0$. Taking tensor product, one has 9 tensor sum terms $\mathcal{H}_{\nu,\mu} = \mathcal{H}_{\nu(1)} \otimes \mathcal{H}_{\mu(2)}$ on $\mathcal{X} = [0,1]^2$, $\nu, \mu = 00, 01, 1$, with RKs $R_{\nu,\mu}(x,y) = R_{\nu(1)}(x(1),y(1))R_{\mu(2)}(x(2),y(2))$. The 4 subspaces with $\nu, \mu = 00, 01$ are of one-dimension each, and can be lumped together as $\mathcal{H}_0 = \mathcal{N}_j$. The other 5 subspaces can be indexed as $\mathcal{H}_\beta$, $\beta = 1, \ldots, 5$. For the ANOVA decomposition, $\mathcal{H}_{00,00}$ contains the constant, $\mathcal{H}_{01,00} \oplus \mathcal{H}_{1,00}$ contains the $x_{(1)}$ main effect, $\mathcal{H}_{00,01} \oplus \mathcal{H}_{00,1}$ contains the $x_{(2)}$ main effect, and the other 4 contain the interaction. Excluding the 4 subspaces corresponding to the interaction, one gets an additive model. See, e.g., Gu (2002, Example 2.8).

2.3 Computation

When the term $L(\eta|\text{data})$ involves only function evaluations $[x_i]\eta = \eta(x_i)$ such as in (2) and (5), it can be shown that the minimizer of (1) has an expression $\eta(x) = \sum_{\nu=1}^m d_\nu \phi_\nu(x) + \sum_{j=1}^q c_j R_j(x, x)$, where $\{\phi_\nu\}$ is a basis of $\mathcal{N}_j$. When $L(\eta|\text{data})$ involves further terms such as the integrals in (3) and (4), the exact minimizer of (1) does not appear to be computable. In either case, it can be shown that the minimizer of (1) in a space $\mathcal{N}_j \oplus \text{span}\{R_j(z_j)\}$, $j = 1, \ldots, q$ shares the same asymptotic convergence rates as the minimizer in $\mathcal{H}$, where $\{z_j\}$ is a random subset of $\{x_i\}$ and $q$ can be as small as $O(n^{2/3})$ for (tensor product) cubic splines; see, e.g., Gu & Wang (2003) and Gu & Kim (2002). Plugging into (1) the expression

$$\eta(x) = \sum_{\nu=1}^m d_\nu \phi_\nu(x) + \sum_{j=1}^q c_j \left( \sum_{\beta=1}^q \theta_\beta R_\beta(z_j, x) \right)$$

and fixing $\theta_\beta$ and $\lambda$, one solves for the coefficients $d_\nu$ and $c_j$. The ANOVA terms are easily extracted by regrouping the terms of (6). For the selection of the smoothing parameters $\theta_\beta$ and $\lambda$, various cross-validation scores have been developed to target the Kullback-Leibler losses derived in the various stochastic settings; see, e.g., Craven & Wahba (1979), Xiang & Wahba (1996), and Gu & Wahba (1996).

Most of the regression, density and hazard models discussed here are implemented in the R package gss by the author; simple illustrations of the usage of gss can be found in the examples of Section 4.

3. DIAGNOSTICS

We are now ready to derive the diagnostics in the settings of regression, density estimation, and hazard estimation. We start with a discussion of the general method.

3.1 General method

Suppose the estimation of $\eta$ has been done in a space $\mathcal{H}$, but in fact $\eta \in \mathcal{H}^* \subset \mathcal{H}$. The task is to devise tools to assess the plausibility of the null hypothesis that $\eta \in \mathcal{H}^*$. In a parametric analysis,
the likelihood ratio tests or the like serve as the primary tool for the purpose. In nonparametric settings, the null $H^*$ is typically infinite-dimensional, and consequently the sampling distributions of likelihood ratio statistics are no longer available, hence one has to look elsewhere.

