Biometrika (2012), **xx**, x, *pp*. 1–9

Printed in Great Britain

On p-values for smooth components of an extended generalized additive model

BY SIMON N. WOOD
MATHEMATICAL SCIENCES, UNIVERSITY OF BATH, BATH BA2 7AY U.K.
s.wood@bath.ac.uk

SUMMARY

The problem of testing smooth components of an extended generalized additive model for equality to zero is considered. Confidence intervals for such components exhibit good across the function coverage probabilities if based on the approximate result $\hat{f}(i) \sim N\{f(i), V_f(i,i)\}$, where f is the vector of evaluated values for the smooth component of interest, and V_f is the covariance matrix for f according to the Bayesian view of the smoothing process. It is therefore proposed to test the null hypothesis $f = 0$, using Wald type tests based on the statistic $T_r = \hat{f}^{\mathrm{T}} V_f^{r-} \hat{f}$, where V_f^{r-} is the rank r Moore–Penrose pseudoinverse of V_f, generalized to non-integer r. Consideration of the structure of T_r suggests setting r to the effective degrees of freedom of $\hat{f}$. Efficient computation of the p-values is considered. The method complements previous work by applying beyond the Gaussian case, while considering tests of zero effect, rather than testing the parametric hypothesis given by the null space of the component's smoothing penalty. The proposed p-values are routine and efficient to compute from a fitted model, without requiring extra model fits or simulation of the null distribution. Simulation results suggest improvement on possible alternative methods.

Keywords: hypothesis test, model selection, p-spline, semi-parametric regression, spline.

1. INTRODUCTION

Consider the extended generalized additive model (Hastie & Tibshirani, 1990)

$$g(\mu_i) = \sum_j A(i,j)\theta_j + \sum_j L_{ij}\mathcal{F}_j, \tag{1}$$

where y_i is a univariate response from some exponential family distribution with scale parameter ϕ and mean, μ_i, dependent on predictor variables via a parametric model matrix, A, with unknown coefficients, θ, and some unknown smooth functions, $\mathcal{F}_j$, of one or more variables. The L_{ij} are bounded linear functionals, in the simplest case $L_{ij}\mathcal{F}_j = \mathcal{F}_j(x_i)$, for example. Each $\mathcal{F}_j$ can be expanded in terms of known basis functions, $b_{jk}(x)$, usually from a reduced rank spline basis, which may differ between $\mathcal{F}_j$. So $\mathcal{F}_j(x) = \sum_k b_{jk}(x)\beta_{jk}$, where the β_{jk} are unknown parameters. A smoothing penalty, $\beta_j^{\mathrm{T}} \tilde{S}_j \beta_j$, is associated with each $\mathcal{F}_j$. Then (1) can be written as $g(\mu) = X\beta$, where X contains A and the evaluated $L_{ij}b_{jk}$, while β contains θ and the β_{jk}, in corresponding order. Defining S_j to be a matrix whose only non-zero block is given by $\tilde{S}_j$, such

that $\beta_j^{\mathrm{T}}\tilde{S}_j\beta_j \equiv \beta^{\mathrm{T}}S_j\beta$, and letting l denote the log likelihood, then β can be estimated as

$$\hat{\beta} = \arg\max_{\beta} \left\{ l(\beta) - \frac{1}{2}\sum_j \rho_j \beta^{\mathrm{T}} S_j \beta \right\},$$

where the ρ_j are smoothing parameters controlling the fit-smoothness and variance-bias trade-offs, and are often estimated by generalized cross validation or marginal likelihood maximization. Under a Bayesian view of the smoothing process, in which the smoothing penalty is induced by an improper Gaussian prior on β, then $\hat{\beta}$ is also the mode of the posterior density of β, and in the large sample limit, or exactly in the case of Gaussian y, $\beta|y \sim N(\hat{\beta}, V_\beta)$ where $V_\beta = (X^{\mathrm{T}}WX + \sum_j \rho_j S_j)^{-1}\phi$, W is diagonal with $W(i,i)^{-1} = v(\mu_i)g'(\mu_i)^2$ and $v(\mu)$ is the variance function of the exponential family distribution. The influence matrix for such a model is $H = XV_\beta X^{\mathrm{T}}W/\phi$. Generally $\dim(\beta) \leq n$, the number of y_i. For details see, for example, Wood (2006) and Wood (2011).

