Wood, S. N., & Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. Biometrics. DOI: 10.1111/biom.12666

 DOI: 10.1111/biom.12666

A Generalized Fellner-Schall Method for Smoothing Parameter Optimization with Application to Tweedie Location, Scale and Shape Models

Simon N. Wood* **and Matteo Fasiolo**

School of Mathematics, University of Bristol, Bristol, U.K.
**email:* simon.wood@bath.edu

Summary. We consider the optimization of smoothing parameters and variance components in models with a regular log likelihood subject to quadratic penalization of the model coefficients, via a generalization of the method of Fellner (1986) and Schall (1991). In particular: (i) we generalize the original method to the case of penalties that are linear in several smoothing parameters, thereby covering the important cases of tensor product and adaptive smoothers; (ii) we show why the method's steps increase the restricted marginal likelihood of the model, that it tends to converge faster than the EM algorithm, or obvious accelerations of this, and investigate its relation to Newton optimization; (iii) we generalize the method to any Fisher regular likelihood. The method represents a considerable simplification over existing methods of estimating smoothing parameters in the context of regular likelihoods, without sacrificing generality: for example, it is only necessary to compute with the same first and second derivatives of the log-likelihood required for coefficient estimation, and not with the third or fourth order derivatives required by alternative approaches. Examples are provided which would have been impossible or impractical with pre-existing Fellner-Schall methods, along with an example of a Tweedie location, scale and shape model which would be a challenge for alternative methods, and a sparse additive modeling example where the method facilitates computational efficiency gains of several orders of magnitude.

Key Words: Fisheries; Smoothing parameter; REML; GAMLSS; Sparse additive model.

1. Introduction

This article is about a simple method for estimating the smoothing parameters and certain other variance parameters of models with a regular log likelihood, subject to quadratic penalization. The method generalizes the algorithm of Fellner (1986) and Schall (1991), by extending the range of smooth model terms with which it can deal, and generalizing beyond the GLM setting to models with any Fisher regular likelihood. The advantage of the Fellner-Schall algorithm is that it offers a simple explicit formula by which smoothing and variance parameters can be iteratively updated to optimize the model restricted likelihood, using essentially the same quantities anyway required in order to estimate the model coefficients. This has led to variants of it being use with smooth additive models, by Rigby and Stasinopoulos (2014) amongst others. However, the original method lacks generality, applying only to smooth terms each having a single smoothing parameter, so that tensor product smooth interactions and adaptive smoothers can not be employed. Rodríguez-Álvarez et al. (2015) partially remove this restriction for some tensor product smooths, but what we propose here is both simpler and more general. Furthermore, the original method only applies to GLM type likelihoods, with application beyond that setting relying on treating linearized approximations as Gaussian. Again what we propose is simpler and more general. Finally the original method derivations, while plausible, do not prove that the algorithm steps each increase the restricted likelihood, nor offer any insight into convergence rates. We address these issues, thereby largely removing the objection that Fellner-Schall smoothing parameter updates were somewhat ad hoc and insufficiently general.

In part, we are motivated by problems in fisheries stock assessment. For example, Figure 1a shows data from a 2010 survey for mackerel eggs off the coast of western Europe. Such surveys are undertaken in order to help estimate the mass of spawning adults that must be present, and generalized additive models provide suitable spatial models for the mean egg density. As with most fisheries data, the egg counts tend to be highly over-dispersed relative to a Poisson distribution, and a Tweedie (1984) distribution based model can offer a much better fit: the variance of a Tweedie random variable y_i, with mean μ_i, is given by $\text{var}(y_i) = \phi\mu_i^p$ where ϕ and p (here $1 < p < 2$) are parameters. An important biological feature is that mackerel are known to favor spawning grounds close to the continental shelf edge, for which the 200 m depth contour offers a reasonable proxy. However, if mackerel are responding to sea depth, there is no good reason to suppose that this response leads only to a change in the mean density of eggs in the water column: other aspects of the distribution shape are also likely to be affected, and a reasonable model would allow the parameters p and ϕ to vary smoothly as sea depth varies.

In principle such a model lies in the GAMLSS class of Rigby and Stasinopoulos (2005) and could be estimated using

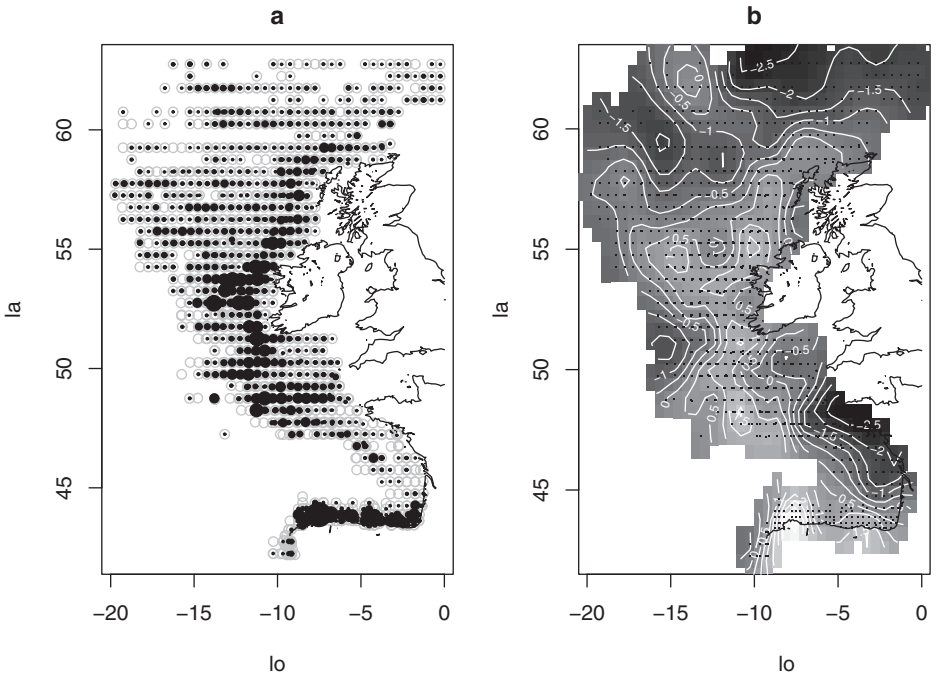

Figure 1. a. Mackerel (*Scomber scombrus*) egg data from the 2010 survey. Gray circles are survey locations, black circles are proportional to the 4th root of egg count. b. Image and contour plot of the spatial effect from the Tweedie location scale and shape model described in Section 6.

Wood et al. (2016). However, there is no publicly available software for estimating a Tweedie location scale and shape model. The problem is that the normalizing constant of the Tweedie density is a function of p and μ and is computable only by summing an infinite series "from the middle" (Dunn and Smyth, 2005). Wood et al. (2016) show how to obtain first and second derivatives of the log density with respect to p and μ, in a numerically stable way, but for covariate dependent p and ϕ the Wood et al. (2016) method would require the corresponding third and fourth derivatives as well. Hence, it would be useful to have a smoothing parameter estimation method that is general enough to encompass a Tweedie location scale and shape model, while avoiding the need for higher derivatives of the log density.

To introduce the smoothing parameter estimation problem in more detail, first consider the simple case of a Gaussian additive model for a univariate response variable

$$y_i = \mathbf{A}_i\boldsymbol{\theta} + \sum_j g_j(x_{ji}) + \epsilon_i \tag{1}$$

where $\mathbf{A}_i$ is the i^{th} row of a parametric model matrix, $\boldsymbol{\theta}$ is a vector of unknown coefficients, g_j is a smooth function of (possibly multivariate) covariate x_j, and the ϵ_i are independent $N(0,\sigma^2)$ random deviates. The g_j can be represented using reduced rank spline bases, with associated quadratic penalties penalizing departure from smoothness during fitting. For example, $g_j(x) = \sum_k b_{jk}(x)\gamma_{jk}$, where the b_{jk} are spline basis functions and the γ_{jk} are coefficients: the associated smoothing penalty is then $\lambda_j\boldsymbol{\gamma}_j^{\mathsf{T}}\mathcal{S}_j\boldsymbol{\gamma}_j$, where $\mathcal{S}_j$ is a fixed matrix, and is usually rank deficient because some functions are treated as "completely smooth." λ_j is a smoothing parameter, controlling the strength of penalization during fitting. In general, each g_j may have several penalties. We denote single elements of a vector/matrix, $\mathbf{x}/\mathbf{X}$, by x_i/X_{ij} (not $\mathbf{x}_i/\mathbf{X}_{ij}$).

