Smoothing parameter uncertainty in general smooth models

Sonja Greven, Fabian Scheipl

1 Introduction

In their article "Smoothing parameter and model selection for general smooth models", Wood, Pya and Säfken make a significant and impressive contribution to the development of general smooth models and specifically to estimation, inference and model selection for these models. Given the popularity of the R package `mgcv` (Wood, 2011) to date, the newly developed methods and their implementation in `mgcv` are likely to shape the routine use of smooth models - for more general model classes than were previously available - in the future.

The developed framework is very flexible and can be used even beyond the models covered in the paper. For example, by shifting the functional structure to an appropriate smooth additive predictor (Scheipl et al., 2015), we have used methods for smooth models (Wood, 2011) to develop flexible functional additive mixed models for functional data. The new extensions in the discussed article then allowed us to extend this approach to "generalized" functional data (Scheipl et al., 2016; Greven and Scheipl, 2016), where the observations can be thought of as coming from some non-Gaussian process with underlying smoothness assumption and e.g. negative binomial, t- or Beta marginal distributions. It is a large advantage for researchers and users of flexible regression models that the `mgcv` package provides such a highly performant, innovative and well-documented implementation of state-of-the-art methodology.

While many aspects in the paper are worthy of attention, we focus on smoothing parameter uncertainty and applaud the authors' effort to develop methods taking it into account for statistical inference. As covariances and confidence intervals for the regression coefficients are motivated from a Bayesian viewpoint, we initially focus on the implicit prior assumptions and their effects on the posteriors (Section 2) before coming back to the effects of smoothing parameter uncertainty on coefficient uncertainty in Section 3.

To facilitate discussion and intuition, we use a running example of a nonparametric regression model

$$y_i = m(x_i) + \varepsilon_i,\ \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n. \tag{1}$$

In simulations, we take x_i to be equidistant in $[0, 1]$ and the true $m(x_i) = d\exp(2x_i)$ for some d. Let $m(\cdot)$ be parameterized using a penalized spline basis expansion such that the coefficient vector $\boldsymbol{\beta}$ projected into the penalty null-space corresponds to a straight line for $m(\cdot)$, while the projection into the span of the penalty matrix captures deviations from linearity. This example has only one log-smoothing

parameter $\rho_1 \equiv \rho$ controlling smoothness of $m(\cdot)$, with $\rho \to \infty$ leading to a linear estimate for $m(\cdot)$ and $\rho \to -\infty$ yielding an unpenalized least squares fit.

2 Prior assumptions and posteriors

The authors motivate their approach for smoothing parameter uncertainty and model selection using a Bayesian viewpoint. We thus think it informative to take a look at the implicit prior assumptions of the approach and their effects on the posteriors.

Results in the paper are with respect to $\boldsymbol{\rho}$, where each entry is a log-smoothing parameter $\rho_j = \log(\lambda_j)$. We can view the penalized log-likelihood in equation (1) of the paper as the joint or the PQL-approximate log-likelihood (Ruppert et al., 2003, pp. 204-216) of a mixed model with (usually improper) normal prior density $f(\boldsymbol{\beta}|\boldsymbol{\rho}) \propto \exp(-\frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{S}^\lambda \boldsymbol{\beta})$ and $\widehat{\boldsymbol{\beta}}$ (as maximizer of (1)) as posterior mode maximizing $f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\rho})$ for a given $\boldsymbol{\rho}$. If there is only one penalty associated with each subvector $\boldsymbol{\beta}_j$ of $\boldsymbol{\beta}$, i.e. different blocks are non-zero for different $\boldsymbol{S}^j$, λ_j in $\boldsymbol{S}^\lambda = \sum_j \lambda_j \boldsymbol{S}^j$ corresponds to the inverse of a variance parameter τ_j^2 associated with the random effect $\boldsymbol{\beta}_j$ and controlling deviations of $g_j(\cdot)$ from the null space of the penalty. Estimated model coefficients $\widehat{\boldsymbol{\beta}}$ converge to estimates in the null space of the penalty matrix $\boldsymbol{S}^j$ if $\rho_j \to \infty$ and thus $\lambda_j \to \infty$ or $\tau_j^2 \to 0$.

Note that in a parameterization with respect to τ_j^2, a function $g_j(\cdot)$ can be estimated to lie exactly in the null space of the penalty (e.g. an exactly linear function in our running example) when $\widehat{\tau}_j^2 = 0$ is on the boundary of the parameter space $[0, \infty)$ for τ_j^2. This case corresponds to $\widehat{\rho}_j \to \infty$ in the $\rho_j \in (0, \infty)$ parameterization, which can never be exactly estimated, although the differences in $\widehat{\boldsymbol{\beta}}$ to a very large $\widehat{\rho}_j$ (very small $\widehat{\tau}_j^2$) are negligible. Considering the alternative τ_j^2 parameterization also makes explicit that there is a boundary issue for $\tau_j^2 = 0$ or $\rho_j \to \infty$ that is well-known to cause non-standard asymptotic behavior e.g. when testing whether $g_j(\cdot)$ lies in the null-space of the penalty (Crainiceanu and Ruppert, 2004; Greven and Crainiceanu, 2013).

