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

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

where β_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 $f_j(x) = (\mathrm{I}_n - 1_n 1_n^t) B_j(x)\beta_j = \tilde{B}_j \beta_j$ with $B_j(x)$ a B-spline basis function of degree p based on a large number of k_j

knots $\tau_j = \{\tau_{j,1} < \ldots < \tau_{j,k_j}\}$ 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 $\tilde{B}_j\beta_j = \tilde{B}_j(F_b^j b_j + F_u^j u_j) = X_j b_j + Z_j u_j$ in such a way that $(F_u^j)^t F_b^j = (F_b^j)^t D_j F_b^j = 0$ and $(F_u^j)^T D_j F_u^j = I_{k_j+p+1-q}$, 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 $u_{js} \sim \mathcal{N}\{0, \sigma^2_{u_j}(\tau_{j,s})\}$, $s = 1, \ldots, k_j$ and that the variance processes $\sigma^2_{u_j}(\tau_j)$ and $\sigma^2(\tilde{x})$ are smooth functions leads to a linear mixed model. More precisely, we define a hierarchical mixed model

$$\begin{aligned} Y &= \beta_0 + \sum_{j=1}^{d}(X_j b_j + Z_j u_j) + \varepsilon,\ \varepsilon|v \sim N(0,\ \sigma^2\Sigma_\varepsilon),\ u_j|c_j \sim N(0,\ \Sigma_{u_j}), \\ \Sigma_\varepsilon &= \mathrm{diag}\{\exp(X_v\gamma + Z_v v)\}, \quad v \sim N(0,\ \sigma_v^2 I_{k_v}), \\ \Sigma_{u_j} &= \mathrm{diag}\{\exp(X_{w_j}\delta + Z_{w_j} w_j)\}, \quad w_j \sim N(0,\ \sigma^2_{w_j} I_{k_{w_j}}), \end{aligned}$$

where X_v, Z_v, X_{w_j} and Z_{w_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

$$T_j = \max_{x\in[0,1]} \left(|Z\hat{u}_j| / \sqrt{\mathrm{Var}\{Z\hat{u}_j\}} \right)$$

where $\mathrm{Var}\{Z\hat{u}_j\}$ is the variance of $\hat{g}_j(x_j)$ with respect to the *conditional distribution* of Y treating u_j as fixed. That is, $\mathrm{Var}\{Z\hat{u}_j\} = \sigma^2 \mathrm{diag}(S_j(x)\Sigma_\varepsilon S_j(x)^t)$ where $S_j(x) = Z_j(x)\{Z_j^t\Sigma_\varepsilon^{-1}(I - S_{-j})Z_j + \sigma^2\Sigma_{u_j}^{-1}\}^{-1} Z_j^t(I - S_{-j})\Sigma_\varepsilon^{-1}$ with $S_{-j} = C_{-j}(C_{-j}^t\Sigma_\varepsilon^{-1}C_{-j} + \Lambda_{-j})^{-1}C_{-j}^t\Sigma_\varepsilon^{-1}$ where

$C_{-j} = [X_1, Z_1, X_2, Z_2, \ldots, X_{j-1}, Z_{j-1}, X_j, X_{j+1}, Z_{j+1}, \ldots, X_d, Z_d]$ and
$\Lambda_{-j} = \text{blockdiag}(\Lambda_1, \Lambda_2, \ldots, \Lambda_{j-1}, \text{diag}(0_q), \Lambda_{j+1}, \ldots, \Lambda_d)$ with $\Lambda_j = \sigma^2 \text{blockdiag}(0_q, \Sigma_{u_j}^{-1})$.

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

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

where $\text{Cov}(\hat{u}_j - u_j) = \{Z_j^t \Sigma_\varepsilon^{-1}(I_n - S_{-j})Z_j + \sigma^2 \Sigma_{u_j}^{-1}\}^{-1}$ 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 $\ell_j(x) = \{Z_j^t \Sigma_\varepsilon^{-1}(I - S_{-j})Z_j + \sigma^2 \Sigma_{u_j}^{-1}\}^{-1/2} Z_j^t(x)$. 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

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

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 $H_0: \ f_j(x_j) = f_j^0(x_j) + g_j(x_j)$ versus $H_1: \ f_j(x_j) \neq f_j^0(x_j) + g_j(x_j) \ \ \forall x_j \in [0,1]$ 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 RLRT 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

