Comments on "Smoothing parameter and model selection for general smooth models" by S. N. Wood, N. Pya, and B. Säfken

Thomas W. Yee

September 1, 2016

Thomas W. Yee, Department of Statistics, University of Auckland, Auckland 1142, New Zealand (E-mail: *t.yee@auckland.ac.nz*)

The authors are congratulated for this significant and detailed contribution, which extends the exceedingly popular GAM technique to higher realms of generality and functionality. It is interesting to see how GAMs have developed over the last two decades or so. Not surprisingly, as authors have become more ambitious, the level of sophistication has escalated and this paper is no exception. For a given model, the paper conveys the substantial amount of work needed to implement automatic smoothing parameter selection based on LAML—these involve fourth-order derivatives, unremitting attention to the numerical analysis of computations, and careful programming—so that the code runs robustly and efficiently on real data. Inference and asymptotic properties are also considered here, as well as a framework that allows a rich variety of smoothing based on low-rank penalized splines. It is unfortunate that the vast majority of researchers in the statistical sciences are still content with pen and paper (or word processor), and do not offer the 'full service' of developing new methodology from beginning to end, which includes a user-friendly software implementation that people can use straightaway.

Early work starting in the mid-1980s led by Hastie and Tibshirani developed GAMs based on backfitting, and were largely confined to the exponential family. Smoothing parameter selection was difficult for this. Over the last $1\frac{1}{2}$ decades the first author has led the charge of automating smoothing parameter selection (and dispensing of backfitting), resulting in several methods such as UBRE/GCV optimization in conjunction with PIRLS, ML- and REML-based methods, and now refined LAML-based methods. However, these works were also largely confined to the exponential family, bar the present paper. In my own work I have, in the large part, had the strong conviction of breaking out of the shackles of the exponential family from the outset, and the handling of multiple linear predictors. Doing so really does open one up to a lot more of the statistical universe. Current work with C. Somchit and C. Wild on developing automatic smoothing parameter selection for

the vector generalized additive model (VGAM) class, which is very general, is very briefly described in Section 1.1 and a working implementation should hopefully appear within the next 12 months. This means that we have attempted to reach one of the goals of the present paper (hereafter referred to as 'WPS'), albeit, approaching from a different direction.

The comments below are shared between the paper and some selected supplementary appendices.

1 GENERAL COMMENTS

The authors have done a valiant job developing LAML estimation for general settings and implementing several important regression models. Some computational tricks and inferential by-products have been found along the way. With the possibility to handle multiple additive predictors η_j now, is it possible to constrain some of these linearly? For example, in the case of two of them, can one fit $\eta_1 = a + b\,\eta_2$ where a and b are estimated too? There are several reasons for pursuing this idea. This is a special case of a reduced-rank regression where the overall regression coefficients are subject to a rank restriction. There are several benefits, such as a lower computational cost, increased parsimony of parameters, interpretation in terms of latent variables, and low-dimensional views in the case of a rank-2 model. Some particularly useful regression models arise as special cases, such as a negative binomial regression with variance function $\mu + \delta_1 \mu^{\delta_2}$ (known as the NB-P model in the count literature). In the case of the multinomial logit model, when the number of classes and number of explanatory variables is even moderate then the number of regression coefficients becomes sizeable, hence some form of dimension reduction becomes warranted. Such a model was called the stereotype model by Anderson (1984). Additionally, it would be very useful if $\eta_j = f_j(\nu)$ were developed for regular models, where $\nu = \boldsymbol{c}^T \boldsymbol{x}$ is an optimal linear

combination of the explanatory variables. For this the exponential family case is referred to as a single-index model and it is a special case of a generalized additive index model (GAIM; Chen & Samworth 2016). Some of these ideas are described in Yee (2015).

Section 6 raises the issue of cost–benefit when some simulation results show that occasionally the new method gives only a small improvement in statistical performance relative to some less complicated methods. I have found that some prudence is required when deciding whether a certain complex procedure is worth implementing, especially when it involves a large expenditure of effort. There is a parallel to be seen from the 1980s and 1990s when smoothing was a very active field and yet it's probably safe to say that only a very small fraction of this work is seen to be used nowadays. My opinion is that mgcv would have a greater impact per unit effort if more families were added rather than developing further techniques for automatic smoothing parameter selection. An exception to this would be techniques that are surprisingly simple, such as the Fellner–Schall method (Wood & Fasiolo 2016) [personal communication] which only requires the first two derivatives. This promising method needs full development.