For estimation, the Laplace approximation (LAML) of the log marginal likelihood criterion (2)

$$\log \int f(\boldsymbol{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\rho}) d\boldsymbol{\beta} = \log f(\boldsymbol{y}|\boldsymbol{\rho})$$

is maximized with respect to $\boldsymbol{\rho}$. This can be seen as approximately maximizing the posterior density

$$f(\boldsymbol{\rho}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\rho}) f(\boldsymbol{\rho}) \propto f(\boldsymbol{y}|\boldsymbol{\rho})$$

under independent improper uniform priors $\mathcal{U}(-\infty, \infty)$ for each ρ_j, $f(\boldsymbol{\rho}) \propto 1$. This prior corresponds to an improper prior on $(0, \infty)$ for either λ_j or τ_j^2 with density proportional to $1/\lambda_j$ respectively $1/\tau_j^2$ and can lead to an improper posterior $f(\boldsymbol{\rho}|\boldsymbol{y})$ (Gelman, 2006). Note that the posterior under proper $\mathcal{U}[a_j, b_j]$ priors has the same mode $\widehat{\boldsymbol{\rho}}$ as for improper $\mathcal{U}(-\infty, \infty)$ priors as long as $\widehat{\rho}_j \in [a_j, b_j]$ for all j. However, if the marginal likelihood is monotonically increasing in ρ_j, $\mathcal{U}[a_j, b_j]$ priors yield b_j as the posterior mode for ρ_j and not infinity.

Since the posterior distribution of $\boldsymbol{\rho}$ with $\mathcal{U}[a_j, b_j]$ priors, $a_j, b_j \in [-\infty, \infty]$, has density

$$f(\boldsymbol{\rho}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\rho}) I(a_j \leq \rho_j \leq b_j \; \forall \; j) = \exp(\mathcal{V}_r(\boldsymbol{\rho})) I(a_j \leq \rho_j \leq b_j \; \forall \; j), \tag{2}$$

where the marginal likelihood $\exp(\mathcal{V}_r(\boldsymbol{\rho}))$ is defined as a function of $\boldsymbol{\rho}$ analogous to equation (2) of the article, it is informative to look at the shape of the Laplace approximation $\exp(\mathcal{V}(\boldsymbol{\rho}))$. Figure 1 shows $\exp(\mathcal{V}(\rho))$ as a function of ρ for four simulations from model (1) with $n = 200$, $d = 0.1$ and $\sigma = 0.5$, estimating a smooth function for $m(\cdot)$ using the `gam` defaults in `mgcv` with `method = "REML"`. From left to right, estimates $\widehat{\rho}$ are increasing. In the rightmost panel the LAML is maximized for $\rho \to \infty$. For comparison, the normal posterior densities for ρ as given by (6) in the paper, using $\widehat{\rho}$ and $\boldsymbol{V}_\rho$ as returned by `mgcv` as mean and variance, are scaled to have the same maximum values and are overlaid in red. We can see that in all cases 1) the non-diminishing mass for $\rho \to \infty$ leads to an improper posterior that cannot be normalized, even though this is hardly visible in the leftmost panel with smallest $\widehat{\rho}$. In particular, the LAML stays practically constant for increasing ρ after a certain point that roughly results in a linear fit for $m(\cdot)$. 2) The normal approximation of equation (6) is reasonable locally near $\widehat{\rho}$ in the three leftmost panels, but does not hold for $\rho \to \infty$ or for the strongly regularized function estimate in the rightmost panel.

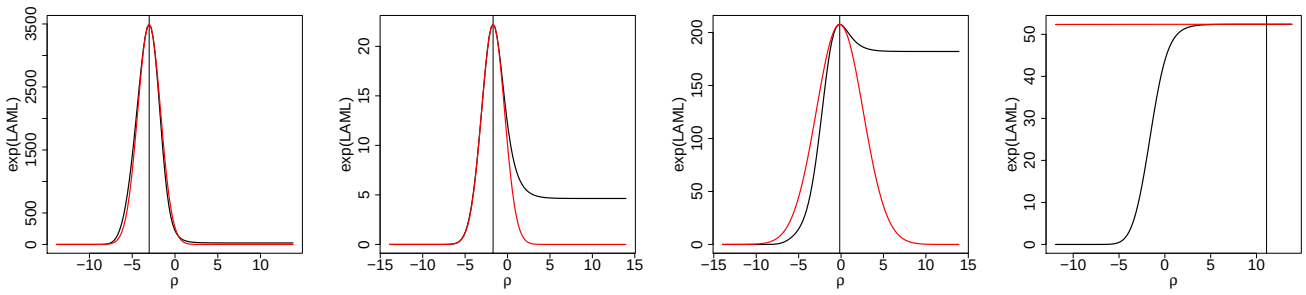

Figure 1: The LAML criterion $\exp(\mathcal{V}(\rho))$ (in black), as a function of ρ for four settings with increasing $\widehat{\rho}$ (vertical lines, as returned by `mgcv`). In the right panel, the LAML is maximized for $\rho \to \infty$. Red lines indicate the normal posterior density for ρ as given by (6) in the paper, scaled to have the same maximum value and using $\widehat{\rho}$ and $\boldsymbol{V}_\rho$ as returned by `mgcv` as mean and variance, respectively.