It is well established (e.g., Kimeldorf and Wahba, 1970; Silverman, 1985; Ruppert et al., 2003) that smoothing penalties can be viewed as resulting from improper Gaussian prior distributions on the spline coefficients, in which case (1) can be re-written as a linear mixed effects model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \ \boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{S}_\lambda^- \sigma^2) \text{ and } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2), \tag{2}$$

where σ^2 and $\boldsymbol{\lambda}$ are parameters, $\boldsymbol{\beta}$ is a coefficient vector containing $\boldsymbol{\theta}$ and the coefficients for each smooth term, and $\mathbf{X}$ is an $n \times p$ model matrix, containing $\mathbf{A}$ and the evaluated basis functions of the smooth terms. $\mathbf{S}_\lambda$ is a positive semi-definite precision matrix, with Moore-Penrose pseudoinverse $\mathbf{S}_\lambda^-$. Let $\mathbf{S}_j$ be $\mathcal{S}_j$ padded out with zeroes, so that $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\beta} = \boldsymbol{\gamma}_j^{\mathsf{T}}\mathcal{S}_j\boldsymbol{\gamma}_j$, where $\boldsymbol{\gamma}_j$ is the coefficient vector for g_j. Then $\mathbf{S}_\lambda = \sum_j \lambda_j\mathbf{S}_j$ (some g_j may each be penalized by several terms in this summation, so that some summation terms, $\lambda_j\mathbf{S}_j$, are themselves replaced by summations $\sum_k \lambda_{jk}\mathbf{S}_{jk}$). The null space of $\mathbf{S}_\lambda$ is interpretable as the space of model fixed effects, whereas the range space is the space of random effects. Obviously, other simple Gaussian random effect terms can be included in the model in addition to smooth functions.

Fellner (1986) developed a simple iteration for updating $\boldsymbol{\lambda}$ in order to maximize the restricted marginal likelihood of (2), for the special case in which $\mathbf{S}_\lambda = \sum_j \lambda_j\mathbb{I}_j$, the $\mathbb{I}_j$ being identity matrices with most of their diagonal entries zeroed, and no non-zero entries in common between different $\mathbb{I}_j$. Schall

(1991) extended this to generalized linear mixed models. Here, we first give a simple generalization of the Fellner-Schall method that applies to any model with the structure (2), including smooth additive models in which the smoother terms each have multiple smoothing parameters. We also show why the method improves the restricted marginal likelihood at each step, which is something not revealed by the conventional derivations of the original method. In the additive Gaussian setting our main result is the update formula

$$\lambda_j^* = \hat{\sigma}^2 \frac{\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) - \mathrm{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}}{\hat{\boldsymbol{\beta}}^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol{\beta}}}\lambda_j,$$

where

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n - \mathrm{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\}}.$$

We also consider updates in the case of any model giving rise to a regular likelihood, but with the previously described prior distribution structure on $\boldsymbol{\beta}$, resulting in the general update (8) in Section 4: generalized linear mixed models are a special case. The update formula is iteratively alternated with evaluation of $\hat{\boldsymbol{\beta}}$ given the current $\boldsymbol{\lambda}$ estimates.

The rest of the article is structured as follows. We first consider the case of Gaussian additive models, deriving a Fellner-Schall type update that can deal with terms with multiple smoothing parameters using a derivation that shows, by construction, that the update must increase the model restricted marginal likelihood. We then study the method in the context of updating one smoothing parameter from a model with several smoothing parameters, showing that it takes longer steps than the EM algorithm, or the most obvious acceleration of the EM algorithm, while not overshooting the maximum of the restricted marginal likelihood, at least in the large sample limit. The update is then generalized to the case of any Fisher-regular likelihood, at the cost of a large sample approximation borrowed from the PQL method. Finally, we present simple examples which were not possible with previous Fellner-Schall methods, before returning to the Tweedie location scale and shape model for the Mackerel data.

2. The Gaussian Case Update

This section derives the update in a manner that gives a conveniently general form, and readily generalizes further. The next section provides theoretical insight into why it is effective. For model (2), the improper log joint density of the data, $\mathbf{y}$, and coefficients, $\boldsymbol{\beta}$, can be written as,

$$\log f_\lambda(\mathbf{y}, \boldsymbol{\beta}) = -\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_\lambda\boldsymbol{\beta}}{2\sigma^2} + \log|\mathbf{S}_\lambda/\sigma^2|_+/2 + c,$$

where $|\mathbf{S}_\lambda|_+$ denotes the product of the non-zero eigenvalues of $\mathbf{S}_\lambda$ and we use c to denote a parameter independent constant, which may vary from expression to expression. Following Wood (2011), the log restricted marginal likelihood can conveniently be written as,

$$l_r(\boldsymbol{\lambda}) = -\frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2 + \hat{\boldsymbol{\beta}}_\lambda^\mathsf{T}\mathbf{S}_\lambda\hat{\boldsymbol{\beta}}_\lambda}{2\sigma^2} + \log|\mathbf{S}_\lambda/\sigma^2|_+/2 - \log|\mathbf{X}^\mathsf{T}\mathbf{X}/\sigma^2 + \mathbf{S}_\lambda/\sigma^2|/2 + c,$$

where $\hat{\boldsymbol{\beta}}_\lambda = \mathrm{argmax} f_\lambda(\mathbf{y}, \boldsymbol{\beta})$ for a given $\boldsymbol{\lambda}$. Expressing the joint density and l_r in this way is the key to straightforwardly obtaining a general update formula. Given that $\partial(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \boldsymbol{\beta}^\mathsf{T}\mathbf{S}_\lambda\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\hat{\boldsymbol{\beta}}_\lambda} = \mathbf{0}$, by definition of $\hat{\boldsymbol{\beta}}_\lambda$, we have

$$\frac{\partial l_r}{\partial \lambda_j} = \mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j)/2 - \mathrm{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}/2 - \hat{\boldsymbol{\beta}}_\lambda^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol{\beta}}_\lambda/(2\sigma^2).$$

So $\partial l_r/\partial\lambda_j$ will be negative if $\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) - \mathrm{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\} < \hat{\boldsymbol{\beta}}_\lambda^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol{\beta}}_\lambda/\sigma^2$, indicating that λ_j should be decreased. If the inequality is reversed then $\partial l_r/\partial\lambda_j$ is positive, indicating that λ_j should be increased. If the inequality becomes an equality then $\partial l_r/\partial\lambda_j = 0$ and λ_j should not be changed. A final requirement of any update is that λ_j should remain positive, but by Theorem 1 or Remark 1, below, $\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) - \mathrm{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\} \geq 0$, while $\hat{\boldsymbol{\beta}}_\lambda^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol{\beta}}_\lambda \geq 0$ by the positive semi-definiteness of $\mathbf{S}_j$. Hence a simple update that meets all four requirements is

$$\lambda_j^* = \sigma^2 \frac{\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) - \mathrm{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}}{\hat{\boldsymbol{\beta}}_\lambda^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol{\beta}}_\lambda}\lambda_j, \tag{3}$$

with λ_j^* set to some pre-defined upper limit if $\hat{\boldsymbol{\beta}}_\lambda^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol{\beta}}_\lambda$ is so close to zero that the limit would otherwise be exceeded. Formally $\boldsymbol{\Delta} = \boldsymbol{\lambda}^* - \boldsymbol{\lambda}$ is an ascent direction for l_r, by Taylor's theorem and the fact that $\Delta^\mathsf{T}\partial l_r/\partial\boldsymbol{\lambda} > 0$, unless $\boldsymbol{\lambda}$ is already a turning point of l_r. To formally guarantee that the update increases l_r requires step-length control, for example, we use the update $\boldsymbol{\delta} = \boldsymbol{\Delta}/2^k$, where k is the smallest integer ≥ 0 such that $l_r(\boldsymbol{\lambda} + \boldsymbol{\delta}) > l_r(\boldsymbol{\lambda})$.

Two terms in the update have the potential to be of $O(p^3)$ floating point cost, but $\mathrm{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}$ can re-use the Cholesky factor of $\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_\lambda$, which is anyway required to estimate $\hat{\boldsymbol{\beta}}_\lambda$, while the block diagonal nature of $\mathbf{S}_\lambda$ means that in reality $\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j)$ has $O(q_j^3)$ computational cost, where q_j ($\ll p$, typically) is the number of coefficients affected by $\mathbf{S}_j$. Under the conditions of the original Fellner–Schall proposal, $\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) = \mathrm{rank}(\mathbf{S}_j)/\lambda$ and we recover exactly the Fellner–Schall update, albeit with a slightly more computationally tractable expression. The update relies on the following (stated for a slightly more general $\mathbf{S}_\lambda$ than we use), which is the key to the generalization beyond singly penalized smooth terms.

