A Volume-of-tube based Test for Penalized Splines Estimators

Wiesenfarth, Manuel  
*University of Goettingen, CRC Poverty, Equity & Growth in Developing Countries*  
Wilhelm-Weber-Str. 2  
37073 Göttingen, Germany  
E-mail: mwiesen@gwdg.de

Krivobokova, Tatyana  
*University of Goettingen, CRC Poverty, Equity & Growth in Developing Countries*  
Wilhelm-Weber-Str. 2  
37073 Göttingen, Germany  
E-mail: tkrivob@gwdg.de

Sperlich, Stefan  
*University of Geneva, Department of Econometrics*  
40 Bd du Pont d’Arve  
CH-1211 Genve 4, Suisse  
E-mail: stefan.sperlich@unige.ch

Introduction

We propose a simple and fast approach to testing polynomial regression versus a general non-parametric alternative modeled by penalized splines. For the construction of the test we exploit novel results on simultaneous confidence bands using the approximation to the tail probability of maxima of Gaussian processes by the volume-of-tube formula (see Krivobokova et al., 2010, and Sun, 1993). Besides allowing for the incorporation of smooth curves that enter an additive model, are spatially heterogeneous (see Krivobokova et al., 2008) and are estimated from heteroscedastic data, the test can also be used for investigating the statistical significance of certain features in a curve, such as dips and bumps. Further advantages include very good small sample properties and the analytical availability, i.e. no computationally intensive procedures such as bootstrapping (as in Härdle et al. (2004), for example) are necessary and results are obtained virtually instantly. In particular, this test is preferable to F-type tests (for example as used in R package *mgcv*, Wood, 2006), which tend to underestimate p-values when smoothing parameters are estimated. In simulations we show that the proposed test performs competitively compared to restricted likelihood ratio tests (RLRT, see Crainiceanu et al., 2005) and thus provides a convenient alternative. The method is implemented in the R package *AdaptFitOS*, making it readily available for practitioners. For the related simultaneous confidence bands, see Wiesenfarth et al. (2010).

Estimation with Penalized Splines

We consider the model

\[ Y_i = \beta_0 + \sum_{j=1}^{d} f_j(x_{ji}) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2(\tilde{x}_i)), \quad i = 1, \ldots, n, \]

where \( \beta_0 \) is an intercept and covariates are assumed to be scaled to the unit interval, i.e. \( x_{j1}, \ldots, x_{jn} \in [0, 1], j = 1, \ldots, d \) without loss of generality. Further, we allow for heteroscedasticity by allowing the residual variance to vary with one of the covariates or some linear combination of them denoted by \( \tilde{x} \).

To estimate unknown smooth functions \( f_j \) with penalized splines, we represent \( f_j(x) = (I_n - 1_n 1_n') B_j(x) \beta_j = \tilde{B}_j \beta_j \) with \( B_j(x) \) a B-spline basis function of degree \( p \) based on a large number of \( k_j \) knots.
knots \( \tau_j = \{ \tau_{j,1} < \ldots < \tau_{j,k_j} \} \) such that the approximation bias will be small enough. Identifiability is ensured by using the centered basis matrices \( \tilde{B}_j \).

The degree of smoothness of each function \( f_j(x_j) \) is allowed to vary with \( x_j \) with a small smoothing parameter for values of \( x_j \) where the function is wiggly and large smoothing parameter where it is smooth. A procedure that allows the function to be spatially inhomogeneous in such a way is said to be locally adaptive. To estimate such complex functions we employ the mixed models representation of penalized splines. To do so, we decompose each \( \tilde{B}_j \beta_j = \tilde{B}_j(F_0^T b_j + F_0^T u_j) = X_j b_j + Z_j u_j \) in such a way that \((F_0^T F_0)^{-1} D_j F_0 = 0 \) and \((F_0^T D_j F_0)^{-1} = I_{k_j + p + 1 - q} \), where \( D_j \) is such that \( \int_0^1 [\tilde{B}_j(x) \beta_j(q)]^2 dx = \beta_j^T D_j \beta_j \). Now assuming that \( u_0 \sim N(0, \sigma_{u_j}^2(\tau_j,s)) \), \( s = 1, \ldots, k_j \) and that the variance processes \( \sigma_{u_j}^2(\tau_j) \) and \( \sigma^2(\tilde{x}) \) are smooth functions leads to a linear mixed model. More precisely, we define a hierarchical mixed model

\[
Y = \beta_0 + \sum_{j=1}^d (X_j b_j + Z_j u_j) + \varepsilon \sim N(0, \Sigma) , u_j|c_j \sim N(0, \Sigma_u_j) ,
\]