We propose some heuristic diagnostics based on Kullback-Leibler geometry; the definitions of Kullback-Leibler distance $KL(^\cdot,^\cdot)$ between functions in various stochastic settings will be given in Sections 3.2-3.4. Denote by $\hat{\eta}$ the estimate of $\eta$ obtained by minimizing (1) in $H$, with the smoothing parameters selected by cross-validation. Let $\tilde{\eta}$ be the Kullback-Leibler projection of $\hat{\eta}$ in $H^*$, the minimizer of $KL(\hat{\eta}, \eta)$ for $\eta \in H^*$, and let $\eta_c$ be the constant model; $\eta_c$ is the maximum likelihood estimate of $\eta = \eta_0$ for the regressions of (2) and (5) and for the hazard estimation of (4), and is $\eta = 0$ for the density estimation of (3). As will be shown shortly, it holds in many settings, and approximately so in others, that

$$KL(\hat{\eta}, \eta_c) = KL(\hat{\eta}, \tilde{\eta}) + KL(\tilde{\eta}, \eta_c),$$

(7)

where $KL(\hat{\eta}, \eta_c)$ is the “total entropy” of $\hat{\eta}$ and $KL(\tilde{\eta}, \eta_c)$ is the “preserved entropy” by the null $H^*$; $KL(\hat{\eta}, \eta_c)$ quantifies the amount of structure there is in $\eta$ and the ratio $1 - \rho = KL(\hat{\eta}, \eta_c)/KL(\tilde{\eta}, \eta_c)$ depicts how much of that structure actually sits in the null $H^*$. We shall use the ratio $\rho = KL(\tilde{\eta}, \eta)/KL(\tilde{\eta}, \eta_c)$ as a model diagnostic; a small value of $\rho$ indicates the lack of necessity for the extra complexity beyond the null $H^*$.

To perceive such a geometric inferential tool in contrast to the classical hypothesis testing, let us consider a standard linear model $Y = \mathbf{1}\beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$ with the null $H^* = \{\beta_2 = 0\}$, where $X_1$ and $X_2$ are matrices and $\beta_1$ and $\beta_2$ vectors. The definition of (8) below yields $KL(\tilde{\eta}, \eta) = n^{-1}\sum_{i=1}^n(\tilde{\eta}(x_i) - \eta(x_i))^2$ for Gaussian regression, so one has $KL(\hat{\eta}, \eta_0) = n^{-1}\sum_{i=1}^n(\hat{Y}_i - \tilde{Y})^2$ and $KL(\tilde{\eta}, \eta_0) = n^{-1}\sum_{i=1}^n(\tilde{Y}_i - \tilde{Y})^2$, where $\tilde{Y} = \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}$ with $\tilde{X} = (1, X_1, X_2)$ and $\tilde{Y} = \tilde{X}_1(\tilde{X}_1^T\tilde{X}_1)^{-1}\tilde{X}_1^T\tilde{Y}$ with $\tilde{X}_1 = (1, X_1)$. It follows that

$$\rho = \frac{\sum_{i=1}^n(\hat{Y}_i - \tilde{Y})^2 - \sum_{i=1}^n(\tilde{Y}_i - \tilde{Y})^2}{\sum_{i=1}^n(\hat{Y}_i - \tilde{Y})^2} = \frac{\sum_{i=1}^n(\hat{Y}_i - \tilde{Y})^2}{\sum_{i=1}^n(\hat{Y}_i - \hat{Y})^2} = \frac{SSR(X_2|X_1)}{SSR(X_1, X_2)},$$

with $X_1$ and $X_2$ indicating groups of predictors; note that neither the variance of $\epsilon$ nor the sample size is referenced in $\rho$. When $\rho = 0.02$, one may well feel comfortable settling with $\beta_2 = 0$, although $\beta_2$ could be statistically significant due to a small error variance or a large sample size. On the other hand, with $\rho = 0.10$ as the sole clue, one would most likely keep $\beta_2$ in the model, but $\beta_2$ could well be statistically insignificant with a large error variance or a small sample size.

To an extent, the $\rho$ ratios are like $p$-values for tests and their calibration is like the selection of significance levels for tests, except that geometry replaces error probability as the logical basis of inference. In particular, the routine practice in a testing situation, that of aligning the test statistics on the common scale of $p$-values through null distributions, appears unnecessary for the $\rho$ ratios.