For the conditional posterior of $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho}$, we know that it is asymptotically Gaussian $\mathcal{N}(\widehat{\boldsymbol{\beta}}_\rho, \boldsymbol{V}_\beta(\boldsymbol{\rho}))$ for each given $\boldsymbol{\rho}$, cf. equation (5) of the discussed article, where we have made the dependence of $\boldsymbol{V}_\beta$ on $\boldsymbol{\rho}$ explicit in our notation. The asymptotic marginal posterior of $\boldsymbol{\beta}|\boldsymbol{y}$ thus is generated similarly to a scale mixture of multivariate normal distributions with density

$$f(\boldsymbol{\beta}|\boldsymbol{y}) = \int f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho}) f(\boldsymbol{\rho}|\boldsymbol{y}) d\boldsymbol{\rho}, \tag{3}$$

where $f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho})$ is the $\mathcal{N}(\widehat{\boldsymbol{\beta}}_\rho, \boldsymbol{V}_\beta(\boldsymbol{\rho}))$ density and the approximate shape of $f(\boldsymbol{\rho}|\boldsymbol{y})$ is shown in Figure 1.

To look at the marginal posterior of $\boldsymbol{\beta}$, we use the `jagam` function (Wood, 2016) in `mgcv`, which allows to extract model specifications of a generalized additive model (GAM) and to fit the model using full Bayesian inference via MCMC in JAGS (Plummer, 2016). We simulate again from running example (1) with $n = 100$ and $\sigma = 1$. The prior for the log-smoothing parameter ρ is set to $\rho \sim \mathcal{U}[-12, 12]$, since only proper priors are possible here, but estimates $\widehat{\rho}$ are always smaller than 12. Figure 2 shows a strongly nonlinear fit for $d = 2$ (top), and an almost linear fit for $d = 0.1$ (bottom). Note that the two estimates (left panels) from `mgcv` and `jagam` are not identical, as `mgcv` uses the posterior mode of $f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho})$ at the posterior mode $\widehat{\boldsymbol{\rho}}$ of $f(\boldsymbol{\rho}|\boldsymbol{y})$, while `jagam` uses the posterior mean of $f(\boldsymbol{\beta}|\boldsymbol{y})$. Confidence/credible bands and the posterior densities (right panels) are however based on the marginal posterior $f(\boldsymbol{\beta}|\boldsymbol{y})$ for both, with differences due to the normal approximation of the marginal posterior in `mgcv`. We can see that the posterior $f(\boldsymbol{\beta}|\boldsymbol{y})$ is close to normal in the strongly nonlinear setting, where $\widehat{\rho}$ is relatively small. For the almost linear estimate, however, where $\widehat{\rho}$ is large, most coefficients show a spiky posterior with most mass close to zero and heavier tails than a Gaussian, while one coefficient exhibits a bimodal distribution. Here, the normal approximation is not very close and the resulting confidence band (bottom left) of `mgcv` is noticeably narrower than the credible band from `jagam`.

3 Smoothing parameter uncertainty

Consider now the posterior covariance of $\boldsymbol{\beta}$, which is used for confidence band construction and in the conditional AIC. The usual uncorrected covariance $\boldsymbol{V}_\beta = \boldsymbol{V}_\beta(\boldsymbol{\rho})$ is based on the conditional posterior $f(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\rho})$ with $\widehat{\boldsymbol{\rho}}$ plugged in. The authors propose a corrected covariance $\boldsymbol{V}'_\beta$ to account for smoothing parameter uncertainty, based on approximating the marginal posterior $f(\boldsymbol{\beta}|\boldsymbol{y})$ using a linear Taylor expansion.