THEOREM 1. *Let* $\mathbf{B}$ *be a positive definite matrix and* $\mathbf{S}_\lambda$ *be a positive semi-definite matrix of the same dimension, parameterized by* $\boldsymbol{\lambda}$*, and with a null space that is independent of the value of* $\boldsymbol{\lambda}$*. Let positive semi-definite matrix* $\mathbf{S}_j$ *denote the derivative of* $\mathbf{S}_\lambda$ *with respect to* λ_j*. Then* $tr(\mathbf{S}_\lambda^- \mathbf{S}_j) - tr\{(\mathbf{B} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\} > 0$.

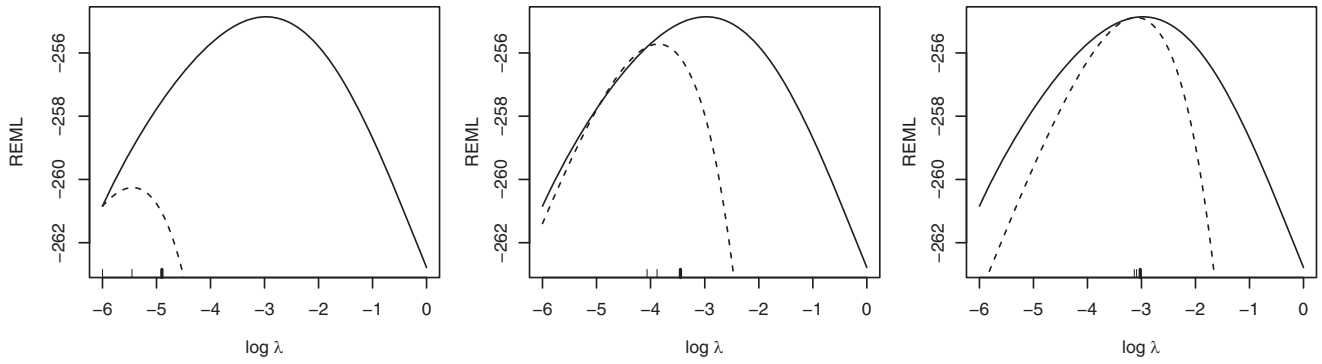

Figure 2. Alternate steps of update (3) for a rank 20 cubic spline smoother of Gaussian data. Each panel shows the log restricted likelihood as a continuous curve, while the EM Q-function is plotted as a dashed curve, shifted to match the log restricted likelihood at each step's start. The two thin ticks on the x axis show the start of the step and the maximum of the Q function. The thick black tick is update (3).

Proof. Let $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}$ be the eigen-decomposition of $\mathbf{B}$. If $\mathbf{S}'_\lambda = \mathbf{\Lambda}^{-1/2}\mathbf{U}^{\mathsf{T}}\mathbf{S}_\lambda\mathbf{U}\mathbf{\Lambda}^{-1/2}$ while $\mathbf{S}'_j = \mathbf{\Lambda}^{-1/2}\mathbf{U}^{\mathsf{T}}\mathbf{S}_j\mathbf{U}\mathbf{\Lambda}^{-1/2}$ then it follows that $\mathrm{tr}\{(\mathbf{B}+\mathbf{S}_\lambda)^{-1}\mathbf{S}_j\} = \mathrm{tr}\{(\mathbf{I}+\mathbf{S}'_\lambda)^{-1}\mathbf{S}'_j\}$, while $\mathrm{tr}(\mathbf{S}_\lambda^{-}\mathbf{S}_j) = \mathrm{tr}(\mathbf{S}_\lambda'^{-}\mathbf{S}'_j)$, where $\mathbf{S}_\lambda'^{-} = \mathbf{\Lambda}^{1/2}\mathbf{U}^{\mathsf{T}}\mathbf{S}_\lambda^{-}\mathbf{U}\mathbf{\Lambda}^{1/2}$. Now form the second eigen-decomposition $\mathbf{S}'_\lambda = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathsf{T}}$. We have that $\mathrm{tr}\{(\mathbf{I}+\mathbf{S}'_\lambda)^{-1}\mathbf{S}'_j\} = \mathrm{tr}\{(\mathbf{I}+\mathbf{D})^{-1}\mathbf{V}^{\mathsf{T}}\mathbf{S}'_j\mathbf{V}\}$, while $\mathrm{tr}(\mathbf{S}_\lambda'^{-}\mathbf{S}'_j) = \mathrm{tr}(\mathbf{D}^{-}\mathbf{V}^{\mathsf{T}}\mathbf{S}'_j\mathbf{V})$. Let s_i denote the i^{th} diagonal element of $\mathbf{V}^{\mathsf{T}}\mathbf{S}'_j\mathbf{V}$. By the conditions of the theorem the null space of $\mathbf{S}_\lambda$ is independent of $\boldsymbol{\lambda}$, and hence $s_i = 0$ if $D_{ii} = 0$. So if $M = \{i : s_i \neq 0\}$, $\mathrm{tr}(\mathbf{S}_\lambda^{-}\mathbf{S}_j) = \sum_{i\in M} s_i/D_{ii}$ while $\mathrm{tr}\{(\mathbf{B}+\mathbf{S}_\lambda)^{-1}\mathbf{S}_j\} = \sum_{i\in M} s_i/(D_{ii}+1)$. Since all the D_{ii} in the summations are positive, by the positive semi-definiteness of $\mathbf{S}_\lambda$ and the definition of M, then the terms in the second summation are each smaller than the corresponding term in the first, and the result is proved.

Remark 1. A similar result also follows if $\mathbf{B}$ is positive semi-definite, but $\mathbf{B}+\mathbf{S}_\lambda$ is positive definite. Define $\mathbf{B}_\delta = \mathbf{B} + \delta\mathbf{I}$ for $\delta \geq 0$ and $\Delta(\delta) = \mathrm{tr}(\mathbf{S}_\lambda^{-}\mathbf{S}_j) - \mathrm{tr}\{(\mathbf{B}_\delta+\mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}$. $\Delta(\delta)$ is continuous with continuous finitely bounded derivative w.r.t. δ for any $\delta \geq 0$ and by Theorem 1 $\Delta(\delta) > 0$ for any $\delta > 0$, hence $\Delta(0) \geq 0$. This is relevant for random effects models for which $\mathbf{X}$ may be rank deficient, while $\mathbf{X}^{\mathsf{T}}\mathbf{X}+\mathbf{S}_\lambda$ is not.

The variance parameter, σ^2, can be estimated directly for any $\boldsymbol{\lambda}$ by setting the derivative of l_r with respect to σ^2 to zero and solving to obtain $\hat\sigma^2 = \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2/[n - \mathrm{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\mathbf{S}_\lambda)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\}]$, which is then substituted for σ^2 in (3).

3. Comparison with the EM Algorithm and Newton Optimization

The update (3) can be viewed as a crude approximation to an EM update (Dempster et al., 1977). Specifically, the EM Q-function for model (2) has the form

$$Q_{\lambda'}(\boldsymbol{\lambda}) = -\frac{\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\lambda'}\|^2 + \hat{\boldsymbol{\beta}}_{\lambda'}^{\mathsf{T}}\mathbf{S}_\lambda\hat{\boldsymbol{\beta}}_{\lambda'}}{2\sigma^2} + \log|\mathbf{S}_\lambda/\sigma^2|_+/2 - \mathrm{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\mathbf{S}_{\lambda'})^{-1}\mathbf{S}_\lambda\}/2, \quad (4)$$

and (3) would be the exact maximizer of Q, if $\mathrm{tr}(\mathbf{S}_\lambda^{-}\mathbf{S}_j) - \mathrm{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\mathbf{S}_{\lambda'})^{-1}\mathbf{S}_j\} \propto 1/\lambda_j$.

In fact, update (3) systematically makes larger changes to $\boldsymbol{\lambda}$ than the EM update, as illustrated in Figure 2. For insight into why this happens, consider updating a single λ_j relating to a block $\lambda_j\mathbf{S}_j$ of $\mathbf{S}_\lambda$, so that $\mathrm{tr}(\mathbf{S}_\lambda^{-}\mathbf{S}_j) = k/\lambda_j$, where $k = \mathrm{rank}(\mathbf{S}_j)$. Then defining $\gamma = \mathrm{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\mathbf{S}_{\lambda'})^{-1}\mathbf{S}_j\}$ and $b = \hat{\boldsymbol{\beta}}_{\lambda'}^{\mathsf{T}}\mathbf{S}_j\hat{\boldsymbol{\beta}}_{\lambda'}/\sigma^2$, (3) seeks λ_j to solve

$$k/\lambda_j = b + \gamma\lambda'_j/\lambda_j, \quad (5)$$

whereas an EM step seeks λ_j to solve

$$k/\lambda_j = b + \gamma. \quad (6)$$

If $k/\lambda_j > b+\gamma$ then λ_j has to be increased from λ'_j under either update. It has to be increased by more under (3), because $\gamma\lambda'_j/\lambda_j$ decreases monotonically from γ as λ_j increases from λ'_j. A similar argument shows that, if $k/\lambda_j < b+\gamma$, then the required reduction in λ_j is larger under (3) than under EM. Figure 3 shows the EM update root finding problem as a dashed curve, and the update (3) root finding problem as a solid curve, for the same set up illustrated in Figure 2.