This paper is concerned with testing the null hypotheses $\mathcal{F}_j = 0$, for any j, within this framework. The key idea is to base the test statistic on the same distributional result that yields well calibrated confidence intervals for the $\mathcal{F}_j$, namely that $\hat{f}_j(i) \sim N\{f_j(i), V_{f_j}(i,i)\}$, approximately, where V_{f_j} is the Bayesian covariance matrix for f_j, the vector of $\mathcal{F}_j$ evaluated at the observed covariate values. If X_j is the matrix such that $\hat{f}_j = X_j\hat{\beta}$, then $V_{f_j} = X_j V_\beta X_j^{\mathrm{T}}$. The proposal is to use a statistic similar to the obvious choice, $T_r = \hat{f}_j^{\mathrm{T}} V_{f_j}^{r-} \hat{f}_j$, where $V_{f_j}^{r-}$ is a rank r pseudoinverse of V_{f_j}. The main problem is then to choose r appropriately. Naive choices lead to failures to produce good power or even the correct null distribution for p-values, as shown in figure 1, but investigation of the structure of T_r suggests a usable choice for r.

Existing work on the testing problem includes Cox et al. (1988), Liu & Wang (2004), Zhang & Lin (2003), Crainiceanu et al. (2005), Scheipl et al. (2008) and Nummi et al. (2011), but has usually focused on the hypothesis that $\mathcal{F}_j$ is a simple polynomial in the null space of the smoothing penalty associated with $\mathcal{F}_j$. Practitioners are as often interested in tests of whether a term should be included in the model at all, and here existing work has limitations. The exact test of Crainiceanu et al. (2005) applies to a Gaussian model with a single smooth term and so fails to cover most of the cases of interest here. Cantoni & Hastie (2002) provide an alternative for the Gaussian additive model case, but at $O(n^3)$ computational cost, and under the assumption that interest is in comparing two pre-specified degrees of freedom for a term. It would be useful to have a zero effect test applicable in the multi term generalized additive case when smoothing parameters have been estimated, and for this to have substantially less than $O(n^3)$ cost.

2. Tests from well calibrated intervals

2·1. *A Wald statistic*

Consider a single smooth component, $\hat{\mathcal{F}}_j(x)$, with a non-zero dimensional penalty null-space, evaluated at a value x_i chosen randomly from the observed values of x. Nychka (1988), with component-wise extension by Marra & Wood (2012), shows that the approximation,

$$\hat{\mathcal{F}}_j(x_i) \sim N\{\mathcal{F}_j(x_i), V_{f_j}(i,i)\}, \tag{2}$$

is well founded and leads to confidence intervals for $\mathcal{F}_j$ with close to nominal across the function coverage probabilities, including in the case when the smoothing parameter for $\mathcal{F}_j$ has been estimated as part of model fitting. The Nychka derivation shows that the Bayesian covariance matrix, V_{f_j}, can be viewed as including a squared bias component, treated as random across the function.

By including both bias and variance components, the intervals have coverage probabilities that are relatively insensitive to the values of the bias-variance controlling smoothing parameters.

The success of the Wahba (1983) intervals based on (2) suggests basing Wald tests on the same result. This requires the joint distribution of a vector, $\hat{f}_j$, of $\hat{\mathcal{F}}_j(x_i)$ values, while (2) only provides the corresponding marginal distributions. Ruppert et al. (2003, §6.4) show that (2) corresponds to $\hat{f}_j \sim N(f_j, V_{f_j})$ when the $\mathcal{F}_j$ are random functions, re-sampled from their prior distribution with each replication of the data, but further work is required for the usual case of $\mathcal{F}_j$ assumed fixed under such replication.

Let X_j denote a rank p matrix such that $\hat{f}_j = X_j\hat{\beta}$. p is the rank of the basis expansion used for f_j. Assume that X_j has $p(p+1)/2$ or more rows. We know that the covariance matrix of $\hat{f}_j$ must have the form $X_j V'_\beta X_j^{\mathrm{T}}$, where V'_β is the covariance matrix of $\hat{\beta}$, and from (2)

$$X_j(i,\cdot)V_\beta X_j(i,\cdot)^{\mathrm{T}} = X_j(i,\cdot)V'_\beta X_j(i,\cdot)^{\mathrm{T}} \tag{3}$$

for all i, so that the variances implied by V'_β match those from (2). Using standard results on Kronecker products (e.g. Harville, 1997, Chapter 16), (3) becomes

$$X_j \otimes_r X_j \mathrm{vec}(V_\beta) = X_j \otimes_r X_j \mathrm{vec}(V'_\beta), \tag{4}$$

where $\otimes_r$ denotes the row-wise Kronecker product, so that the i^{th} row of $X_j \otimes_r X_j$ is $X_j(i,\cdot) \otimes X_j(i,\cdot)$. Since $X_j \otimes_r X_j$ has rank $\geq p(p+1)/2$, by construction, and V_β is symmetric, (4) can only hold if $V'_\beta = V_\beta$. So the covariance matrix of $\hat{f}_j$, corresponding to (2), is $V_{f_j} = X_j^{\mathrm{T}} V_\beta X_j$, although the distribution of $\hat{f}_j$ has not been shown to be multivariate Gaussian.