Equation (6) of the paper postulates an asymptotic posterior distribution $\boldsymbol{\rho}|\boldsymbol{y} \sim N(\widehat{\boldsymbol{\rho}}, \boldsymbol{V}_\rho)$ in the interior of the parameter space, where $\boldsymbol{V}_\rho$ is the inverse of the Hessian of the negative log marginal likelihood $-\mathcal{V}_r$ with respect to $\boldsymbol{\rho}$. The boundary case corresponds to $\widehat{\rho}_j \to \infty$, with $\widehat{\rho}_j$ values treated as 'working infinity' during estimation when $\partial\mathcal{V}/\partial\rho_j \approx \partial^2\mathcal{V}/\partial\rho_j^2 \approx 0$. In this case, the authors propose to substitute a Moore-Penrose pseudoinverse of the Hessian. Consequences are most easily discussed for the case of a scalar ρ. When $\widehat{\rho} \to \infty$, the Hessian of the negative log marginal likelihood is $-\partial^2\mathcal{V}_r/\partial\rho^2|_{\rho=\widehat{\rho}} \approx -\partial^2\mathcal{V}/\partial\rho^2|_{\rho=\widehat{\rho}} \to 0$. The Moore-Penrose pseudoinverse of 0 is again 0, i.e. $\boldsymbol{V}_\rho \to 0$. Thus, the posterior for ρ collapses to a point mass at $\widehat{\rho} \approx \infty$ (corresponding to a point mass at $\tau^2 \approx 0$). In this case, $\boldsymbol{J}$ and $\frac{\partial}{\partial\rho}\boldsymbol{R}_\rho$ also converge to $\boldsymbol{0}$, formula (7) gives $\boldsymbol{V}'_\beta \to \boldsymbol{V}_\beta$ and the corrected and uncorrected covariances coincide. This is also practically confirmed in simulations for our running example: When the estimated effective degrees of freedom approach 2, the differences in the two estimated covariance matrices approach zero.

Thus, smoothing parameter uncertainty is not accounted for near the boundary of the parameter space, where coverage is most of an issue (Marra and Wood, 2012). For very large smoothing parameters, confidence bands are effectively based on the submodel defined by the penalty nullspace. For example, in our running example, if $m(\cdot)$ is estimated to be linear, confidence bands are computed based on the linear submodel, ignoring the possibility of truly nonlinear functions. Confidence bands are too narrow