Figure 3 also illustrates the equivalent problem for the restricted marginal likelihood itself, which can be viewed as solving the same problem as the EM update, but with both b and γ being functions of λ: the dependence of b on λ is indirect via $\hat{\boldsymbol{\beta}}_\lambda$, but the dependence of γ is direct. This suggests using an accelerated EM update seeking to solve

$$k/\lambda_j = b + \gamma(\lambda_j), \quad (7)$$

where $\gamma(\lambda_j) = \mathrm{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X}+\mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}$. This takes longer steps than the original EM update because, like $\gamma\lambda'_j/\lambda_j$, $\gamma(\lambda_j)$ decreases monotonically from γ as λ_j increases form λ'_j (see the dashed curve in Figure 3). Update (3) also results in longer update steps than this accelerated EM step, as Figure 3 suggests and the theorem following demonstrates.

THEOREM 2. *Consider updating a single λ_j corresponding to a diagonal block $\lambda_j\mathbf{S}_j$ of $\mathbf{S}_\lambda$. Update (3) takes a longer step than the equivalent accelerated EM update.*

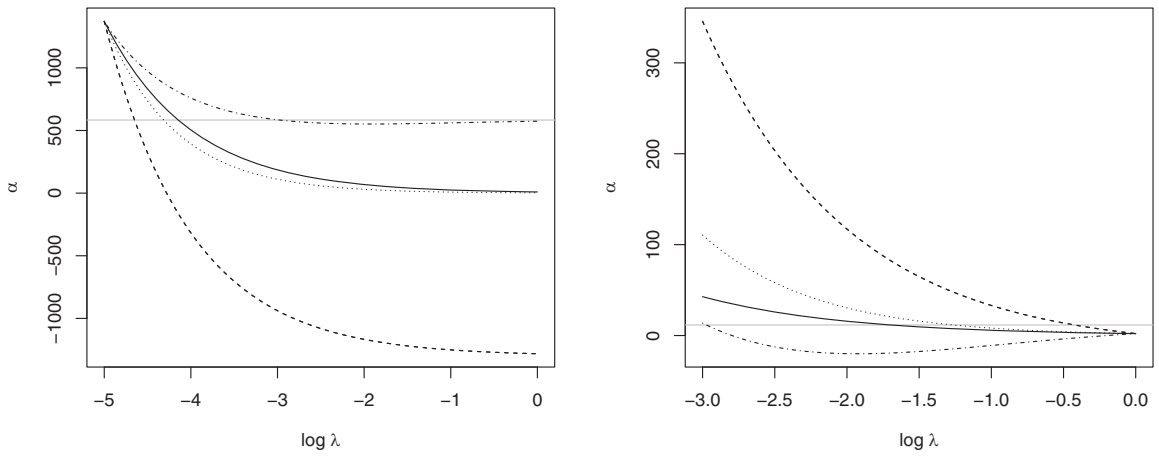

Figure 3. Illustration of the root finding problem corresponding to the various updates discussed in Section 3, for the same modeling problem underlying Figure 2. The gray horizontal line is the constant b. The right plot corresponds to $\log\lambda'_j = -5$ and the right to $\log\lambda'_j = 0$. The EM update corresponds to the point at which the dashed curve crosses the b line: root finding problem (6). The accelerated EM update corresponds to where the dotted curve crosses the b line: root finding problem (7). Update (3) corresponds to where the solid curve crosses the b line: root finding problem (5). The REML optimum is where the dot-dashed curve crosses the b line.

Proof. Under the stated conditions $\mathrm{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) = k/\lambda_j$ where $k = \mathrm{rank}(\mathbf{S}_j)$. Let $\gamma(\lambda_j) = \mathrm{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}$ and $\alpha(\lambda_j) = k/\lambda_j - \gamma(\lambda_j)$. The accelerated EM step seeks λ_j such that $\alpha(\lambda_j) = b$ where $b = \hat{\boldsymbol{\beta}}^\mathsf{T}_{\lambda'}\mathbf{S}_j\hat{\boldsymbol{\beta}}_{\lambda'}/\sigma^2$, increasing λ_j if $\alpha(\lambda_j) > b$ and decreasing λ_j if $\alpha(\lambda_j) < b$. Update (3) is exactly equivalent to seeking λ_j such that $\alpha'(\lambda_j) = b$, where $\alpha'(\lambda_j) = k/\lambda_j - \gamma'(\lambda_j)$ and $\gamma'(\lambda_j) = \gamma(\lambda'_j)\lambda'_j/\lambda_j$. By definition $\alpha'(\lambda'_j) = \alpha(\lambda'_j)$, so to prove the result it suffices to prove that $\alpha'(\lambda_j) > \alpha(\lambda_j)$ when $\lambda_j > \lambda'_j$ and $\alpha'(\lambda_j) < \alpha(\lambda_j)$ when $\lambda_j < \lambda'_j$. Canceling k/λ_j terms this is equivalent to proving $\gamma'(\lambda_j) < \gamma(\lambda_j)$ when $\lambda_j > \lambda'_j$ and $\gamma'(\lambda_j) > \gamma(\lambda_j)$ when $\lambda_j < \lambda'_j$. Now let $\mathbf{S}_{-j} = \sum_{i\neq j}\lambda_i\mathbf{S}_i$, and let $\mathbf{B}$ be any matrix such that $\mathbf{B}^\mathsf{T}\mathbf{B} = \mathbf{S}_{-j}$. Consider the QR decomposition $(\mathbf{X}^\mathsf{T}, \mathbf{B}^\mathsf{T})^\mathsf{T} = \mathbf{QR}$ and form the symmetric positive semi-definite eigen-decomposition $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T} = \mathbf{R}^{-T}\mathbf{S}_j\mathbf{R}^{-1}$. Routine manipulation shows that $\gamma(\lambda_j) = \sum_{i=1}^k \Lambda_{ii}/(1+\lambda_j\Lambda_{ii})$. It follows that $\gamma'(\lambda_j) = \sum_{i=1}^k \Lambda_{ii}/(\lambda_j/\lambda'_j + \lambda_j\Lambda_{ii})$. Hence $\gamma'(\lambda_j) < \gamma(\lambda_j)$ if $\lambda_j > \lambda'_j$ and $\gamma'(\lambda_j) > \gamma(\lambda_j)$ if $\lambda_j < \lambda'_j$, proving the result.

Taking longer steps than a plain or accelerated EM algorithm would be of limited utility if those steps overshot the maximum of the restricted likelihood and require repeated step-length control, especially when close to the optimum. In practice such overshoot does not occur. The following theorem offers some insight into the reasons. It requires two technical assumptions.

Assumption 1: If $\mathbf{Q}_1$ is the first n rows of $\mathbf{Q}$ from the proof of Theorem 2 and $\mathbf{a} = \mathbf{U}^\mathsf{T}\mathbf{Q}_1^\mathsf{T}\mathbf{y}$, then $a_i^2 = O_p(n^{\beta_i})$ where $\beta_i > 0$ for all i, and β_i is the minimum β' such that $a_i^2 = O_p(n^{\beta'})$.

The assumption is less obscure than it at first appears. Let $\hat{\boldsymbol{\mu}}_0 = \mathbf{X}\hat{\boldsymbol{\beta}}$, when $\lambda_j = 0$, so that $\hat{\boldsymbol{\mu}}_0 = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_j)^{-1}\mathbf{X}\mathbf{y} = \mathbf{Q}_1\mathbf{U}\mathbf{U}^\mathsf{T}\mathbf{Q}_1^\mathsf{T}\mathbf{y}$. Now let $\mathbf{C} = \mathbf{Q}_1\mathbf{U}$, so that $\hat{\boldsymbol{\mu}}_0 = \mathbf{C}\mathbf{C}^\mathsf{T}\mathbf{y} = \sum_i \hat{\boldsymbol{\mu}}_i$, where $\hat{\boldsymbol{\mu}}_i = \mathbf{C}_{\cdot i}\mathbf{C}_{\cdot i}^\mathsf{T}\mathbf{y}$. The assumption that $\mathbf{y}^\mathsf{T}\hat{\boldsymbol{\mu}}_i = O_p(n)$, and that 1 is the lowest power of n for which this holds, is essentially equivalent to assuming that no model component is orthogonal to $\mathbb{E}(\mathbf{y})$, but since $a_i = \mathbf{C}_{\cdot i}^\mathsf{T}\mathbf{y}$ it is also equivalent to Assumption 1 with $\beta_i = 1$.