Hence the Wald statistic corresponding to (2) is

$$T_r = \hat{f}_j^{\mathrm{T}} V_{f_j}^{r-} \hat{f}_j \tag{5}$$

where $V_{f_j}^{r-}$ is a rank r pseudoinverse of V_{f_j} or the generalization thereof discussed below. The rank, r, must be chosen, but naive choices lead to the poor test performance shown in figure 1.

2·2. *A well behaved Wald statistic*

To understand the failures of figure 1 requires investigation of the structure of T_r. For clarity, consider the simplified model $y_i = \mathcal{F}(x_i) + \epsilon_i$, where $\mathcal{F}$ is represented by a rank p penalized spline type smoother and the ϵ_i are independent $N(0, \sigma^2)$. Let $f^{\mathrm{T}} = \{\mathcal{F}(x_1), \ldots, \mathcal{F}(x_n)\}$ and H be the smoother matrix, such that $\hat{f} = Hy$. Without loss of generality assuming $\sigma^2 = 1$, we have that the covariance matrix of $\hat{f}$ corresponding to (2) is $V_f = H$. Now let the zero truncated eigen decomposition of this matrix be $V_f = U\Lambda U^{\mathrm{T}} = H$, where Λ is the $p \times p$ diagonal matrix of non-zero eigenvalues, λ_i, and U the p column matrix of corresponding eigenvectors, the i^{th} of which is u_i. We have $\hat{f} = U\Lambda U^{\mathrm{T}} y$. The u_i form a basis for f, so that $\hat{f} = \sum_{i=1}^{p} \hat{\gamma}_i u_i$, where $\hat{\gamma}_i = \lambda_i u_i^{\mathrm{T}} y$. The non-zero eigenvalues, λ_i, are all ≤ 1, and when the smoothing parameter is zero are all exactly 1. So $u_i^{\mathrm{T}} y$ is the i^{th} basis coefficient in the absence of penalization, and λ_i is the shrinkage factor for that coefficient applied by the smoother.

Now consider T_r with $r = p$. We can write

$$T_p = \hat{f}^{\mathrm{T}} V_f^{p-} \hat{f} = d^{\mathrm{T}} d = \sum_{i=1}^{p} d_i^2$$

where

$$d = \Lambda^{-1/2} U^{\mathrm{T}} \hat{f} = \Lambda^{-1/2} U^{\mathrm{T}} U \Lambda U^{\mathrm{T}} y = \Lambda^{-1/2} \Lambda U^{\mathrm{T}} y.$$

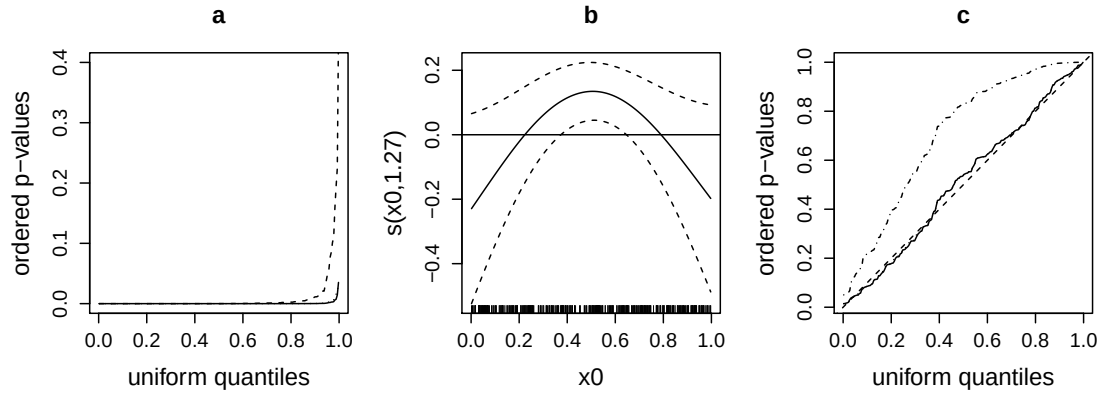

Fig. 1. Some poor p-value examples each based on model (6) from §3 with centred rank 10 cubic splines used to represent the smooth curves. a) Low power when the rank, r, of T_r is set to the numerical rank of V_f, thereby including many highly weighted, heavily penalized terms as discussed in §2·2. The dashed curve shows the ordered p-values for non-null term f_0, from §3, using this T_r, against uniform quantiles. The other two almost indistinguishable curves are for T_r of §2·2, and §3 alternative b. High power gives curves to the lower right of this plot. b) A case where simple rounding of the effective degrees of freedom gives low power. The panel shows the smooth estimate of the quadratic truth, f_0, from §3. T_r with r set to the rounded effective degrees of freedom of the term, rounds down and gives a p-value of 0.89, compared to 0.045 using §2. c) Poor distribution under the null hypothesis when r is the rounded up effective degrees of freedom. The dash-dot curve shows the ordered p-values from rounding up, against uniform quantiles. The dashed line is the ideal line, while the solid and dotted lines are from the T_r variants as in panel a. The poor null distribution results from rounding effective degrees of freedom very close to 1 up to 2, so that the statistic is dominated by a term penalized almost to zero.