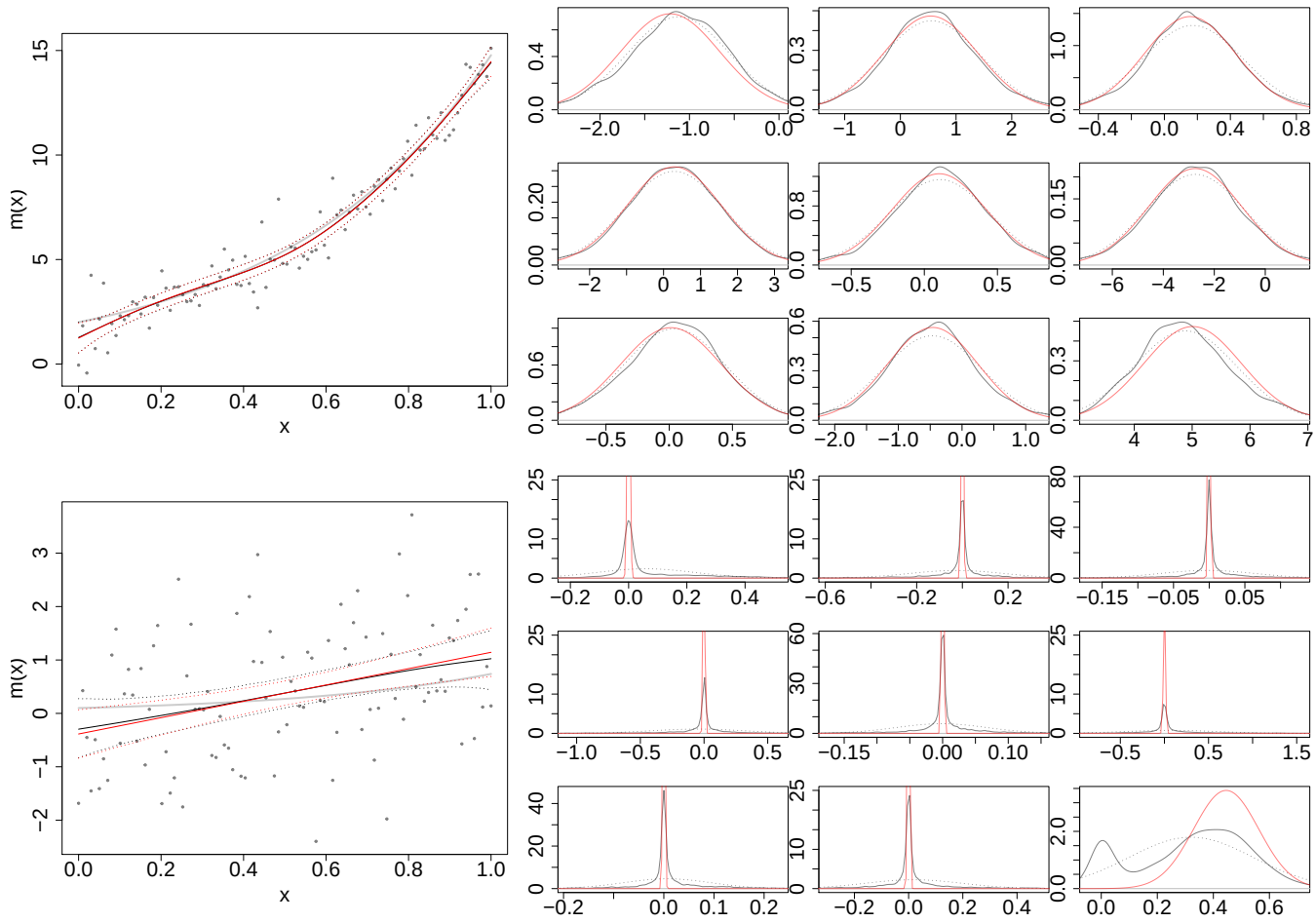

Figure 2: The marginal posterior densities of $\boldsymbol{\beta}|\boldsymbol{y}$ from `jagam`, separately for each component of $\boldsymbol{\beta}$ (nine small panels, in black), with normal distributions fitted to the posterior sample overlaid in dashed black, and the Gaussian approximation based on the smoothing parameter uncertainty corrected covariance $\boldsymbol{V}'_\beta$ from `mgcv` in red for the estimate shown in the large panel (Bayesian estimate and credible intervals in black, estimate and confidence interval from `mgcv` in red, true function in grey, data as dots). Shown are two simulations from running example (1) with strong nonlinearity $d = 2$ (top) and small nonlinearity $d = 0.1$ (bottom). Vertical axis cropped in some panels for clarity.

and linear in this case. This leads to undercoverage for functions $m(\cdot)$ that are not linear, but close enough that there is a relevant probability of estimating a linear $\widehat{m}(\cdot)$, cf. Figure 3. For strongly nonlinear functions, linear estimates rarely occur and do not noticeably change coverage. For truly linear functions, computing confidence bands based on a linear submodel also does not lead to undercoverage. Thus, while coverage is not affected in most cases, for functions that are only slightly nonlinear relative to the noise level, there is a noticeable dip in coverage ("phase transition" between nonlinear and linear models).

Since the linear Taylor approximation does not work well on the boundary of the parameter space for $\widehat{\rho}_j \to \infty$, we sketch two possible alternatives. In order to work with a normal approximation for the marginal posterior of $\boldsymbol{\beta}$ that incorporates uncertainty in $\boldsymbol{\rho}$, we can use the law of total covariance

$$\mathrm{Cov}(\boldsymbol{\beta}|\boldsymbol{y}) = \mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\mathrm{Cov}(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\rho})] + \mathrm{Cov}_{\boldsymbol{\rho}|\boldsymbol{y}}[\mathrm{E}(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\rho})] = \mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\boldsymbol{V}_\beta(\boldsymbol{\rho})] + \mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\widehat{\boldsymbol{\beta}}_\rho \widehat{\boldsymbol{\beta}}_\rho^\top] - \mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\widehat{\boldsymbol{\beta}}_\rho]\,\mathrm{E}_{\boldsymbol{\rho}|\boldsymbol{y}}[\widehat{\boldsymbol{\beta}}_\rho]^\top,$$

where expectations are with respect to the posterior distribution $f(\boldsymbol{\rho}|\boldsymbol{y})$ given in (2).

Alternatively, abandoning the normality assumption which might be questionable near the boundary, see Figure 2, we can directly use (3) to write $f(\boldsymbol{\beta}|\boldsymbol{y})$ as a continuous mixture of $\mathcal{N}(\widehat{\boldsymbol{\beta}}_\rho, \boldsymbol{V}_\beta(\boldsymbol{\rho}))$ densities. As $\boldsymbol{\beta}$ is multi-dimensional and we are usually interested in linear combinations of $\boldsymbol{\beta}$, it is easier to work with $f(\boldsymbol{x}^\top\boldsymbol{\beta}|\boldsymbol{y}) = \int f(\boldsymbol{x}^\top\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\rho}) f(\boldsymbol{\rho}|\boldsymbol{y}) d\boldsymbol{\rho}$ for some relevant vector $\boldsymbol{x}$, which is a one-dimensional scale mixture of $\mathcal{N}(\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}_\rho, \boldsymbol{x}^\top\boldsymbol{V}_\beta(\boldsymbol{\rho})\boldsymbol{x})$ densities with mixing distribution given in (2).