### 3.2 Exponential family regression

For regression with exponential family responses as in (2), one has

$$KL(\hat{\eta}, \eta) = \frac{1}{n}\sum_{i=1}^n\{\mu(x_i)(\theta(\hat{\eta}(x_i)) - \theta(\eta(x_i))) \cdot (\hat{\theta}(\hat{\eta}(x_i)) - \hat{\theta}(\eta(x_i))) \},$$

(8)

where $\mu(x) = (db/d\theta)(x) = E[Y|x]$; see, e.g., Gu (2002, Section 5.2.1). Substituting in the expression (6) with $m$ and $g$ the versions associated with $H^*$, the minimization of (8) can be solved via Newton iteration. The resulting $KL(\hat{\eta}, \tilde{\eta})$ depends on the $\theta_\beta$’s hidden in (6), and an outer loop of optimization shall be performed to minimize $KL(\hat{\eta}, \tilde{\eta})$ with respect to the $\theta_\beta$’s.
Note that after substituting in (6) one is no longer projecting into \( \mathcal{H}^* \) but only into some \((m + q)\)-dimensional subspace, indexed by \( \theta_\beta \), therein. The outer loop optimization serves to locate the nearest (to \( \tilde{\eta} \)) of such \((m + q)\)-dimensional subspaces.

Setting \( \eta = \tilde{\eta} + \alpha(\tilde{\eta} - \eta_c) \) in (8) for \( \alpha \) real, differentiating with respect to \( \alpha \), and evaluating at \( \alpha = 0 \), one has

\[
\frac{1}{n} \sum_{i=1}^{n} (\tilde{\mu}(x_i) - \tilde{\mu}(x_i)) h(x_i)(\tilde{\eta}(x_i) - \eta_c(x_i)) = 0,
\]

(9)

where \( h = (d\theta/d\eta)|_{\tilde{\eta}} \); note that \( \tilde{\eta} \) minimizes (8) in \( \mathcal{H}^* \) and \( \tilde{\eta} - \eta_c \in \mathcal{H}^* \). It is easy to verify that

\[
\text{KL}(\tilde{\eta}, \eta_c) = \text{KL}(\tilde{\eta}, \tilde{\eta}) + \text{KL}(\tilde{\eta}, \eta_c) + \frac{1}{n} \sum_{i=1}^{n} (\tilde{\mu}(x_i) - \tilde{\mu}(x_i))(\theta(\tilde{\eta}(x_i)) - \theta(\eta_c(x_i)))
\]

(10)

for the Kullback-Leibler distance defined in (8). When \( \eta = \theta \) is the canonical parameter, (9) and (10) yields (7); this covers the cases of Gaussian regression, logistic regression, and Poisson regression. When \( \eta \neq \theta \), (7) may hold approximately depending on how accurate the first order approximation, \((\tilde{\mu} - \mu)(\theta(\tilde{\eta}) - \theta(\eta_c)) \approx (\tilde{\mu} - \mu)\tilde{h}(\tilde{\eta} - \eta_c) \), is.

3.3 Density estimation

For density estimation as in (3), one has

\[
\text{KL}(\tilde{\eta}, \eta) = \mu_{\tilde{\eta}}(\tilde{\eta} - \eta) - \log \int_{X} e^{\tilde{\eta}(x)} dx + \log \int_{X} e^{\eta(x)} dx,
\]

(11)

where \( \mu_{\tilde{\eta}}(h) = \int_{X} h(x)e^{\tilde{\eta}(x)} dx / \int_{X} e^{\tilde{\eta}(x)} dx \); see, e.g., Gu (2002, Section 6.3.1). One can again substitute the expression of (6) into (11) and perform an inner loop Newton iteration nested under an outer loop optimization with respect to the \( \theta_\beta \)'s. Similar to (9), one has \( \mu_{\tilde{\eta}}(\tilde{\eta} - \eta_c) = \mu_{\tilde{\eta}}(\tilde{\eta} - \eta_c) \), which yields (7) for the Kullback-Leibler distance defined in (11).