Assumption 2: In the notation of the proof of Theorem 2, $\lambda\Lambda_{ii} = O_p(n^{\alpha_i})$, where α_i is an unknown real constant and is the minimum α'_i such that $\lambda\Lambda_{ii} = O_p(n^{\alpha'_i})$.

This simply assumes that each $\lambda\Lambda_{ii}$ has some polynomial dependence on n, but not that we know what it is.

THEOREM 3. *Let the setup be as Theorem 2 and $\hat{\lambda}_j$ denote the maximizer of the restricted likelihood with respect to λ_j. Given assumptions 1 and 2, for an initial λ_j sufficiently close to $\hat{\lambda}_j$, then as $n \to \infty$ the update, λ_j^*, given by (3) is either between λ_j and $\hat{\lambda}_j$, or tends to $\hat{\lambda}_j$.*

Proof. Dropping the subscript j, let $\rho = \log\lambda$, and let λ denote the j^{th} smoothing parameter at the start of the updates. Consider again the root finding problems equivalent to the update (3) and to maximization of the restricted marginal likelihood. Applying Taylor's theorem to the components of these root finding problems, we have that, for λ sufficiently close to $\hat{\lambda}$,

$$\frac{k}{\lambda} - \frac{k}{\lambda}(\hat{\rho} - \rho) - \gamma(\lambda) - \left(\frac{\mathrm{d}\gamma}{\mathrm{d}\rho} + \frac{\mathrm{d}b}{\mathrm{d}\rho}\right)(\hat{\rho} - \rho) = b(\lambda),$$

where the derivatives are evaluated at the initial value, ρ, and

$$\frac{k}{\lambda} - \frac{k}{\lambda}(\rho^* - \rho) - \gamma(\lambda) + \gamma(\lambda)(\rho^* - \rho) = b(\lambda).$$

So, if $\gamma(\lambda) \leq \delta(\lambda) = -(\mathrm{d}\gamma/\mathrm{d}\rho + \mathrm{d}b/\mathrm{d}\rho)$, then $\lambda < \lambda^* \leq \hat{\lambda}$. Also, if $\lambda\gamma(\lambda) \to 0$ and $\lambda\delta(\lambda) \to 0$, as $n \to \infty$, then $|\rho^* - \hat{\rho}| \to 0$.

Now consider the actual behavior of $\gamma(\lambda)$ and $\delta(\lambda)$. Using the QR and eigen-decomposition steps from the proof of

Theorem 2, some routine manipulation yields

$$\gamma(\lambda) = \frac{1}{\lambda}\sum_i \frac{\lambda\Lambda_{ii}}{1+\lambda\Lambda_{ii}}$$

and

$$\delta(\lambda) = \frac{1}{\lambda}\sum_i \frac{\lambda\Lambda_{ii}}{1+\lambda\Lambda_{ii}}\left\{\frac{(1+2a_i^2)\lambda\Lambda_{ii}+\lambda^2\Lambda_{ii}^2}{1+2\lambda\Lambda_{ii}+\lambda^2\Lambda_{ii}^2}\right\}.$$

So the i^{th} term of δ will be larger that the i^{th} term of γ if $\lambda\Lambda_{ii} > (2a_i^2-1)^{-1}$: if $\lambda\Lambda_{ii} = O_p(n^{\alpha_i})$ in accordance with Assumption 2, then this dominance occurs in the $n\to\infty$ limit when $\alpha_i > -\beta_i$. Furthermore, if $\alpha_i < -\beta_i/2$, then the i^{th} terms of $\gamma(\lambda)\lambda$ and $\delta(\lambda)\lambda$ both tend to zero in the large sample limit. So in the large sample limit, sufficiently close to $\hat\lambda$, there are only two non-exclusive possibilities: $\gamma(\lambda) < \delta(\lambda)$ so that λ^* lies between λ and $\hat\lambda$, and/or all the terms in the $\delta(\lambda)$ and $\gamma(\lambda)$ summations tend to zero, so that $\lambda\delta(\lambda), \lambda\gamma(\lambda)\to 0$ and $|\lambda^*-\hat\lambda|\to 0$.

The solution of the linearized root-finding problem corresponding to the restricted likelihood maximization is the Newton method update. Since the theorem indicates that the updates take steps no longer than Newton's method, then a corollary of Theorem 3 is that iteration of update (3) will converge no faster than Newton's method, asymptotically, and may converge more slowly. Obviously, this slower convergence in terms of number of step required is offset by the fact that less computation is required per steps.

4. Beyond the Linear Gaussian Case and Alternatives to REML

Now consider replacing the Gaussian log likelihood with another log likelihood, l, meeting the Fisher regularity conditions, so that the improper log joint density becomes

$$\log f_\lambda(\mathbf{y},\boldsymbol\beta) = l(\boldsymbol\beta) - \boldsymbol\beta^\mathsf{T}\mathbf{S}_\lambda\boldsymbol\beta/2 + \log|\mathbf{S}_\lambda|_+ + c,$$

and in the large sample limit $\boldsymbol\beta|\mathbf{y}\sim N(\hat{\boldsymbol\beta}_\lambda, \mathbf{V}_\lambda)$ where $\mathbf{V}_\lambda^{-1} = \mathcal{H}_\lambda$ or $\mathbb{E}\mathcal{H}_\lambda$ and $\mathcal{H}_\lambda = -\partial^2 l/\partial\boldsymbol\beta\partial\boldsymbol\beta^\mathsf{T} + \mathbf{S}_\lambda$. Newton's method can be used to find $\hat{\boldsymbol\beta}_\lambda$, with the usual modifications to guarantee convergence (e.g., Wood, 2015, Section 5.1.1). Following Wood et al. (2016), the log Laplace approximate marginal likelihood in this case is conveniently expressed as

$$l_r = l(\hat{\boldsymbol\beta}_\lambda) - \hat{\boldsymbol\beta}_\lambda^\mathsf{T}\mathbf{S}_\lambda\hat{\boldsymbol\beta}_\lambda/2 + \log|\mathbf{S}_\lambda|_+/2 - \log|\mathcal{H}_\lambda|/2 + c.$$

Defining $\mathbf{H} = -\partial^2 l/\partial\boldsymbol\beta\partial\boldsymbol\beta^\mathsf{T}$, we have

$$\frac{\partial l_r}{\partial\lambda_j} = -\hat{\boldsymbol\beta}_\lambda^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol\beta}_\lambda/2 + \mathrm{tr}(\mathbf{S}_\lambda^-\mathbf{S}_j)/2 - \mathrm{tr}\{\mathbf{V}_\lambda\mathbf{S}_j\}/2 - \mathrm{tr}\{\mathbf{V}_\lambda\partial\mathbf{H}/\partial\lambda_j\}/2.$$

The direct dependence of $\mathbf{H}$ on λ_j is inconvenient. However, the PQL and performance oriented iteration methods for $\boldsymbol\lambda$ estimation of Breslow and Clayton (1993) and Gu (1992) both neglect the dependence of $\mathbf{H}$ on $\boldsymbol\lambda$, on the basis that it anyway tends to zero in the large sample limit. If we follow these precedents, then the development follows the Gaussian case and the update is

$$\lambda_j^* = \frac{\mathrm{tr}(\mathbf{S}_\lambda^-\mathbf{S}_j) - \mathrm{tr}\{\mathbf{V}_{\lambda'}\mathbf{S}_j\}}{\hat{\boldsymbol\beta}_\lambda^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol\beta}_\lambda}\lambda_j. \tag{8}$$

If $\partial^2 l/\partial\boldsymbol\beta\partial\boldsymbol\beta^\mathsf{T}$ is independent of $\boldsymbol\lambda$ at finite sample size, as is the case for some distribution – link function combinations in a generalized linear model setting, then the update is guaranteed to increase l_r under step size control, but otherwise this is not the case, and in practice the $\boldsymbol\lambda$ estimate no longer exactly maximizes l_r.

Theorem 1, required to guarantee that $\lambda_j^* > 0$, will hold if $\mathbf{V}_\lambda$ is based on the expected Hessian of the negative log likelihood, but if it is based on the observed Hessian, then this must be positive definite for the theorem to hold. Hence, if the observed Hessian is not positive definite, then the expected Hessian, or a suitable nearest positive definite matrix to the observed Hessian, should be substituted.

As in the Gaussian case, a link to the EM update can again be established via an approximate Q function, obtained by taking a second order Taylor expansion of l around $\hat{\boldsymbol\beta}_\lambda$, and using the large sample distribution of $\boldsymbol\beta|\mathbf{y}$:

$$Q^*_{\lambda'}(\boldsymbol\lambda) = l(\hat{\boldsymbol\beta}_{\lambda'}) - \hat{\boldsymbol\beta}_{\lambda'}^\mathsf{T}\mathbf{S}_\lambda\hat{\boldsymbol\beta}_{\lambda'}/2 + \log|\mathbf{S}_\lambda|_+/2 - \mathrm{tr}(\mathbf{V}_{\lambda'}\mathbf{S}_\lambda)/2 - \mathrm{tr}(\mathbf{V}_{\lambda'}\partial^2 l/\partial\boldsymbol\beta\partial\boldsymbol\beta^\mathsf{T})/2.$$

The final term is then neglected, again following the PQL type assumption.