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

with

$$\begin{aligned}
\mu_1(x_1, x_2, x_3) &= \varphi_1 f_1(x_1) + x_2(1 - x_2) + f_2(x_2) + x_3 + f_{32}(x_3) \\
\mu_2(x_1, x_2, x_3) &= f_1(x_1) + x_2(1 - x_2) + \varphi_2 f_2(x_2) + x_3 + f_{32}(x_3) \\
\mu_3(x_1, x_2, x_3) &= f_1(x_1) + x_2(1 - x_2) + f_2(x_2) + x_3 + \varphi_3 f_{32}(x_3)
\end{aligned}$$

with $\varphi_j \in [0; 0.6]$, $j = 1, 2, 3$ 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 $f_1^*(x_1) = \varphi_1 f_1(x_1)$, $f_2^*(x_2) = x_2(1 - x_2) + \varphi_2 f_2(x_2)$ and $f_3^*(x_3) = x_3 + \varphi_3 f_{32}(x_3)$, respectively. To do so, B-spline bases with $(p = 1, \ q = 1)$, $(p = 5, \ q = 3)$ and $(p = 3, \ q = 2)$, respectively, are used.

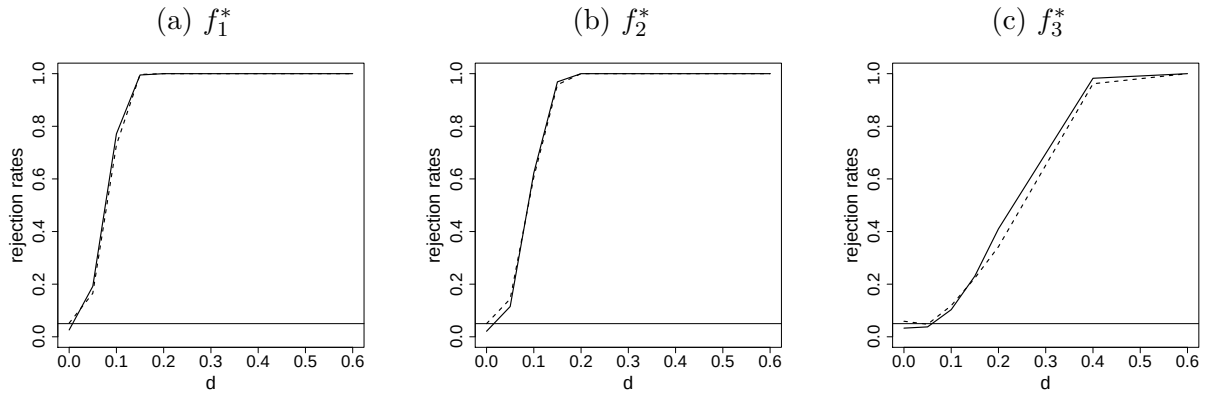

Figure 1: Empirical power curves of the proposed test (solid lines) and RLR test (dashed lines)

Further, $\sigma = 0.33$, $n = 300$, $k_j = 40$, $j = 1, 2, 3$ and $k_{w3} = 5$ 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

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. The Annals of Statistics, 31(6):1852–1884.

Crainiceanu, C., Ruppert, D., Claeskens, G. and Wand, M. (2005). Exact likelihood ratio tests for penalised splines. Biometrika, 92(1):91.

Härdle, W., Huet, S., Mammen, E. and Sperlich, S. (2004). Bootstrap inference in semiparametric generalized additive models. Econometric Theory, 20:265–300.

Hotelling, H. (1939). Tubes and spheres in n-spaces, and a class of statistical problems. American Journal of Mathematics, 61(2):440–460.

Krivobokova, T., Crainiceanu, C. and Kauermann, G. (2008). Fast adaptive penalized splines. Journal of Computational and Graphical Statistics, 17(1):1–20.

Krivobokova, T., Kneib, T. and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. Journal of the American Statistical Association, 105(490):852–863.

Ruppert, D., Wand, M., and Carroll, R. (2003). Semiparametric Regression. Cambridge University Press, Cambridge, U.K.

Scheipl, F. (2010). RLRsim: Exact (Restricted) Likelihood Ratio tests for mixed and additive models. R package version 2.0-5.

Sun, J. (1993). Tail probabilities of the maxima of Gaussian random fields. The Annals of Probability, 21(1):34–71.

Wiesenfarth, M., Krivobokova, T. and Klasen, S. (2010). Simultaneous Confidence Bands for Additive Models with Locally Adaptive Smoothed Components and Heteroscedastic Errors. Courant Research Centre: Poverty, Equity and Growth-Discussion Papers 50.