So $d_i = \hat{\gamma}_i \lambda_i^{-1/2}$ for $i = 1, \ldots, p$, and the test statistic is a weighted sum of squares of the basis coefficients, $\hat{\gamma}_i$, where the weights are highest for the most heavily penalized coefficients. In consequence, tests based on T_p suffer a serious loss of power under all but light penalization, since T_p is then dominated by the components most heavily penalized towards zero, at the expense of the unequivocally non-zero components. Those components of f for which there is least evidence in the data, are those most heavily weighted in the statistic. Figure 1a provides one illustration of such a loss of power, while figure 1c shows a more dramatic example, where even the null distribution is spoiled.

This problem is avoided if the most heavily penalized components are dropped from T_p. To this end, consider the number of components that should be retained in order to obtain an optimal unpenalized approximation to the penalized estimate of f. If $\|\cdot\|$ is the Euclidean norm, then from standard results, the rank $k < p$ approximation to $\hat{f}$ minimizing $\max_{y \neq 0} \|f_k - \hat{f}\|/\|y\|$, and linear in y, is $\hat{f}_k = H_k y$, where $H_k = U_k \Lambda_k U_k^{\mathrm{T}}$, U_k is the first k columns of U and Λ_k is the diagonal matrix of the k largest eigenvalues of H. U_k does not depend on the smoothing parameter for f, so it also provides the minimax optimal unpenalized rank k basis for f. Let $\tilde{f}_k = U_k U_k^{\mathrm{T}} y$, the un-penalized version of $\hat{f}_k$, and since $\tilde{f}_k$ is not subject to smoothing bias, also consider the smoothing bias corrected penalized estimate, $\tilde{f} = \hat{f} + (\hat{f} - H\hat{f}) = (2H - HH)y$. Seeking to minimize the mean square error in approximating $\tilde{f}$ by $\tilde{f}_k$, suggests choosing k to minimize $\Delta = \mathrm{tr}\{\mathrm{cov}(\tilde{f}_k - \tilde{f})\}$. Defining $\Lambda' = 2\Lambda - \Lambda^2$, we have that $\Delta = k - 2\sum_{i=1}^{k} \Lambda'(i,i) + \mathrm{tr}(\Lambda'^2)$, which is minimized by k such that $\Lambda'(k,k) \geq 0.5$ and $\Lambda'(k+1,k+1) < 0.5$. Given that the $\Lambda'(i,i)$ form a sigmoidal decreasing sequence between 1 and 0, then $k \approx \tau = \mathrm{tr}(\Lambda')$. τ is one version of the effective degrees of freedom of $\hat{f}$.

Simply using T_k as the test statistic can lead to the loss of power illustrated in figure 1b, when term estimates are close to functions in the null space of the smoothing penalty, as d_i carrying

important information can then be dropped. One way to avoid both this dropping of important terms, and the overweighting of highly penalized terms, is to relax the requirement for integer degrees of freedom in the test statistic. Instead use $r = \tau$ in a generalized T_r, which is well defined for non-integer r, varies smoothly with r, but recovers a conventional Wald statistic when $r = \tau$ is integer.

In particular a generalized T_r is sought which, given (2), has null distribution χ^2_r when r is integer, but for non-integer r still has $E(T_r) = r$ and $\mathrm{var}(T_r) = 2r$, under the null hypothesis. One way to achieve this is by a slight generalization of V_f^{r-} to

$$V_f^{r-} = U \begin{bmatrix} \lambda_1^{-1} & & & & \\ & \ddots & & & \\ & & \lambda_{k-2}^{-1} & & \\ & & & B & \\ & & & & 0 \end{bmatrix} U^{\mathrm{T}}, B = \tilde{\Lambda}\tilde{B}\tilde{\Lambda}^{\mathrm{T}},\ \tilde{\Lambda} = \begin{bmatrix} \lambda_{k-1}^{-1/2} & 0 \\ 0 & \lambda_k^{-1/2} \end{bmatrix},\ \tilde{B} = \begin{bmatrix} 1 & \rho \\ \rho & \nu \end{bmatrix},$$

$k = \lfloor r \rfloor + 1$, $\nu = r - k + 1$ and $\rho = \{\nu(1-\nu)/2\}^{1/2}$. Hence, if $\delta_1 = (d_1, \ldots, d_{k-2})^{\mathrm{T}}$ and $\delta_2 = (d_{k-1}, d_k)^{\mathrm{T}}$, then $T_r = \delta_1^{\mathrm{T}}\delta_1 + \delta_2^{\mathrm{T}}\tilde{B}\delta_2$, and, given (2), routine manipulation confirms that this T_r has the desired properties under the null hypothesis.