The example of Section 7 is known as the proportional odds model, or cumulative logit model, where there are parallelism constraints applied to the η_js for each explanatory variable bar the intercept. It is a special case of a nonparallel cumulative logit model whereby $\eta_j = \alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}$. Would an unwarranted parallelism assumption explain the less than perfect behaviour in the lower tail of Figure 6? Alternatively, it might be remedied by choosing a link function with a heavier tail or allowing asymmetry, such as a cauchit or complementary log-log link.

In the fuel efficiency example of Section 8 one might wish the smooths to be monotonic unless it is strongly believed that some interaction exists—this applies specifically for Figure 8(e). Would monotonic P-splines be more suitable? And how easy would it be to

constrain the smooths of Figure 8(a) and (d) to be equal so that their difference is some constant? In Section 1.1 we analyze these data using some new methodology.

For a general additive model is it possible to have smoothing parameters from, e.g., $f_4(x_4)$ and $f_6(x_6)$ constrained to be equal? An application of this might be to smooth two variables whose support are equal.

1.1 VGAMS with P-splines

Although WPS convey automatic additive smoothing to potentially a very large class of models, a simpler alternative for most models described in the appendix is to consider them within the VGAM framework and utilize Wood (2004). Here are some sketch details, which makes use of the notation of Yee (2015). For the VGAM class with constraint matrices $\mathbf{H}_k$,

$$\boldsymbol{\eta}_i \;=\; \mathbf{H}_1\,\boldsymbol{\beta}^*_{(1)} + \sum_{k=2}^{d} \mathbf{H}_k\,\boldsymbol{f}^*_k(x_{ik}), \tag{1}$$

where $\boldsymbol{f}^*_k(x_k) = (f^*_{(1)k}(x_k), \ldots, f^*_{(\mathcal{R}_k)k}(x_k))^T$ is a $\mathcal{R}_k$-vector of smooth functions of x_k to be estimated. Each vector of component functions in (1) generates several columns of the model matrix since they are linear combinations of B-spline basis functions, therefore

$$\boldsymbol{\eta}_i \;=\; \mathbf{H}_1\,\boldsymbol{\beta}^*_{(1)} + \sum_{k=2}^{d} \mathbf{H}_k\,\mathbf{X}^*_{[ik]}\,\boldsymbol{\beta}^*_{[k]} \tag{2}$$

for some submatrix $\mathbf{X}^*_{[ik]}$. Equation (2) can be further bolstered by allowing terms of the form $\mathbf{H}_k \left(\ldots, f^*_{(j)k}(x_{ikj}) \ldots\right)$ in $\boldsymbol{\eta}_i$, i.e, η_j-specific values of a covariate (known as the $\boldsymbol{x}_{ij}$ or `xij` facility), but details are not given here. We are maximizing the penalized log-likelihood

$$\ell(\boldsymbol{\beta}^*) \;=\; \sum_{i=1}^{n} \ell_i\{\eta_1(\boldsymbol{x}_i), \ldots, \eta_M(\boldsymbol{x}_i)\} - J(\boldsymbol{\lambda})$$

for a suitably regular model, where the penalty term is

$$J(\boldsymbol{\lambda}) \;=\; \sum_{k=2}^{d} \boldsymbol{\beta}_{[k]}^{*T} \left\{ \mathbf{P}_k^* \otimes \operatorname{diag}(\lambda_{(1)k}, \ldots, \lambda_{(\mathcal{R}_k)k}) \right\} \boldsymbol{\beta}_{[k]}^* \;=\; \boldsymbol{\beta}^{*T} \mathbf{P}^* \boldsymbol{\beta}^*$$

with $\mathbf{P}_k^* = \mathbf{D}_k^{*T} \mathbf{D}_k^*$, and $\mathbf{D}_k$ is the matrix representation of the δth-order differencing operator Δ^δ applied to the B-spline coefficients since the knots for x_k are equidistant.

For a response $\boldsymbol{y}$, the computations are performed by augmenting $\boldsymbol{y}$, the large model matrix $\mathbf{X}_{\text{VAM}}$ comprising blocks of $\mathbf{X}_{[ik]}^*$ and the weight matrices $\mathbf{W} = \operatorname{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$:

$$\boldsymbol{y}' \;=\; \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0}_\varphi \end{pmatrix}, \quad \mathbf{X}_{\text{PVAM}} \;=\; \begin{pmatrix} \mathbf{X}_{\text{VAM}} \\ \widetilde{\mathbf{X}} \end{pmatrix}, \quad \mathbf{W}' \;=\; \operatorname{diag}(\mathbf{W},\, \mathbf{I}_\varphi), \tag{3}$$

for some dimension φ and where $\mathbf{P}^* = \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}}$ with $\widetilde{\mathbf{X}} =$

$$\Big(\mathbf{O},\; \operatorname{diag}\Big(\mathbf{D}_2^* \otimes \operatorname{diag}(\lambda_{(1)2}^{1/2}, \ldots, \lambda_{(\mathcal{R}_2)2}^{1/2}), \ldots, \mathbf{D}_d^* \otimes \operatorname{diag}(\lambda_{(1)d}^{1/2}, \ldots, \lambda_{(\mathcal{R}_d)d}^{1/2}) \Big) \Big). \tag{4}$$