In the case of a penalized generalized linear model, the general update (8) becomes

$$\lambda_j^* = \phi\frac{\mathrm{tr}(\mathbf{S}_\lambda^-\mathbf{S}_j) - \mathrm{tr}\{(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}+\mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}}{\hat{\boldsymbol\beta}_\lambda^\mathsf{T}\mathbf{S}_j\hat{\boldsymbol\beta}_\lambda}\lambda_j,$$

where $\mathbf{W}$ is the diagonal matrix of weights at convergence of the usual penalized iteratively re-weighted least squares iteration used to find $\hat{\boldsymbol\beta}_\lambda$, and ϕ is the scale parameter, which can be substituted by an estimate using the obvious equivalent of $\hat\sigma^2$. Under the original Fellner–Schall restrictions, this update corresponds to the Schall update for generalized linear mixed models.

Given the Bayesian motivation for the smoothing penalty inducing Gaussian priors that leads to the restricted marginal likelihood criterion, the preceding method can be viewed as an empirical Bayes procedure. However, similarly convenient updates are readily computed for directly frequentist criteria. For example, the AIC criteria, $-2l+2\tau$ where $\tau = \mathrm{tr}(\mathbf{V}_\lambda\mathbb{E}\mathbf{H})$ (and $\mathbf{V}_\lambda$ is based on the expected Hessian), leads to the update

$$\lambda_j^* = \frac{\partial\tau}{\partial\lambda_j}\left(\frac{\partial l}{\partial\boldsymbol\beta}\frac{\mathrm{d}\hat{\boldsymbol\beta}}{\mathrm{d}\lambda_j}\right)^{-1}\lambda_j,$$

where $\partial\tau/\partial\lambda_j = -\mathrm{tr}(\mathbf{V}_\lambda\mathbf{S}_j\mathbf{V}_\lambda\mathbb{E}\mathbf{H})$ and $\mathrm{d}\hat{\boldsymbol\beta}/\mathrm{d}\lambda_j = -(\mathbf{H}+\mathbf{S}_\lambda)^{-1}\mathbf{S}_j\hat{\boldsymbol\beta}$. For models in which a deviance, D, can be sensibly

defined, such as GLMs, then an alternative is the GCV criterion $nD/(n-\tau)^2$, which yields the update,

$$\lambda_j^* = \frac{-2D}{n-\tau}\frac{\partial \tau}{\partial \lambda_j}\left(\frac{\partial D}{\partial \boldsymbol{\beta}}\frac{\mathrm{d}\hat{\boldsymbol{\beta}}}{\mathrm{d}\lambda_j}\right)^{-1}\lambda_j.$$

5. Simple Examples

First, consider a simple Gaussian model of the motorcycle data from Silverman (1985), available in the MASS package (Venables and Ripley, 2002) in R (R Core Team, 2014). The data are accelerations of the head of a crash test dummy against time. An adaptive smooth, as described in Wood (2011), is appropriate for smoothing the acceleration data against time, with the degree of smoothness of a P-spline (Eilers and Marx, 1996) varying smoothly with time. The smooth used has five smoothing parameters with the penalties acting on overlapping subsets of the 40 model coefficients, thereby violating the structural conditions on S_λ required by previously published Fellner-Shall iterations.The smooth was estimated using the method presented here and by the quasi-Newton variant of the method of Wood (2011) (so both methods have the same leading order computational cost per iteration). Starting from all smoothing parameters set to 1, and without step length control, the new method converged in 39 steps, as against 32 for the quasi-Newton method. The fits are identical to graphical accuracy with equal effective degrees of freedom of 12.22. See Figure 4.

The second example is a Cox proportional hazards model for time to recurrence of colon cancer for $n = 929$ patients in a chemotherapy trial (Moertel et al., 1995), available in the survival package (Therneau, 2015) in R. In this case previously published Fellner–Schall methods would only be usable by fitting an equivalent Poisson model to artificial data at an $O(n)$ multiplication of the computational cost, which is impractically uncompetitive with existing methods. This cost inflation is avoided by using update (8). The linear predictor for the Cox regression had parametric effects for whether the colon was perforated or not, obstructed or not, and whether the tumor had adhered to neighboring organs. In addition, a 3 level factor indicated the control group, treatment with one drug of interest or treatment with a drug combination. Smooth effects of age were included separately for males and females along with a smooth effect for number of affected lymph nodes. For this example, the new iteration, without step length control, converged in 15 steps, compared to 16 steps for direct quasi-Newton optimization using the methods of Wood et al. (2016). The parametric model coefficients differ only in the 4th significant digit, while differences in the estimated smooth effects are also small, as shown in Figure 4.

Finally, consider a model fitting problem that is completely infeasible using pre-existing methods, but is of a type that is increasingly common in many areas of high throughput

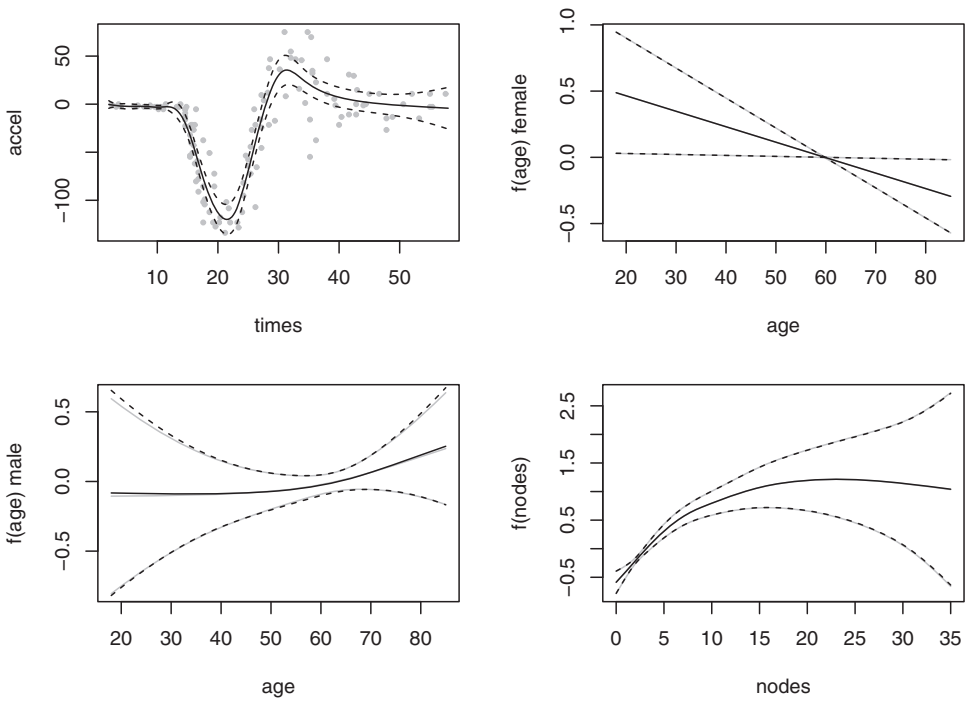

Figure 4. Top left: an adaptive smoother fitted to the motorcycle data using the proposed method. A fit by direct restricted marginal likelihood maximization is indistinguishable. Previous Fellner–Schall methods could not be used for this example, as it lacks the required special structure of S_λ. Other panels: estimated smooth effects for the colon cancer survival model. The estimates using full Laplace approximate restricted marginal partial likelihood are shown in gray, with the new method estimates overlaid in black. The intervals shown in all panels are 95% Bayesian intervals as discussed in Wood (2006), for example.