2·3. *The distribution of* T_r

If (2) and the null hypothesis hold exactly then $E(d) = 0$, while $\mathrm{cov}(d) = I$. The statistic T_r is based on d_1 to d_k, which then tend to independent $N(0,1)$ by the multivariate central limit theorem of Lindeberg (1922), or by Ruppert et al. (2003, §6.4) if $\mathcal{F}$ is a frequentist random effect, and are in any-case marginally $N(0,1)$ with zero covariance. It follows that in the large sample limit under (2) and the null hypothesis, $T_r \sim \chi^2_r$, if r is integer, while for non-integer r,

$$T_r \sim \chi^2_{k-2} + \nu_1\chi^2_1 + \nu_2\chi^2_1,$$

where $\nu_1 = \{\nu + 1 + (1-\nu^2)^{1/2}\}/2$ and $\nu_2 = \nu + 1 - \nu_1$ are the eigenvalues of $\tilde{B}$. The cumulative distribution function of such a weighted sum of χ^2 random variables can reliably be evaluated by the method of characteristic function inversion of Davies (1980). The possibility of $\nu_2 \ll 1$ can make the series of Ruben (1962), or the integral of Imhof (1961) too slow here. Alternatively a gamma$(r/2, 2)$ approximation can be used, which can be made less crude by employing Liu et al. (2009) for the upper tail. The viable alternatives produce similar simulation results and are compared in the right panel of figure 3.

When the scale parameter has been estimated, the p-value must be computed as $\mathrm{pr}(\chi^2_{k-2} + \nu_1\chi^2_1 + \nu_2\chi^2_1 > t_r\chi^2_\kappa/\kappa)$ where κ is the residual degrees of freedom used to compute the scale estimate, and t_r is the observed T_r. This probability can readily be computed by quadrature, given access to the cumulative distribution function of a weighted sum of χ^2 random variables. The results here hold equally well for components $\mathcal{F}_j$ as for a single $\mathcal{F}$.

2·4. *Efficient computation of* T_r

Direct naive formation of $V_{f_j}^{r-}$ has $O(n^3)$ computational cost, but this can be reduced to at most $O(np^2)$. For computational purposes we are interested in $\hat{f}_j = X_j\hat{\beta} = \bar{X}_j\hat{\beta}_j$ and $V_{f_j} = X_jV_\beta X_j^{\mathrm{T}} = \bar{X}_jV_{\beta_j}\bar{X}_j^{\mathrm{T}}$, where V_{β_j} is the Bayesian covariance matrix of β_j, the coefficient vector of f_j, and $\bar{X}_j$ contains the non zero columns of X_j. Forming the QR decomposition