For general responses the above can be embedded within a PIRLS algorithm that uses working responses and working weight matrices, and then the GCV/UBRE is minimized. This could be performed by performance-oriented iteration or by outer iteration, as described by Wood (2006). The advantage of the above approach over WPS is its relative simplicity and fewer requirements such as only needing second-order derivatives.

Applying an implementation of this in the VGAM R package to the fuel efficiency data gives Figure 1. The family function `binormal(zero = NULL)` was used, which specifies $\eta_1 = \mu_1$, $\eta_2 = \mu_2$, $\eta_3 = \log \sigma_{11}$, $\eta_4 = \log \sigma_{22}$, $\eta_5 = \log\{(1+\rho)/(1-\rho)\}$, so that the covariances can be modelled with covariates (`weight` and `hp` here). There is substantial agreement between the fitted means and Figure 8(a,b,d,e), but with the functions decreasing monotonically here as expected. However, the plots c–e,h–j suggest that the intercept-only assumption of

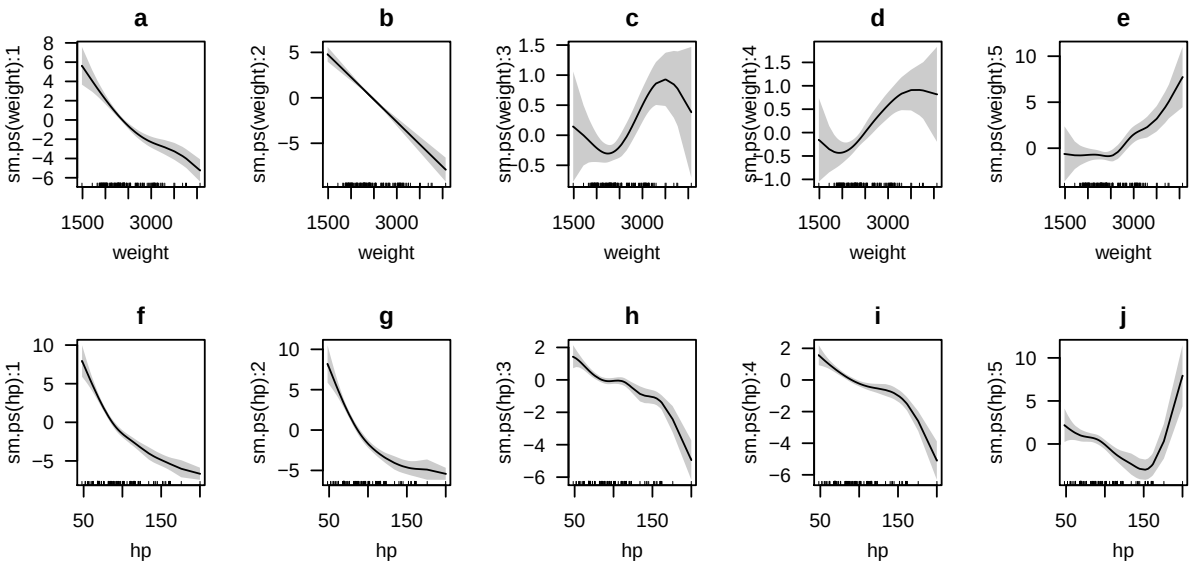

Figure 1: Fitted bivariate normal regression applied to the fuel efficiency data.

the covariances made by WPS is in doubt; the fitted component functions appear nonlinear and, e.g., the variability decreases with increasing `hp`.

In Figure 1 it would be straightforward using VGAM to constrain plots a,b to differ by a constant (known or unknown), and similarly for plots f,g, for example,

$$\mathbf{H}_1 = \mathbf{I}_5, \quad \mathbf{H}_2 = \mathbf{H}_3 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & \mathbf{I}_3 \end{pmatrix}$$

in the unknown case.

2 ON SOME APPENDICES

2.1 On Tweedie Models (Appendix J)

Some series expansions could be exploited, e.g., for large y the digamma function $\psi(y) = \log y - 1/(2y) - \sum_{k=1}^{\infty} B_{2k}/(2ky^{2k}) \sim \log y - (2y)^{-1} - (12y^2)^{-1}$ where B_k is the kth Bernoulli number. Then $\partial \log W_j / \partial p$ involves the difference between two terms which can make use of this series. Likewise for the $j^2\,\psi'(-j\alpha)$ term involving the trigamma function.