\[
\Sigma = \operatorname{diag}\{ \exp(X_0 \gamma + Z_0 \varepsilon) \} , \quad \varepsilon \sim N(0, \Sigma_{\varepsilon}),
\]

\[
\Sigma_{u_j} = \operatorname{diag}\{ \exp(X_{u_j} \delta + Z_{u_j} \varepsilon) \} , \quad \varepsilon_j \sim N(0, \Sigma_{u_j}),
\]

where \( X_0, Z_0, X_{u_j}, Z_{u_j} \) are obtained by decomposing the spline bases in the same fashion as above, but based on smaller numbers of knots. All parameters of this model can be estimated from the corresponding (restricted) likelihood including locally adaptive smoothing parameters \( \lambda_j(\tau_j) = \sigma^2/\sigma^2_{u_j}(\tau_j) \) penalizing the integrated squared \( q \)-th derivative of the spline function.

To avoid numerically intensive computations, we follow Krivobokova et al. (2008) who suggested to use the Laplace approximation of the likelihood in the case of locally adaptive smoothing with homoscedastic errors which can analogously be extended to the heteroscedastic case.

**Goodness-of-Fit Test**

The difficulties when conducting inference in nonparametric regression (testing and simultaneous confidence bands) are caused by the fact that all nonparametric estimators are biased and the smoothing parameters are estimated from the data, introducing extra variability. Krivobokova et al. (2010) discussed simultaneous confidence bands and showed that using the mixed models representation of penalized splines in combination with the volume-of-tube formula the bias is automatically corrected for and the variability due to estimated smoothing parameter is negligible for sufficiently large \( n \). In this paper, we make use of these results and construct a goodness-of-fit test.

To do so, we define the test problem by the hypotheses \( H_0 : f_j(x_j) = f_{j,0}(x_j) \) and \( H_1 : f_j(x_j) = f_{j,0}(x_j) + g_j(x_j) \) \( \forall x_j \in [0,1] \) with \( f_{j,0}(x_j) \) a polynomial of degree \( q - 1 \) and \( g_j(x_j) \) an unspecified deviation. Further, we choose the B-spline basis such that \( f_{j,0}(x_j) = X_j b_j \). Then, testing for polynomial regression versus a general nonparametric alternative is equivalent to testing \( H_0 : f_j(x_j) = X_j b_j \) versus \( H_1 : f_j(x_j) = X_j b_j + Z_j u_j \) or equivalently \( H_0 : Z_j u_j = 0 \). The idea is to exploit the orthogonality of \( X_j b_j \) and \( Z_j u_j \) and to construct a simultaneous confidence band around the deviation from the parametric fit \( g_j(x_j) = Z_j u_j \). Then, the test procedure corresponds to checking whether the confidence band uniformly encloses the zero line coinciding with the test statistic

\[
T_j = \max_{x_j \in [0,1]} \left( \frac{|Z\hat{u}_j|}{\sqrt{\text{Var}\{Z\hat{u}_j\}}} \right)
\]

where \( \text{Var}\{Z\hat{u}_j\} \) is the variance of \( \hat{g}_j(x_j) \) with respect to the conditional distribution of \( Y \) treating \( u_j \) as fixed. That is, \( \text{Var}\{Z\hat{u}_j\} = \sigma^2 \operatorname{diag}(S_j(x) \Sigma \Sigma_j(x)^T) \) where \( S_j(x) = Z_j(x)\{Z_j^\Sigma \Sigma_j^{-1}(I - S_{-j})Z_j + \sigma^2 \Sigma^{-1}\}^{-1}Z_j^T(I - S_{-j})\Sigma^{-1} \) with \( S_{-j} = C_{-j}(C_{-j}^T \Sigma_{-1} C_{-j} + \Lambda_{-j})^{-1} C_{-j}^T \Sigma_{-1} \) where
\begin{align*}
C_{-j} &= [X_1, Z_1, X_2, Z_2, \ldots, X_{j-1}, Z_{j-1}, X_j, X_{j+1}, Z_{j+1}, \ldots, X_d, Z_d] \\
\Lambda_{-j} &= \text{blockdiag}(\Lambda_1, \Lambda_2, \ldots, \Lambda_{j-1}, \text{diag}(0_q), \Lambda_{j+1}, \ldots, \Lambda_d) \text{ with } \Lambda_j = \sigma^2 \text{blockdiag}(0_q, \Sigma_{u_j}^{-1}).
\end{align*}