To compute pointwise level $(1-\alpha)$ confidence bands for either approach, we first define a grid $\boldsymbol{\rho}_r$, $r = 1, \ldots, R$, of values covering the prior domain, e.g. for a one-dimensional ρ a grid covering the interval $[a, b]$. We then compute weights $w(\boldsymbol{\rho}_r) = \exp(\mathcal{V}(\boldsymbol{\rho}_r)) / \sum_{r=1}^R \exp(\mathcal{V}(\boldsymbol{\rho}_r))$ to use in numerical integration with respect to the posterior $f(\boldsymbol{\rho}|\boldsymbol{y})$. This requires R refits of the model with $\boldsymbol{\rho}$ fixed at $\boldsymbol{\rho}_r$, $r = 1, \ldots, R$, to obtain $\exp(\mathcal{V}(\boldsymbol{\rho}_r))$ as the LAML for this model fit. For the first approach, we can then approximate the total covariance as

$$\mathrm{Cov}(\boldsymbol{\beta}|\boldsymbol{y}) \approx \sum_{r=1}^R w(\boldsymbol{\rho}_r) \left[\boldsymbol{V}_\beta(\boldsymbol{\rho}_r) + \widehat{\boldsymbol{\beta}}_{\rho_r} \widehat{\boldsymbol{\beta}}_{\rho_r}^\top\right] - \left[\sum_{r=1}^R w(\boldsymbol{\rho}_r) \widehat{\boldsymbol{\beta}}_{\rho_r}\right] \left[\sum_{r=1}^R w(\boldsymbol{\rho}_r) \widehat{\boldsymbol{\beta}}_{\rho_r}\right]^\top,$$

where $\widehat{\boldsymbol{\beta}}_{\rho_r}$ and $\boldsymbol{V}_\beta(\boldsymbol{\rho}_r)$ are obtained from the model output as the estimate and uncorrected covariance when refitting the model with $\boldsymbol{\rho}$ fixed at $\boldsymbol{\rho}_r$. For the second approach, we can approximate $f(\boldsymbol{x}^\top\boldsymbol{\beta}|\boldsymbol{y})$ by a discrete mixture of $\mathcal{N}(\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}_{\rho_r}, \boldsymbol{x}^\top\boldsymbol{V}_\beta(\boldsymbol{\rho}_r)\boldsymbol{x})$ densities, with mixture weights again given by the $w(\boldsymbol{\rho}_r)$. The $\alpha/2$ and $(1-\alpha/2)$ quantiles for this mixture can then be obtained e.g. using the R package `nor1mix` (Mächler, 2016).

Note that both approaches use the posterior density $f(\boldsymbol{\rho}|\boldsymbol{y})$, which is only proper if finite prior limits $[a_j, b_j]$ are used. We have seen in Section 2 that these prior limits, if chosen sufficiently large, do not have a strong influence on the model fit: If there is a maximum of the LAML within these bounds, it is not affected by them, and if the LAML is monotonically increasing for ρ_j, the estimate $\widehat{\boldsymbol{\beta}}$ is hardly affected by further increasing ρ_j once it approximately lies in the null space of the penalty. When using the posterior, however, these chosen boundaries do make a difference. Increasing b_j puts more prior weight on functions $g_j(\cdot)$ close to the penalty null space, while decreasing b_j puts more weight

on functions deviating from the null space, e.g. on linear and nonlinear $m(\cdot)$ in our running example, respectively. Thus, the two proposed alternatives also depend on the chosen prior limits.

We compared the different methods in a small simulation study for our running example with $n \in \{50, 200\}$, $\sigma \in \{0.2, 0.5\}$ and $d \in \{0, 0.1, 0.5, 1, 2\}$ and worked with the prior limits $a = -10$ and $b \in \{4, 6, 8, 10\}$. (Results were not found to be sensitive to the lower limit here, as $\exp(\mathcal{V}(\rho))$ quickly decreases towards zero for small ρ values.) For the grid ρ_r we used $\hat{\rho} + \ln(10)k$ with $k = (-10, -9.9, \ldots, 9.9, 10)$, i.e. an equidistant grid centered on $\hat{\rho}$, truncated to $[a, b]$.

Figure 3 shows results for average coverage across $[0, 1]$ for 12000 replications for $(n, \sigma) = (50, 0.2)$ and $(200, 0.5)$. There is a clear pattern of undercoverage for the uncorrected covariance $\boldsymbol{V}_\beta$ for $d = 0.1$, which is only partly corrected by using the corrected covariance $\boldsymbol{V}'_\beta$. For these two settings, $d = 0.1$ is only slightly nonlinear compared to the noise level, resulting in a high percentage (29% respectively 41%) of (near-)linear estimates with estimated degrees of freedom below 2.1. For $d = 0$, computing confidence bands in the linear subspace leads to nominal or over-coverage, as does the normal approximation for large d values. Using either the total covariance or the mixture approximation to $f(\boldsymbol{\beta}|\boldsymbol{y})$ gives very similar coverage to the corrected covariance independent of b in the case of large d values, where estimates are almost never linear and the normal approximation works well. For small $d = 0.1$, undercoverage is improved compared to the corrected covariance, at the cost of some overcoverage for $d = 0$ (on the order of the overcoverage all methods show for larger d values). Coverage is dependent on upper limit b of the uniform prior for ρ, with smaller b placing more weight on nonlinear functions and thus leading to higher coverage values. We interpret the undercoverage we observe for some large b, i.e., poorly calibrated credible/confidence intervals for $\widehat{m}(x)$, as a result of prior-data-conflict in these settings: large values of b put an inordinate amount of prior weight on (approximately) linear functions, while the observations come from a non-linear data generating process. For $(n, \sigma) = (200, 0.2)$, near-linear estimates occur in only 2% of simulations for $d = 0.1$ due to the higher information content of the data, leading to no undercoverage using $\boldsymbol{V}'_\beta$ and no advantages of the alternative methods. The dip in coverage for $d = 0.1$ is also smaller for $(n, \sigma) = (50, 0.5)$, possibly due to the wide confidence bands in this setting with low signal-to-noise ratio. Based on our limited four simulations, the mixture approximation with a low b value seems to give coverage closest to constant across different values of d. Computing time for the confidence intervals based on the mixture approximation were 2.5 to 3.5 seconds on a laptop computer for $n = 200$ depending on b and without grid optimization or parallelization.