3.4 Hazard estimation

For hazard estimation as in (4), one has

\[
\text{KL}(\tilde{\eta}, \eta) = \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \left\{ e^{\tilde{\eta}(t,U_i)}(\tilde{\eta}(t,U_i) - \eta(t,U_i)) - (e^{\tilde{\eta}(t,U_i)} - e^{\eta(t,U_i)}) \right\} dt;
\]

(12)

see, e.g., Gu (2002, Section 7.2.1). The same procedure can be followed to locate \( \tilde{\eta} \) that minimizes (12) in the nearest \((m + q)\)-dimensional subspace characterized by (6). Similar to (9), one has

\[
\frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \left( e^{\tilde{\eta}(t,U_i)} - e^{\eta(t,U_i)} \right)(\tilde{\eta}(t,U_i) - \eta_c(t,U_i)) dt = 0,
\]

which yields (7) for the Kullback-Leibler distance defined in (12).

For Weibull regression as in (5), with the hazard \((\alpha/t)e^{\alpha \log t - \eta(U)}\) parametric in \( t \) and the location parameter \( \eta(U) \) of \( \log T \) as a function of the covariate \( U \), (12) becomes

\[
\text{KL}(\tilde{\eta}, \eta) = \frac{1}{n} \sum_{i=1}^{n} (X_i^\alpha - Z_i^\alpha) \left\{ -e^{-\alpha \tilde{\eta}(U_i)} - e^{-\alpha \eta(U_i)} - \alpha e^{-\alpha \tilde{\eta}(U_i)}(\tilde{\eta}(U_i) - \eta(U_i)) \right\}.
\]

(13)
Similar to (9), one has
\[
\frac{1}{n} \sum_{i=1}^{n} (X_i - Z_i)(e^{-\alpha \eta(U_i)} - e^{-\alpha \hat{\eta}(U_i)})(\check{\eta}(U_i) - \eta(U_i)) = 0,
\]
which yields (7) for the Kullback-Leibler distance given in (13).

4. EXAMPLES

We now illustrate the diagnostics through a few examples. The computation can be done using the R package \texttt{gss}. R resources are archived at \texttt{cran.r-project.org}, where the source code of base R and that of scores of add-on packages can be found along with installation instructions.

4.1 AIDS incubation

To study the AIDS incubation time, a valuable source of information is in the records of patients who were infected with the HIV virus through blood transfusion, of which the date can be ascertained retrospectively. A data set collected by the Centers for Disease Control and Prevention (CDC) is listed in Wang (1989), which includes the time \(X\) from the transfusion to the diagnosis of AIDS, and the time \(Y\) from the transfusion to the end of study (July 1986), both in months, for 295 individuals. It is clear that \(X \geq Y\), i.e., the data are truncated.

The joint density of \((X, Y)\) on the truncated domain \(\{x \leq y\}\) can be written as
\[
f(x, y) = \frac{e^{\eta(x,y)}}{\int_{x \leq y} e^{\eta(x,y)}},\]
where \(\eta(x,y) = \eta_x(x) + \eta_y(y) + \eta_{x,y}(x,y)\). The observed \(X\) and \(Y\) are dependent on each other due to the truncation, but the exclusion of \(\eta_{x,y}(x,y)\) implies the pre-truncation independence of the incubation time and the infection time. Assuming pre-truncation independence, the analysis of the data was showcased in Gu (2002, Sections 1.4.2, 6.5.3, 6.6.4), and our task here is to assess the adequacy of this assumption using the diagnostics of Section 3.

With the data in an R data frame \texttt{aids} with components \texttt{incu} and \texttt{infe}, a tensor product cubic spline model can be fitted using the following commands.

```r
aids.fit <- ssden(~incu*infe, data=aids, quad=list(pt=quad.pt,wt=quad.wt),
                 domain=data.frame(incu=c(0,100),infe=c(0,100)), seed=2375)
```

The \texttt{seed} option was needed for the results to be reproducible; note that \(z_j\) in (6) is a random subset. The quadrature in \texttt{quad.pt} and \texttt{quad.wt} is a 40 \times 40 uniform grid truncated to \(\{x \leq y\}\), with half weights on the diagonal; the code for the generation of \texttt{quad.pt} and \texttt{quad.wt} can be found in the documentation of \texttt{ssden}.