science:

$$y_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + f_{a(i)}(z_{1i}) + f_{b(i)}(z_{2i}) + f_{s(i)}(z_{3i}) + \epsilon_i,$$

where the $f_{\cdot}$ are smooth functions, x_j and z_j are covariates, and $s(i)$ indicates the subject to which the i^{th} observation belongs. Each subject is associated with one level of each of two crossed factors a and b: $a(i)$ and $b(i)$ indicate the levels of a and b for observation i. There are 10000 subjects each with 110 observations, a and b have 50 and 40 levels, and the smooth functions are each represented with rank 10 spline bases. So the model has 10,092 smooth functions and 100,921 coefficients for 1,100,000 observations. Five smoothing parameters were used, one each for f_1 and f_2, one for the 50 f_a smooths and so on. Estimation of the model coefficients is feasible because the model matrix and penalties are highly sparse. Specifically, consider the sparse Cholesky decomposition $\mathbf{L}^{\mathsf{T}}\mathbf{L} = \mathbf{P}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{S}_\lambda)\mathbf{P}^{\mathsf{T}}$, where $\mathbf{P}$ represents a sparsity preserving pivoting operation (Davis, 2006). Then $\hat{\boldsymbol{\beta}} = \mathbf{P}^{\mathsf{T}}\mathbf{L}^{-1}\mathbf{L}^{-T}\mathbf{P}\mathbf{X}^{\mathsf{T}}\mathbf{y}$, computation of which takes seconds on a mid range laptop (the equivalent computation with dense matrices would take days, even assuming the required terabytes of memory were available). Even with a sparse $\mathbf{X}$, the complexity of the terms involved in conventional Newton based smoothing parameter estimation methods causes unavoidable loss of sparsity (infill) rendering the methods computationally infeasible. In contrast, our Fellner Schall update can be computed without infill. The only potentially difficult term, $\mathrm{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}$, is actually the sum of squares of the elements of the matrix $\mathbf{B}$, where $\mathbf{B} = \mathbf{L}^{-T}\mathbf{P}\mathbf{D}_j$, and $\mathbf{D}_j$ is any sparse matrix such that $\mathbf{D}_j\mathbf{D}_j^{\mathsf{T}} = \mathbf{S}_j$. $\mathbf{D}_j$ can readily be created alongside $\mathbf{S}_j$. Furthermore, $\mathrm{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\} = p - \sum_j \lambda_j \mathrm{tr}\{(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \mathbf{S}_\lambda)^{-1}\mathbf{S}_j\}$ gives the model effective degrees of freedom, required for estimating σ^2.

Figure 5 shows the model estimates when using (3) to iteratively estimate smoothing parameters. Estimation using the `Matrix` library in R took less than 10 minutes with a single CPU core of a mid range laptop. A rough estimate is that a high end workstation using 10 CPU cores would take over 2 months to fit the same model using alternative methods.

6. A Tweedie Location, Scale and Shape Model for Mackerel

We now return to the introduction's motivating example of modeling mackerel (*Scomber scombrus*) egg densities. The data consist of counts of eggs in samples taken from the water column at the sampling stations shown in Figure 1. Available covariates are temperature and salinity at 20 m depth, water volume sampled (an offset), spatial location as longitude and latitude (converted to km east and km north), the identity of the ship collecting the data, and the sea bed depth.

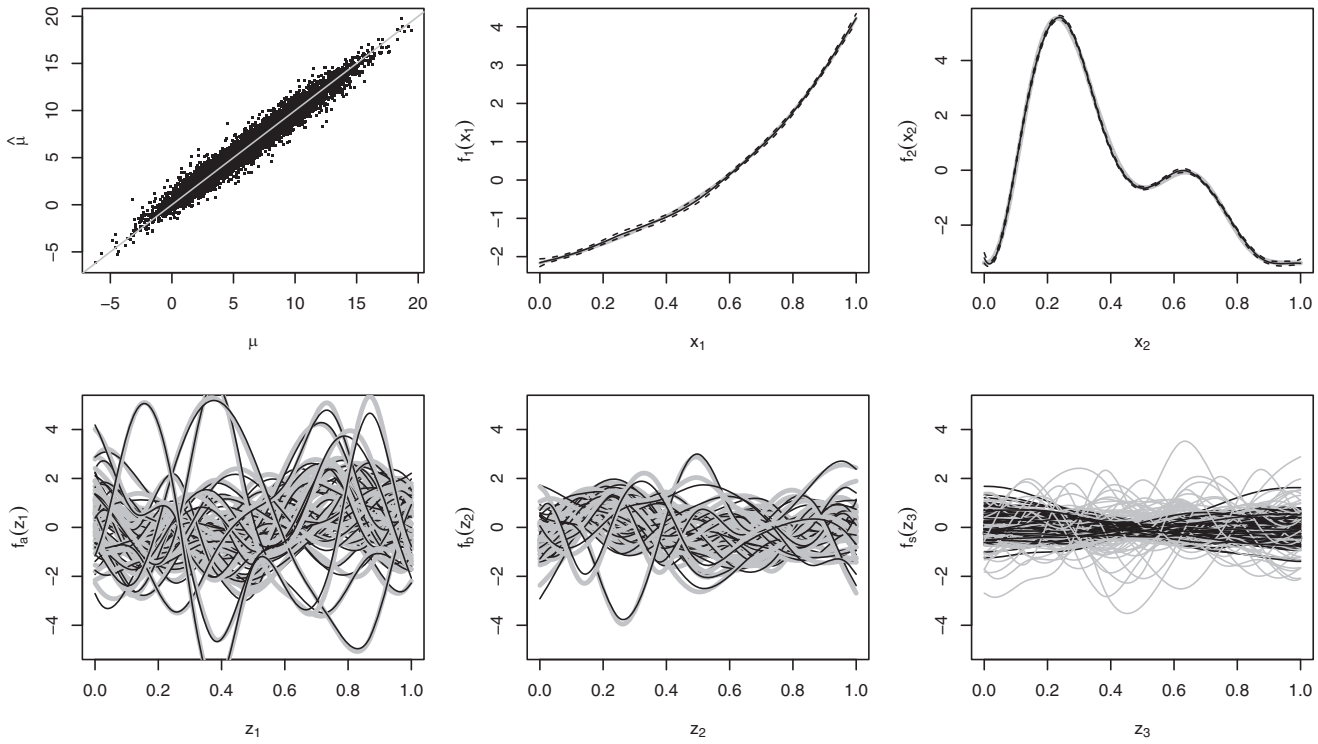

Figure 5. Results of the very large sparse model fit discussed in Section 5. The model has 10092 smooths, with 100921 coefficients, 5 smoothing parameters, and is estimated from 1.1 million (simulated) observations, using less than 10 CPU minutes and 3Gb of memory on a mid-range laptop computer (Dell E6230). Top left: Scatter plot of 1% random sample of fitted values against simulation truth. Top middle: first main effect smooth with 95% intervals, overlaid on simulation truth in gray. Top right: as top middle for second main effects smooth. Bottom row: estimates in black, truth in gray. Bottom left: the 50 smooths conditional on factor a. Bottom middle: the 40 smooths conditional on factor b. Bottom right: 1% sample of the 10000 subject specific smooths: the noise level was too high to estimate these well.

A common theme with data of this type is that the counts are highly over-dispersed relative to a Poisson distribution, but with a mean variance relationship that is less extreme than that suggested by a negative binomial distribution (see e.g., Wood, 2006, Section 5.4.1). A Tweedie (1984) distribution is often a better model, but it would be useful to allow its shape and scale parameters to vary with covariates. Specifically, the Tweedie distribution assumes that the variance of random variable y_i is related to its mean, μ_i via $\text{var}(y_i) = \phi_i \mu_i^{p_i}$, where ϕ_i and p_i are parameters usually taking one fixed value for all i. For the mackerel data it would be useful to allow p_i and ϕ_i to be smooth functions of covariates, particularly sea bed depth, for example, using the model

$$\begin{aligned} \log(\mu_i) &= g_1(\texttt{lo}_i, \texttt{la}_i) + g_2(\texttt{T20}_i) + g_3(\texttt{S20}_i) \\ &\quad + g_4(\texttt{b.depth}^{1/2}) + b_{s(i)} + \log(\texttt{vol}_i), \\ h(p_i) &= g_5(\texttt{b.depth}^{1/2}), \quad \log(\phi_i) = g_6(\texttt{b.depth}^{1/2}), \\ &\quad \texttt{count}_i \sim \text{Tweedie}(\mu_i, p_i, \phi_i). \end{aligned} \tag{9}$$

The g_k are smooth functions, h is a known link function designed to keep $1 < p < 2$, $s(i)$ indicates which ship collected sample i and $b_{s(i)}$ are independent $N(0, \sigma_b^2)$ random effects. We represented the spatial effect using a rank 150 Duchon spline with first order derivative penalization (see Duchon, 1977; Miller and Wood, 2014), and other terms with rank 10 cubic penalized regression splines. The model can be estimated, *given smoothing parameters*, using the Newton iteration detailed in Wood et al. (2016) and available in R package `mgcv`. The estimation of smoothing parameters using Wood et al. (2016) would require the currently unavailable third and fourth derivatives of the Tweedie density. We therefore estimated the smoothing parameters using the iterative update (8).