Figure 4 illustrates the corrected covariance and mixture approximation confidence intervals for our running example. The left panel shows the linear estimate for $m(\cdot)$ and the pointwise confidence band based on the corrected covariance in green. Confidence bands based on the mixture approximation with $a = -10$ and $b = 10$ or $b = 4$ are shown in dark and light blue, respectively. It is clear that these acknowledge the possibility of the true function being nonlinear and are wider in particular in the middle and towards the ends of the $[0, 1]$ interval. This leads to increased coverage of the truly nonlinear function. The right panel shows the mixture approximation to the posterior of $m(x_i)$ for an example x_i in the middle of the interval. Uncertainty in ρ here leads to mixture components (for smaller ρ values) with smaller mean and larger variance. This results in a skewed posterior with long lower tail and thus in a smaller lower bound of the confidence/credible interval. It is also clear from the different locations of the maxima that there is a difference between the maximum of the marginal posterior density $f(\boldsymbol{\beta}|\boldsymbol{y})$

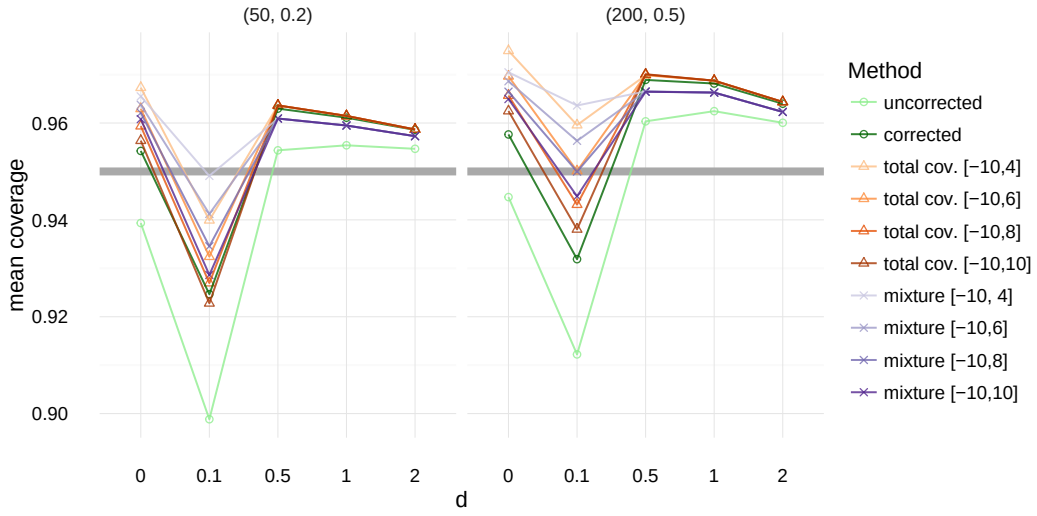

Figure 3: Average coverage across $[0, 1]$ for nominal level $(1 - \alpha) = 0.95$ in running example (1) with $(n, \sigma) = (50, 0.2)$ (left) and $(200, 0.5)$ (right). Coverages based on $\boldsymbol{V}_\beta$ and $\boldsymbol{V}'_\beta$ (circles) are denoted by uncorrected and corrected, respectively. The prior supports for the total covariance (triangles) and mixture approximation (crosses) methods are given in the legend.

(light blue) and the maximum of the conditional posterior density $f(\boldsymbol{\beta}|\boldsymbol{y}, \widehat{\rho})$ (green). Thus, in addition to not being symmetrical, the interval based on the mixture approximation is also not centered on the same value as that based on the corrected covariance.

4 Summary

The proposal by Wood, Pya and Säfken is an important step in the direction of accounting for smoothing parameter uncertainty in inference for $\boldsymbol{\beta}$. While it works well in most settings, for the 'phase transition' between functions in the null space of the penalty and those far from the null space it does not lead to well calibrated inference. Although the problem does not seem to be extremely large, we did see some undercoverage in confidence bands due to the neglect of smoothing parameter uncertainty when the function is effectively estimated to lie in the null space of the penalty ($\widehat{\rho}_j \to \infty$ case).