The model has three terms, \texttt{"incu"}, \texttt{"infe"}, and \texttt{"incu:infe"}. To check the practical significance of the term \texttt{"incu:infe"}, use

```r
project(aids.fit, include=c("incu","infe"))
```

which yields \(\rho = \text{KL}(\hat{\eta}, \check{\eta})/\text{KL}(\check{\eta}, \eta_c) = 0.0166\), i.e., more than 98% of the structure in \(\hat{\eta}\) is preserved in \(\check{\eta}\). This suggests the adequacy of pre-truncation independence.

4.2 Progression of diabetic retinopathy

The Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) is an on-going epidemiological study of a cohort of patients receiving their medical care in an 11-county area in southern Wisconsin, who were first examined in 1980-1982, then again in 1984-1986, 1990-1992, and 1994-1996. A subset derived from the WESDR data is distributed in GRKPACK (Wang 1997), consisting of the baseline measures of duration of diabetes in years, percent of glycosylated hemoglobin, body mass index, and a binary indicator of retinopathy progression at the first follow-up, of 669 patients.
Let \( p(x) = p(x_1, x_2, x_3) \) be the probability of retinopathy progression at the first follow-up given the three covariates. Logistic models of the form \( \log \{ p(x)/(1 - p(x)) \} = \eta(x) \) are considered, where \( \eta(x) \) can be decomposed into the sum of a constant, three main effects, three two-way interactions, and a three-way interaction. An analysis of the data can be found in Gu (2002, Section 5.5.3), where a cubic spline additive model for \( \eta(x) \) was fitted and presented. We now illustrate how the diagnostics of Section 3 can be used to pick the additive model.

With the data in an R data frame \( \texttt{wesdr} \) with components \( \texttt{dur}, \texttt{gly}, \texttt{bmi}, \texttt{ret} \), a tensor product linear spline model can be fitted and the additivity assessed.

\[
\texttt{wesdr.fit0} \leftarrow \texttt{gssanova1(ret~dur*gly*bmi, family="binomial", type="linear", data=wesdr, nbasis=135, seed=2375)}
\]

\[
\texttt{project(wesdr.fit0, include=c("dur","gly","bmi"))}
\]

The Kullback-Leibler projection yields \( \rho = \text{KL}(\hat{\eta}, \tilde{\eta}) = 0 \), indicating that a proportional hazard projection would lose 12% of the structure in \( \hat{\eta} \). The time axis transformation does not affect hazard proportionality or the lack of it.

4.4 Treatments of gastric cancer

The survival times of 90 gastric cancer patients are listed in Moreau, O’Quigley, & Mesbah (1985). Half of the patients were treated by chemotherapy, the other half by chemotherapy combined with radiotherapy. The data are right-censored with no left-truncation.
We again consider a log hazard model of the form $\eta(t, u) = \eta_0 + \eta_t(t) + \eta_u(u) + \eta_{tu}(t, u)$ but with a binary covariate $u \in \{1, 2\}$, where $\eta_0(1) = -\eta_0(2)$ is a constant, $\eta_t(t)$ is the mean log hazard of the two treatments, and $\eta_{tu}(t, 1) = -\eta_{tu}(t, 2)$ is the contrast log hazard; the model was fitted and presented in Gu (2002, Section 7.3.1). The task here is again to assess the hazard proportionality.

With the data in an R data frame `gastric` with components `futime`, `status`, and `trt`, the hazard model can be fitted and evaluated as follows; `trt` is a factor variable.

```r
gastric.fit <- sshzd(Surv(futime,status)~futime*trt, data=gastric, seed=2375)
project(gastric.fit, include=c("futime","trt"))
```

The Kullback-Leibler projection yields $\rho = \text{KL}(\hat{\eta}, \hat{\eta})/\text{KL}(\hat{\eta}, \hat{\eta}_c) = 0.7307$, suggesting that hazard proportionality is strongly violated. Moreau et al. (1985) reported a $p$-value between 0.01 and 0.02 using the goodness-of-fit statistic they developed for the testing of hazard proportionality.