$$\bar{X}_j = Q\begin{bmatrix} R \\ 0 \end{bmatrix}$$

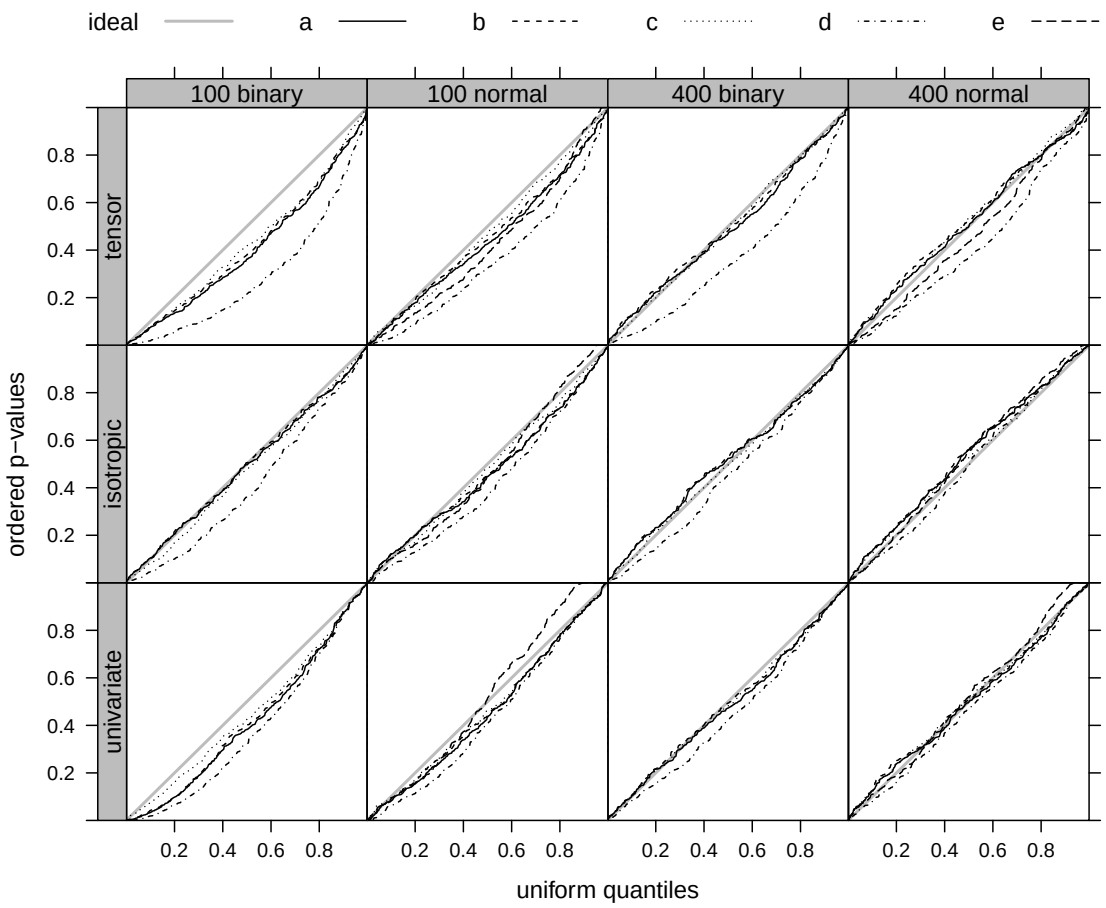

Fig. 2. Quantile-quantile plots for the p-values computed in §3 when the null hypothesis is correct. Only sample sizes 100 and 400 for correlated covariates are shown as these gave the worst results for the method proposed in this paper. Method e can only be computed for the normal data case. For all panels at sample size 400 a,b and c are indistinguishable from uniform in a Kolmogorov-Smirnov test at the 5% level, unlike d and e.

we obtain

$$(X_j V_\beta X_j^{\mathrm{T}})^{r-} = Q \begin{bmatrix} (RV_{\beta_j} R^{\mathrm{T}})^{r-} & 0 \\ 0 & 0 \end{bmatrix} Q^{\mathrm{T}} = Q_1 (RV_{\beta_j} R^{\mathrm{T}})^{r-} Q_1^{\mathrm{T}}$$

where Q_1 is the first p columns of Q, so that $\bar{X}_j = Q_1 R$. Hence,

$$T_r = \hat{f}_j^{\mathrm{T}} V_{f_j}^{r-} \hat{f}_j = \hat{\beta}_j^{\mathrm{T}} R^{\mathrm{T}} Q_1^{\mathrm{T}} Q_1 (RV_{\beta_j} R^{\mathrm{T}})^{r-} Q_1^{\mathrm{T}} Q_1 R \hat{\beta}_j = \hat{\beta}_j^{\mathrm{T}} R^{\mathrm{T}} (RV_{\beta_j} R^{\mathrm{T}})^{r-} R \hat{\beta}_j,$$

which is computationally efficient.

For large datasets, little is usually gained by using the whole of $\bar{X}_j$ to compute T_r, and we might as well use a random sample of n_s of its rows, reducing computational cost to $O(n_s p^2)$. Note that if $F = V_\beta X^{\mathrm{T}} W X / \phi$ then τ_j, the required effective degrees of freedom for $\hat{f}_j$, can be obtained by summing the diagonal elements of $2F - FF$ corresponding to $\hat{\beta}_j$.