We have discussed two alternative approaches to the problem, which seem to improve undercoverage occurring in some of the cases we looked at. To more fully develop either approach would require a more efficient way to choose a suitable grid with the smallest possible number of grid points and a much wider simulation study including models with more than one smoothing parameter and other response distributions. Preliminary results for a simple logistic GAM indicate that the two discussed alternatives seem to be transferable in principle. We also saw that results remain sensitive to the chosen prior limits of the uniform priors for the log-smoothing parameters $\boldsymbol{\rho}$. The best combination of the

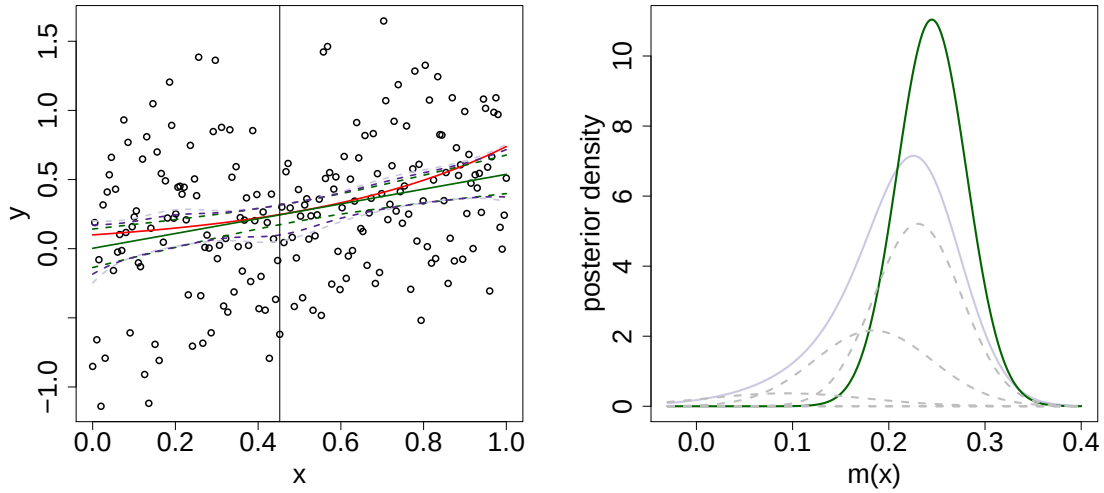

Figure 4: Left: Data (circles) simulated from the running example with $n = 200$, $d = 0.1$ and $\sigma = 0.2$, true function $m(\cdot)$ (red), estimate and pointwise confidence band based on the corrected covariance (green) and confidence bands based on the mixture approximation for $a = -10$ and $b = 10$ (dark blue) or $b = 4$ (light blue). Right: For the x marked by a vertical line on the left, posterior density as given by the normal approximation with corrected covariance (green) and the approximation by a normal mixture distribution with $b = 4$ (light blue). Dashed grey lines show aggregated mixture components with average means and standard deviations as well as total weights within each fifth of the grid for ρ.

discussed alternatives and the Wood, Pya and Säfken approach might be to only compute the mixture approximation when the estimate is close to the penalty null-space, as the corrected covariance seems to work well in all other cases. We think it remains worthwhile to think about alternative approaches for smoothing parameter uncertainty near the boundary as well as the relationship to model selection via the newly proposed conditional AIC, for which the corrected covariance is also used.

References

C. M. Crainiceanu and D. Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1): 165–185, 2004.

A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

S. Greven and C. M. Crainiceanu. On likelihood ratio testing for penalized splines. *AStA Advances in Statistical Analysis*, 97(4):387–402, 2013.

S. Greven and F. Scheipl. A general framework for functional regression modelling. *Statistical Modelling*, to appear, 2016.

M. Mächler. *nor1mix: Normal (1-d) Mixture Models (S3 Classes and Methods)*, 2016. URL `http://CRAN.R-project.org/package=nor1mix`. R package version 1.2-2.

G. Marra and S. N. Wood. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1):53–74, 2012.

M. Plummer. *rjags: Bayesian Graphical Models using MCMC*, 2016. URL `https://CRAN.R-project.org/package=rjags`. R package version 4-6.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

F. Scheipl, A.-M. Staicu, and S. Greven. Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501, 2015.

F. Scheipl, J. Gertheiss, and S. Greven. Generalized functional additive mixed models. *Electronic Journal of Statistics*, 10(1):1455–1492, 2016.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011.

S. N. Wood. Just another gibbs additive modeller: Interfacing jags and mgcv. *arXiv preprint arXiv:1602.02539*, 2016. URL `https://arxiv.org/abs/1602.02539`.