5. SIMULATIONS

To gain some perspective concerning the calibration of the diagnostics, simple simulations of limited scales are presented in this section. The purpose of the simulations is to seek practical insights, but not quite to establish universal decision rules.

5.1 Density estimation

Samples $X_i$, $i = 1, \ldots, n$, were generated from bivariate normal distributions with mean 0, unit variance, and correlation $r$. For each of the six combinations of sample size $n = 100, 200$ and correlation $r = 0, 0.3, 0.5$, twenty replicates were generated. For each replicate, the domain $X$ is rectangular with the marginals extending 5% on each end beyond the data ranges. The domain is mapped to $[0, 1]^2$ and the tensor product cubic spline of Section 2.2 is used in (3). Five different random subsets $\{z_j\} \subset \{X_i\}$ were used to calculate five different fits with interaction, and Kullback-Leibler projections into the additive model space were calculated. The sizes of $\{z_j\}$ were $q = 30$ for $n = 100$ and $q = 33$ for $n = 200$.

The simulation results are summarized in Figure 1. The left frames illustrate the boxplots of $\rho = \text{KL}(\hat{\eta}, \hat{\eta})/\text{KL}(\hat{\eta}, \hat{\eta}_c)$ (of 5 fits) for each of the $6 \times 20$ replicates. The right frames plot the median $\rho$ against the sample correlation of each replicate. It appears that $n = 200$ is much less variable than $n = 100$; it is easier to generate a “bad sample” with a smaller sample size. Threshold values around $\rho = 0.02, 0.03$ seem to provide adequate calibration.

5.2 Logistic regression

For $x_i$ from $U(0, 1)^2$, $i = 1, \ldots, n$, binomial responses were generated through $Y_i \sim \text{Bin}(2, p(x_i))$, where $p(x) = e^{\eta(x)}/(1 + e^{\eta(x)})$ and

$$\eta(x) = 0.3(10^6 x_{(1)}^{11} (1 - x_{(1)})^6 + 10^4 x_{(2)}^{3} (1 - x_{(1)})^{10}) - 2 + b \sin(2\pi x_{(2)}),$$

with $b$ a constant to be specified. For each of the four combinations of sample size $n = 100, 200$ and $b = 0, 1$, twenty replicates were generated. For each replicate, additive cubic spline logistic regression fits were calculated using 5 different subsets $\{z_j\} \subset \{x_i\}$, and Kullback-Leibler projections into the space of univariate functions of $x_{(1)}$ were calculated. The minus log likelihood of (2) is given by

$$L(\eta|\text{data}) = -\frac{1}{n} \sum_{i=1}^{n} \{Y_i \eta(x_i) - 2 \log(1 + e^{\eta(x_i)})\}.$$
Figure 1: Density Estimation Simulation. Top: \( n = 100 \). Bottom: \( n = 200 \). The twenty boxplots corresponding to the twenty replicates for each combination of \( n \) and \( r \) are ordered by the median \( \text{KL}(\hat{\eta}, \bar{\eta})/\text{KL}(\hat{\eta}, \eta_c) \) of 5 fits; the faded horizontal lines represent \( \text{KL}(\hat{\eta}, \bar{\eta})/\text{KL}(\hat{\eta}, \eta_c) = 0.02, 0.03 \).

The sizes of \( \{z_j\} \) were again \( q = 30 \) for \( n = 100 \) and \( q = 33 \) for \( n = 200 \). The simulation results are summarized in Figure 2.