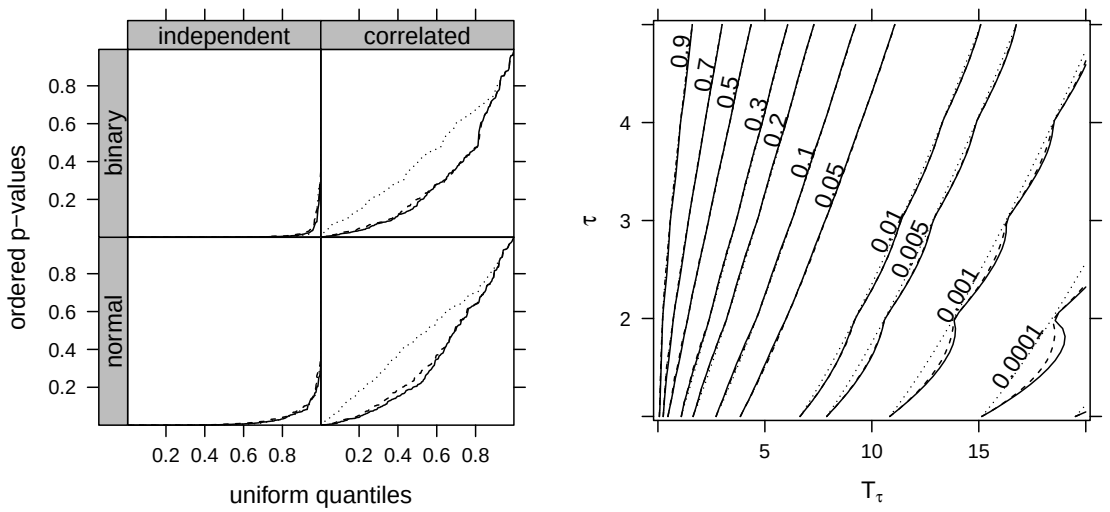

Fig. 3. Left: quantile-quantile plots illustrating the power of variants a, b and c from §3, with line styles as in figure 2. Higher power is to the lower right of each panel. Variant c gives low power with correlated covariates. Right: contour plot of p-values computed with the three methods given in §2·3, over the range of τ and T_τ where important differences are expected. Continuous contours are for the exact method, dotted is the gamma approximation, and dashed is the gamma approximation with upper tail correction.

3. Simulation results

This section uses simulated data to illustrate the performance of the p-values computed using a) T_τ from §2·2 and four possible alternatives: b) T_r based on $r = \lfloor \tau \rfloor$, if $\tau - \lfloor \tau \rfloor < 0.05$, and $r = \lceil \tau \rceil$ otherwise, c) an adaptation of Cox et al. (1988) to the zero effect additive component setting, using the statistic $T'_p = \sum_{i=1}^{p} \lambda_i d_i^2$, d) the ad hoc approach of Wood (2006, §4.8.5) and e) the method of Cantoni & Hastie (2002), but using estimated, rather than fixed, smoothing parameters in the alternative models. Option e only applies in the Gaussian case and is not practical at the largest sample size used. No previously published methods directly cover zero effect p-values for generalized additive model components with multiple estimated smoothing parameters, but the Cox et al. and Cantoni and Hastie approaches are readily adapted to this setting.

Data were simulated from a linear predictor with the structure

$$\eta = f_0(x_0) + f_1(x_1) + f_2(x_2) \tag{6}$$

where $f_0(x) = 8x(1-x)$, $f_1(x) = \exp(2x)$ and $f_2(x) = 2 \times 10^5 x^{11}(1-x)^6 + 10^4 x^3 (1-x)^{10}$, which is multi-peaked. For each replicate, n values of each of covariates x_0 to x_4 were simulated, where n was 100, 400 or 4000. Marginally the observations of each x_j were independent $U(0,1)$. Two alternative correlation settings were used: either all x_j were mutually independent, or the r^2 between x_0 and x_1 was set to 95%, and also between x_2 and x_3. Gaussian and Bernoulli response distributions were used, as contrasting cases. In the Gaussian case the response was given by $y_i = \eta_i + \epsilon_i$ where the ϵ_i were independent $N(0, 3^2)$. In the binary case the y_i were generated as independent Bernoulli random deviates, with mean $\mu_i = \exp(\eta_i - 5)/\{1 + \exp(\eta_i - 5)\}$.

400 replicate data sets were simulated at each sample size at each correlation setting for each response distribution. For each replicate the correct distribution and link function were assumed, but three alternative models for the linear predictor were used. For the first $\eta = \alpha + f_0(x_0) + f_1(x_1) + f_2(x_2) + f_3(x_3)$ was assumed, where the f_j are smooth univariate func-

tions (represented using penalized thin plate regression splines). For the second and third $\eta = \alpha + f_0(x_0) + f_1(x_1) + f_2(x_2) + f_3(x_3, x_4)$ was used, with f_3 a penalized thin plate regression spline for the second alternative, and a two penalty tensor product of cubic regression splines for the third alternative. So in all cases the true f_3 is zero. Smoothing parameter estimation was by maximum likelihood. Restricted maximum likelihood results are similar, while generalized cross validation gave slightly worse results because of a small proportion of seriously under-smoothed components. R 2.14.0 (R Development Core Team, 2010) with mgcv 1.7-14 was used. Each alternative p-value was computed for the spurious term, f_3, for each replicate.