Estimation converged in 13 iterations taking 17 seconds (single core of a mid range laptop computer). In comparison, it took 11 seconds to fit a necessarily over-simplified version of the model, with fixed p and ϕ, using the method of Wood et al. (2016) in R package `mgcv`. The AIC for model (9) was 180 lower than for the fixed p and ϕ version, although residual plots (not shown) are reasonable for both models. The estimated spatial smoother is shown in Figure 1b, while the remaining effects are plotted in Figure 6. Notice how the smooth effects of sea depth all have a pronounced peak at around $\sqrt{200}$, corresponding to the edge of the continental

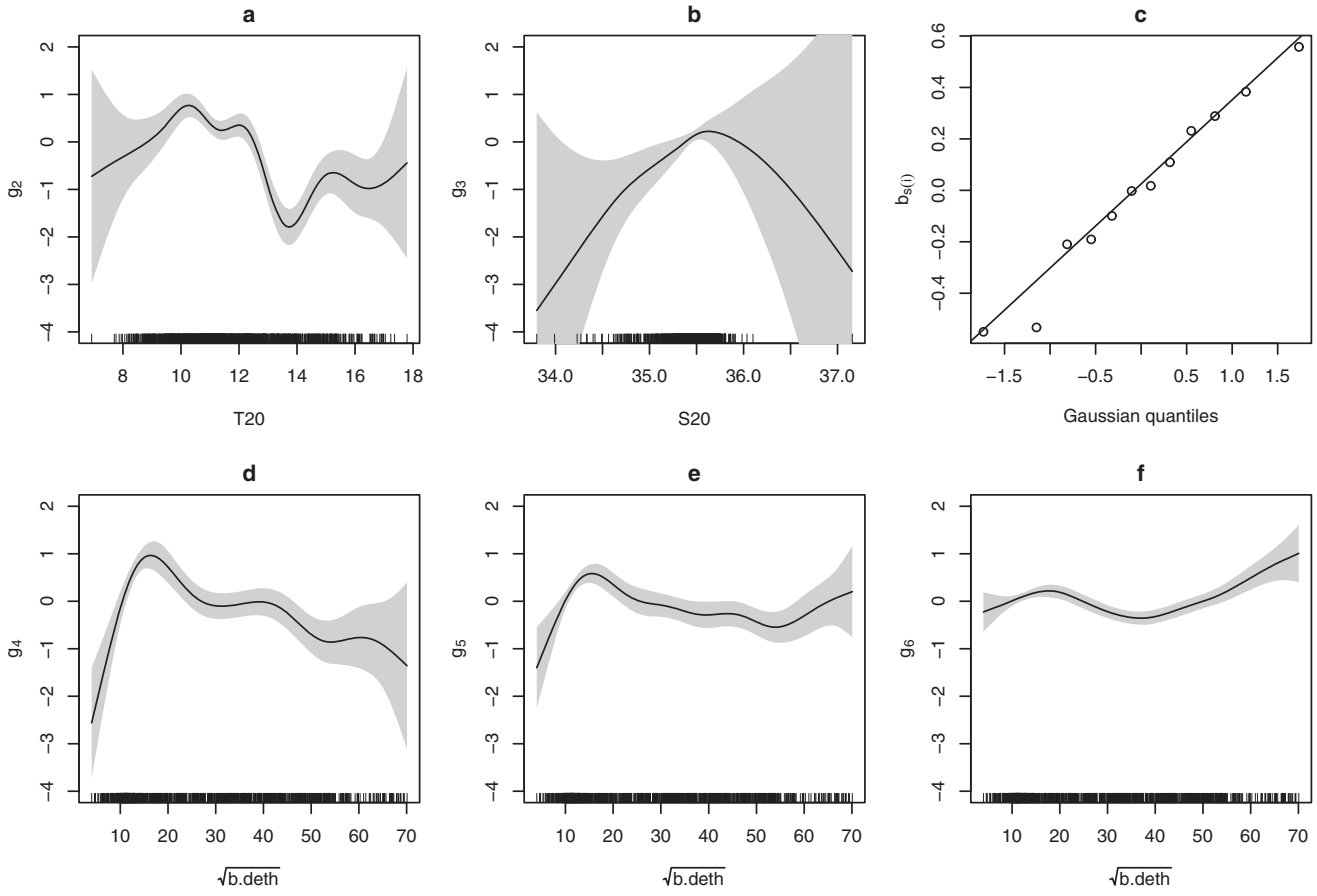

Figure 6. Estimated smooth effects for the Tweedie location scale and shape model of the Mackerel egg survey data discussed in Section 6. Panel c shows a QQ-plot for the predicted ship level random effects. Panels d, e and f are the smooth effects of sea depth for μ, p and ϕ respectively. Notice how they all have a peak close to $\sqrt{200}$, the depth representing the continental shelf edge. The shaded regions are approximate 95% confidence intervals.

shelf. Both egg density and its variability appear to be peaking near the shelf edge.

7. Discussion

Prior to the work reported here, the Fellner–Schall method could only be applied to a subset of the smooth additive models that could be estimated by direct Laplace approximate marginal likelihood maximization. The generalizations introduced here remove this obstacle, and we have also strengthened the theoretical underpinnings of the method. The major advantage of the method is its simplicity: the direct method of Wood et al. (2016) requires evaluation of third or fourth order derivatives of the log likelihood, which are not required by the generalized Fellner–Schall method. In addition, direct optimization of the Laplace approximate marginal likelihood requires nested optimization and implicit differentiation to obtain derivatives of β with respect to λ. Such an approach involves considerable effort if it is to be numerically stable, which is not required by the modified Fellner Schall iteration. The main theoretical cost is that, beyond the Gaussian case, we are forced to make the same simplification that underpins the PQL and performance oriented iteration methods, and neglect the dependence of the Hessian of the log likelihood on the smoothing parameters.

As demonstrated in Sections 5 and 6, our generalized Fellner–Schall method can be applied to cases in which alternative estimation methods would be very difficult to implement, but it also offers advantages in settings which are in principle less numerically taxing. The method can be applied to non-standard smooth models, provided that we can obtain the first and second derivatives of the log-likelihood, which are anyway required for Newton optimization of model coefficients. This greatly simplifies the process of implementing non-standard models for particular applied problems, freeing the modeler from the more onerous aspects of implementation, to concentrate on development of the model itself. To gain insight into the effort saved, the reader might care to compare the expressions for the fourth order and second order derivatives of the generalized extreme value distribution, for example.

Finally, an interesting question raised by the work here, is whether it is possible to reduce the implementation cost even further, by replacing the Hessian of the log-likelihood in the update by a Quasi-Newton approximation, thereby allowing coefficients to be estimated by Quasi-Newton methods, and only requiring first derivatives of the log-likelihood.

8. Supplementary material

The Mackerel data and method are available in R packages `gamair` and `mgcv`, available on CRAN: see `gam`'s `optimizer` argument. A simulation study method comparison is available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

Thanks to Yousra El Bachir for useful comments on an earlier version of this article and to the referees for some very helpful comments. This work was funded by EPSRC grant EP/K005251/1 'Sparse, rank-reduced and general smooth modelling'. The mackerel data are available from ICES Atlantic Anguilla surveys, `http://eggsandlarva.ices.dk`.

References

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Davis, T. A. (2006). *Direct Methods for Sparse Linear Systems*. Philadelphia: SIAM.

Dempster, A. P., Laid, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**, 1–38.

Duchon, J. (1977). Splines minimizing rotation-invariant seminorms in Solobev spaces. In Construction Theory of Functions of Several Variables W. Schemp and K. Zeller (eds.), 85–100. Berlin: Springer.

Dunn, P. K. and Smyth, G. K. (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing* **15** 267–280.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics* **28**, 51–60.

Gu, C. (1992). Cross-validating non-gaussian data. *Journal of Computational and Graphical Statistics* **1**, 169–179.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* **41**, 495–502.

Miller, D. L. and Wood, S. N. (2014). Finite area smoothing with generalized distance splines. *Environmental and Ecological Statistics* **21**, 715–731.

Moertel, C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie,J. A., Tangen, C. M. et al. (1995). Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage iii colon carcinoma: A final report. *Annals of Internal Medicine* **122**, 321–326.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rigby, R. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**, 507–554.

Rigby, R. A. and Stasinopoulos, D. M. (2014). Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. *Statistical Methods in Medical Research* **23**, 318–332.

Rodríguez-Álvarez, M. X., Lee, D.-J. Kneib, T. Durbán, M. and Eilers, P. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing* **25**, 941–957.

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

Schall, R. (1991). Estimation in generalized linear models with random effects *Biometrika* **78**, 719–727.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society B* **47**, 1–53.

Therneau, T. (2015). *A Package for Survival Analysis in S*. http://CRAN.R-project.org/package=survival

Tweedie, M. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions: Proc. Indian Statistical Institute Golden Jubilee International Conference*, 579–604.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton FL: CRC press.

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

Wood, S. N. (2015). Core Statistics. Cambridge: Cambridge University Press.

Wood, S. N., Pya, N. and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models with discussion. *Journal of the American Statistical Association* **111**, 1548–1575.

Received June 2016. *Revised January* 2017.
Accepted January 2017.