Rejection of \( H_0 \) takes place if \( T_j > c_j \). To obtain the critical value \( c_j \) we consider with respect to the marginal distribution of \( Y \) the zero mean Gaussian process

\[
G_j(x) = \frac{Z_j(x)(\hat{u}_j - u_j)}{\sqrt{\text{Cov}(\hat{u}_j - u_j)Z_j(x)^t}} \sim \mathcal{N}(0, 1),
\]

where \( \text{Cov}(\hat{u}_j - u_j) = \{Z_j^t \Sigma_{u_j}^{-1}(I_n - S_{-j})Z_j + \sigma^2 \Sigma_{u_j}^{-1}\}^{-1} \) and

\[
\text{Cov}\{G_j(x_1), G_j(x_2)\} = \left( \frac{\ell_j(x_1)}{\|\ell_j(x_1)\|} \right)^t \left( \frac{\ell_j(x_2)}{\|\ell_j(x_2)\|} \right) = \eta_j^t(x_1) \eta_j(x_2),
\]

with \( \ell_j(x) = \{Z_j^t \Sigma_{u_j}^{-1}(I_n - S_{-j})Z_j + \sigma^2 \Sigma_{u_j}^{-1}\}^{-1/2}Z_j(x) \). Since \( G_j(x) \) is a zero mean Gaussian process, we can apply the volume-of-tube formula (Hotelling, 1939) to obtain \( c_j \) from

\[
\alpha = P \left( \sup_{x \in [0, 1]} |G_j(x)| \geq c_j \right) = \frac{\kappa_j}{\pi} \exp \left( -\frac{c_j^2}{2} \right) + 2 \{1 - \Phi(c_j)\} + o \left( \exp \left( -\frac{c_j^2}{2} \right) \right),
\]

with \( \kappa_j = \int_0^1 \frac{d}{dx} \eta_j(x) |dx| \) as the length of the mixed model manifold and \( \Phi(\cdot) \) the distribution function of a standard normal distribution.

Note that p-values can be obtained easily by calculating the tail probabilities by replacing \( c_j \) by a given \( T_j \) in the volume-of-tube formula. By exploiting the decomposition of the B-spline basis, improved power is obtained compared to the test strategy proposed in Claeskens & Van Keilegom (2003), for example, who build their proposed test on the simultaneous confidence band around \( f_j \) itself with corresponding hypotheses \( H_0 : f_j(x) = f^0_j(x) + g_j(x_j) \) versus \( H_1 : f_j(x) \neq f^0_j(x) + g_j(x_j) \) \( \forall x_j \in [0, 1] \) and rely on local polynomials for estimation and bootstrapping to obtain the critical value. That is, their test procedure corresponds to investigating a simultaneous confidence band around \( f_j(x_j) \) and not around \( g_j(x_j) \).

In the following section, we compare the performance of the proposed test with RLR tests using the simulation based approximation to the RLRRT distribution implemented in the R package RLRsim (Scheipl, 2010).

Tests for feature significance can be obtained by choosing \( q = 2 \) and constructing the test with respect to the first derivative of the function under consideration (see Ruppert et al., 2003, Chapter 6.8) restricting to the interval of interest.

**Simulation Study**

We consider additive models with i.i.d Gaussian errors

\[
Y = \mu_j(x_1, x_2, x_3) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad j = 1, 2, 3
\]

with

\[
\begin{align*}
\mu_1(x_1, x_2, x_3) &= \varphi_1 f_1(x_1) + x_2(1 - x_2) + f_2(x_2) + x_3 + f_3(x_3) \\
\mu_2(x_1, x_2, x_3) &= f_1(x_1) + x_2(1 - x_2) + \varphi_2 f_2(x_2) + x_3 + f_3(x_3) \\
\mu_3(x_1, x_2, x_3) &= f_1(x_1) + x_2(1 - x_2) + f_2(x_2) + x_3 + \varphi_3 f_3(x_3)
\end{align*}
\]

with \( \varphi_j \in [0; 0.6], \quad j = 1, 2, 3 \) corresponding to the separation distances between the null and the alternative. We test for no effect, second degree polynomial and for linearity of the components \( f^*_1(x_1) = \varphi_1 f_1(x_1), \quad f^*_2(x_2) = x_2(1 - x_2) + \varphi_2 f_2(x_2) \) and \( f^*_3(x_3) = x_3 + \varphi_3 f_3(x_3) \), respectively. To do so, B-spline bases with \( (p = 1, \ q = 1), \ (p = 5, \ q = 3) \) and \( (p = 3, \ q = 2) \), respectively, are used.
Further, $\sigma = 0.33$, $n = 300$, $k_j = 40$, $j = 1, 2, 3$ and $k_{w_3} = 5$ are chosen. Three Monte Carlo simulations with 1000 replications each were carried out. Results for $n = 600$ led to the same conclusions. As shown in Figure 1, the power curves of the proposed test and the RLR test are virtually identical.

REFERENCES