As expected the p-value distributions for a-c, T_τ, T_r and T'_p, improve with increasing n and also with decreasing covariate correlation, since high covariate correlation results in high smoothing parameter uncertainty, which is neglected in the methods considered here. The biggest departures from the ideal uniform distribution are for $n = 100$ and $n = 400$ with correlated covariates, and these are the cases shown in figure 2. The T_τ, T_r and T'_p based p-value performance is starting to deteriorate at $n = 100$ for binary data and correlated covariates, but by $n = 400$, the distributions are indistinguishable from nominal. The Wood (2006) statistic, d, underestimates p-values even at sample size 4000, while the Cantoni & Hastie (2002) approach, e, also gives incorrect null distribution, particularly for the uncorrelated covariate cases not shown, presumably as a result of the necessary adaptation of using estimated smoothing parameters.

Having investigated the distribution of the p-values under the null hypothesis, a further simulation was conducted to investigate power for the three approaches giving close to the correct uniform p-value null distribution. This time sample size 400 was used, and Gaussian and binary data simulated as before, except that the linear predictor for the binary case was multiplied by 2. With these settings the estimate of f_0 is interestingly on the border of significance. 200 replicates were generated for each distribution and each correlation setting, and the p-values associated with f_0 were computed using the §2·2 test, a, as well as alternative b and c above. Figure 3 shows the results. The three alternatives show similar performance for uncorrelated covariates, but T'_p suffers serious loss of power, relative to T_r, when the covariates are correlated. Alternative b performs similarly to a, but has the practical drawback of depending discontinuously on smoothing parameters. Overall option a, T_τ, appears to give the best performance.

4. Discussion

It has been demonstrated how effective p-values can be computed for testing smooth components of (1) for equality to zero, based on the results of Nychka (1988) and Marra & Wood (2012). The proposal appears to be the first well founded zero effect test for components of a generalized additive model in which there are several estimated smoothing parameters, albeit that it is conditional on those estimates. It has the practical advantage of being efficiently and routinely computable. Finally, although it is not the primary purpose of this paper, in principle the null hypothesis that a component is in the null space of its penalty can be tested by omitting d_i components corresponding to $\lambda_i = 1$ from T_τ, and reducing the degrees of freedom of the χ^2_{k-2} component of the null distribution accordingly, however the possibility of estimating the resulting T_τ to be zero complicates the study of this approach. The p-values discussed here are implemented in function `summary.gam` of R package `mgcv` from version 1.7-14.

Acknowledgements

This work was part funded by the United Kingdom Higher Education Funding Council, and is part of the research programme of the United Kingdom National Centre for Statistical Ecology. I thank the referees, editor and associate editor for comments which improved the paper.

REFERENCES

CANTONI, E. & HASTIE, T. (2002). Degrees-of-freedom tests for smoothing splines. *Biometrika* **89**, 251–63.

COX, D., KOH, E., WAHBA, G. & YANDELL, B. S. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.* **16**, 113–119.

CRAINICEANU, C., RUPPERT, D., CLAESKENS, G. & WAND, M. P. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika* **92**, 91–103.

DAVIES, R. B. (1980). Algorithm as 155: The distribution of a linear combination of chi-2 random variables. *J. R. Statist. Soc. C* **29**, 323–333.

HARVILLE, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. London: Champan & Hall.

IMHOF, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48**, 419–426.

LINDEBERG, J. W. (1922). Eine neue herleitung des esponentialgesetzes in der wahrscheinlichkeitsrechnung. *Math. Z.* **15**, 211–225.

LIU, A. & WANG, Y. (2004). Hypothesis testing in smoothing spline models. *J. Stat. Comput. Sim.* **74**, 581–597.

LIU, H., TANG, Y. & ZHANG, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comp. Statist. Data Anal.* **53**, 853–856.

MARRA, G. & WOOD, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scand. J. Statist.* **39**, 53–74.

NUMMI, T., PAN, J., SIREN, T. & LIU, K. (2011). Testing for cubic smoothing splines under dependent data. *Biometrics* **67**, 871–875.

NYCHKA, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Am. Statist. Assoc.* **83**, 1134–1143.

R DEVELOPMENT CORE TEAM (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RUBEN, H. (1962). Probability content of regions under spherical normal distributions, iv: The distribution of homogeneous and non-homogeneous quadratic functions of normal variables. *Ann. Math. Statist.* **33**, 542–570.

RUPPERT, D., WAND, M. P. & CARROLL, R. J. (2003). *Semiparametric Regression*. London:Cambridge University Press.

SCHEIPL, F., GREVEN, S. & KÜCHENHOFF, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Comp. Statist. Data Anal.* **52**, 3283–3299.

WAHBA, G. (1983). Bayesian 'confidence intervals' for the cross-validated smoothing spline. *J. R. Statist. Soc. B* **45**, 133–150.

WOOD, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: CRC/Chapman & Hall.

WOOD, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Statist. Soc. B* **73**, 3–36.

ZHANG, D. & LIN, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4**, 57–74.