2.2 On Ordered Categorical Models (Appendix K)

While only the logit link is currently implemented for an underlying logistic distribution, there is an argument `link` taking on the value `"identity"`, which is potentially confusing for the user. Ideally the chain rule could be applied more generally so that a variety of link functions could be catered for. Being based on the multinomial distribution, the ordered categorical model should share some of the computational particulars that the multinomial distribution has.

A non-parallel cumulative link model would be worthwhile but quite challenging to implement, and exacerbated by possible intersecting $\eta_j(\boldsymbol{x}_i)$ that results in out-of-range probabilities. Although the $\mathcal{J}$ contraction over $\boldsymbol{x}^k$ technique would be utilized much to handle the parallel case, it probably would be rather inefficient when the number of levels of the response becomes even moderate.

2.3 On Software Implementation (Appendix M)

The first author's mgcv R package is the most advanced state-of-the-art software for GAMs and the author must be commended for writing and enhancing this over many years.

WPS have made a very good start by identifying a few of the most strategic models and implementing those first, such as the Cox model and zero-altered Poisson distribution. As an author of another large GAM-like R package, I can identify with the mgcv developers on the never-ending task of extending and maintaining the software. This explains why there are currently some limitations in a few family functions (which should hopefully be addressed in due course), for example, the negative binomial has an index parameter (called `theta` in the software) that is restricted to intercept-only and/or required to have a known value. Another example is the multivariate normal whose elements of the variance-covariance matrix are also intercept-only (Section 1.1). Section 3.3 mentions offsets for η_j, however mgcv currently seems unable to handle these when there are multiple linear predictors.

At the risk of being branded as an irritant, here are some suggestions that may be useful.

1. Prediction involving nested data-dependent terms fails, e.g., `s(scale(x))`. Here, a limitation of *safe* prediction is exposed and one solution is *smart* prediction (Yee 2015, Secs.8.2.5,18.6).

2. The ability for family functions to handle multiple responses can be useful, e.g., `gam(cbind(y1, y2) ~ s(x2), family = poisson, pdata)`. However this would entail a considerable amount of work to convert other functions to handle this feature.

3. For `multinom()` and `ocat()` objects the `fitted()` methods function returns the same result as `predict()`. The fitted probabilities for each class are a more natural type of fitted value.

4. Much attention is given to preserving numerical stability. There are instances where the use of `expm1()` is preferable, as is for `log1p()` too.

5. A specific distribution worth investigating in terms of robustness for handling singularities is the skew normal (Azzalini 1985) where for $Y = \lambda_1 + \lambda_2 Z$, with $0 < \lambda_2$ and $Z \sim \mathrm{SN}(\lambda)$ (whose density is $f(z;\lambda) = 2\,\phi(z)\,\Phi(\lambda z)$ for real z and λ), the expected information matrix as a function of $(\lambda_1, \lambda_2, \lambda)$ is singular as $\lambda \to 0$ even though all 3 parameters remain identifiable.

6. A final question or two: using the syntax of

   ```
   gam(list(y1 ~ s(x), y2 ~ s(v), y3 ~ 1, 1+3 ~ s(z)-1), family=mvn(d=3))
   ```

   how might one constraint the intercepts of η_1 and η_2 to be equal? Likewise, how might one constraint the intercepts of η_1 to be twice the value of the intercept of η_2?

References

Anderson, J. A. (1984), 'Regression and ordered categorical variables', *J. Roy. Statist. Soc. Ser. B* **46**(1), 1–30. With discussion.

Azzalini, A. A. (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics* **12**(2), 171–178.

Chen, Y. & Samworth, R. J. (2016), 'Generalized additive and index models with shape constraints', *J. Roy. Statist. Soc. Ser. B* **78**(4), 729–754.

Wood, S. N. (2004), 'Stable and efficient multiple smoothing parameter estimation for generalized additive models', *J. Amer. Statist. Assoc.* **99**(467), 673–686.

Wood, S. N. (2006), *Generalized Additive Models: An Introduction with R*, Chapman and Hall, London.

Wood, S. N. & Fasiolo, M. (2016), 'A generalized Fellner-Schall method for smoothing parameter estimation with application to Tweedie location, scale and shape models', *arXiv preprint arXiv:1606.04802* .

Yee, T. W. (2015), *Vector Generalized Linear and Additive Models: With an Implementation in R*, Springer, New York, USA.