5.3 Hazard estimation

Consider two treatments with hazards

\[ e^{\phi(t,1)} = 1 + 12t^2, \quad e^{\phi(t,2)} = (1 + 12t^2)(1.5 + b(t - .65)). \]

Right censored lifetime data \( X_i = \min(T_i, C_i), \delta_i = I[T_i < C_i], \) \( i = 1, \ldots, n, \) were generated, with \( T_i \) having the given hazards and \( C_i \) following a censoring distribution \( P(C > c) = e^{-2t^3/3}. \) For each of the four combinations of sample size \( n = 100, 200 \) and \( b = 0, 2, \) twenty replicates were generated, with a half-half split of the two treatments \( U = 1, 2; \) the censoring rates were between 14.5% and 31.5%, with the mean around 21%. For each replicate, log hazard models with interaction were fitted using 5 different subsets \( \{z_j\} \subset \{(X_i, U_i)\}, \) and Kullback-Leibler projections into the space of additive models were calculated. The sizes of \( \{z_j\} \) were again \( q = 30 \) for \( n = 100 \) and \( q = 33 \) for \( n = 200 \).

The simulation results are summarized in the top frames of Figure 3; the diagnostic appear much less reliable. A possible explanation might be that hazard estimation is a more challenging problem so the estimation precision is relatively low. To check on the plausibility of the explanation, we repeated the simulation using the same lifetime observations \( T_i \) but with \( C_i = \infty; \) intuitively the estimation precision should be better in the absence of censoring. The results of the censoring-free simulation are summarized in the bottom frames of Figure 3. Relative to the censored cases,
the ratio \( \rho = \frac{KL(\hat{\eta}, \tilde{\eta})}{KL(\hat{\eta}, \eta_c)} \) appears lower for \( b = 0 \) and higher for \( b = 2 \), both to the right directions. The diagnostic seems to be rather reliable in the censoring-free \( n = 200 \) experiments.

6. DISCUSSION

In this article, we have introduced simple diagnostics for the assessment of the necessity of model terms in smoothing spline ANOVA models. The diagnostics are illustrated using real-data examples, and some quantitative perspective is obtained through limited scale simulations. The techniques are implemented in open-source software for ready use by practitioners.

Casting things as mixed-effect models through the Bayes model of penalized least squares regression, the elimination of model terms can be restated as the elimination of variance components, for which the likelihood ratio tests may be employed. The catch here is that the null is on the boundary of the parameter space, so the usual \( \chi^2 \) null distribution for the log likelihood ratio no longer holds. Guo (2002) derived the null distribution as a mixture of \( \chi^2 \)'s and discussed various other aspects of the test; it however remains unclear how this approach may be extended to settings other than Gaussian regression. Tests for additivity were also derived by Barry (1993) and Eubank, Hart, Simpson, & Stefanski (1995), for Gaussian regression assuming sampling points on product grids; the “full factorial design” makes additivity a less elusive infinite dimensional null, but the resulting methods are not as useful in practice.

The diagnostics are designed for the “testing” of possibly infinite dimensional nulls in function estimation settings, but their operational characteristics are different from classical tests. For example, the “rejection rate” may not necessarily increase with the sample size, which can be advantageous, but on the other hand, the procedure relies on the decomposition of the “total entropy” \( KL(\hat{\eta}, \eta_c) \), which could be problematic when the overall “signal” is weak. When the overall “signal” is weak, however, a more relevant problem is the assessment of the overall model rather than part of it.

Classical tests assume unambiguous, prospective characterizations of the “null behavior” of the data in the form of sampling distributions for test statistics, which however are generally unavailable for infinite dimensional nulls in nonparametric settings. The diagnostics introduced here are retrospective in nature, philosophically similar to the diagnostics developed for regression in Gu (1992), where further discussions concerning the contrasts between prospective and retrospective analyses can be found.

The calibration of the diagnostics is a rather subjective issue, but the simulations seem to suggest that a reasonable threshold for \( KL(\hat{\eta}, \tilde{\eta})/KL(\hat{\eta}, \eta_c) \) could be in the range of 0.02 ~ 0.03 for
density estimation and regression. Hazard estimation with censored survival data appears to be a more difficult problem with weaker “signal”, and the diagnostic appears less reliable in the context. Because the general method covers a broad spectrum of settings, it appears impractical to establish universal decision rules via simulations, but customized simulations in targeted applications could prove helpful.

ACKNOWLEDGEMENTS

The author acknowledges the assistance of Jingyuan Wang in the early stage of the work. This research was supported by U.S. National Institutes of Health under grant R33HL68515.

REFERENCES


Received ???
Accepted ???