P-Spline Vector Generalized Additive Models and Its Applications

Chanatda Somchit, Chris Wild, and Thomas Yee

Department of Statistics
University of Auckland

1 December 2015
Outline

1. Introduction to VGLMs and VGAMs
2. P-spline VGAMs
3. Simulation study
4. Examples
5. References
VGLMs and VGAMs are the extension of class of GLMs and GAMs to include a class of multivariate regression models [Yee and Wild, 1996].

- **VGLMs** model each parameter as a linear combination of the covariates,

\[
\eta_j(x) = \beta_j^T x = \sum_{k=1}^{p} \beta_{(j)k} x_k, \quad j = 1, \ldots, M,
\]

- **VGAMs** extend VGLMs to

\[
\eta_j(x) = \sum_{k=1}^{p} f_{(j)k}(x_k), \quad j = 1, \ldots, M.
\]
Introduction to VGLMs and VGAMs (2)

The current class of VGLMs/VGAMs is very large and includes many statistical distributions and models:

- univariate and multivariate distributions,
- categorical data analysis,
- quantile and expectile regression,
- time series, survival analysis,
- extreme value analysis, nonlinear regression,
- reduced-rank regression, ordination, etc.

The underlying algorithm of VGAMs is

- Modified vector backfitting using vector splines.
- But...it is not easy to integrate the automatic numerical procedure used to determine the shape of non-linear terms from the data.
We aim to...

- **integrate an automatic procedure** for estimating the smoothing parameters to the VGAM framework.

To achieve this......

- the ideas of **GAMs based on penalized regression splines** proposed by Marx and Eilers [1998] and Wood [2006] are generalized to the VGAM class.
Therefore, ... 

- we develop **VGAMs based on penalized regression splines** with **P-spline smoothers**, which we call ‘**P-spline VGAMs**’.

As a result, ... 

- P-spline VGAMs can be transformed into the VGLM framework,
- and maximized by **penalized iteratively re-weighted least squares (P-IRLS)**.
- The **computational procedure for the automatic and stable multiple smoothing parameter selection** can be implemented.
- The issue of determining the shape of the smooth terms can be resolved.
The underlying ideas of P-splines are
- to use **B-splines as basis functions**, and a large number of equally-spaced knots are used,
- but to prevent the problem of overfitting, a **discrete approximate wiggleness penalty** is applied to the model fitting objective,

\[ \sum_{s=1}^{S} \left( \Delta^{[d]} a_s \right)^2 = a^T P_{[d]} a, \quad (1) \]

where \( P_{[d]} = D_{[d]}^T D_{[d]} \), \( D_{[d]} \) is the matrix consisting of \( d^{th} \) order difference of the coefficients, and \( a \) is a parameter vector.

- Here is a linear combination of B-spline basis functions:

\[ f(x_i) = \sum_{s=1}^{S} a_s B_s(x_i). \quad (2) \]
The basic structure of P-spline VGAMs is

\[ g_j(\theta_j) = \eta_j(x) = f_{(j)1}(x_1) + \cdots + f_{(j)p}(x_p), \quad j = 1, \ldots, M, \]  

(3)

- where \( f_{(j)k} \) are represented using B-splines, and are centered for identifiability.
- If there is an intercept, \( x_1 = 1 \),
- All P-spline VGAM smooth components are estimated simultaneously.

Allow the P-spline VGAM approach to be more general for use in many more situations
In practice, we may wish to constrain the effects of a single covariate

- to be the same for different $\eta_j$, or
- to have no effect for others.

For example

\begin{align*}
\eta_1 &= \beta_{(1)1} + f_{(1)2}(x_2) + f_{(1)3}(x_3) \\
\eta_2 &= \beta_{(2)1} + f_{(1)2}(x_2).
\end{align*}

(4)
Yee and Wild [1996] introduced ‘constraint matrices’ applied directly to the linear/additive predictors to control how the covariates act – “constraints on the functions”.

These constraints

- are very useful for categorical models, such as, a bivariate odds-ratio model, and the proportional odds model,
- lead the VGLM/VGAM approach to be more general for use in most situations.

We will generalize these ideas to the P-spline VGAM framework.
In general for P-spline VGAMs, we represent the models as

\[ \eta = f_1(x_1) + \cdots + f_p(x_p) \]

\[ = H_1 f_1^*(x_1) + \cdots + H_p f_p^*(x_p), \quad (5) \]

where

- \( H_1, \ldots, H_p \) are known full column-rank ‘constraint matrices’,
- \( f_k^* = (f_{(1)}(x_k), \ldots, f_{(R_k)}(x_k))^T \) is a vector consisting of a possibly reduced set of smooth functions.
- Each smooth term is centered for identifiability.
- No constraints at all, \( H_k = I_M \).
The smooth predictor vector $\eta$ can be now written as

$$\eta(x_i) = \sum_{k=1}^{p} H_k X_{ik}^* \beta_k^*,$$

(6)

where

- $X_{ik}^* = x_{ik}^* \otimes I_{R_k}$, $x_{ik}^* = (B_{k:1}(x_{ik}), \ldots, B_{k:S_k}(x_{ik}))^T$ is a vector of B-splines generated at the values of $x_k$ and $i$th observation,

- $\beta_k^* = \text{vec} \left( \begin{array}{c} \beta_{(1)k}^T \\ \vdots \\ \beta_{(R_k)k}^T \end{array} \right)$, where $\beta_{r_k}^* = (a_{(r_k)k:1}, \ldots, a_{(r_k)k:S_k})^T$, for $r_k = 1, \ldots, R_k$. 
Let $\beta^* = (\beta_1^T, \ldots, \beta_p^T)^T$ be a vector containing all of the possibly reduced sets of B-spline coefficients in the models,

$$X_{vam} = \left( (X_{am} \tilde{e}_1) \otimes H_1 \mid (X_{am} \tilde{e}_2) \otimes H_2 \mid \cdots \mid (X_{am} \tilde{e}_p) \otimes H_p \right),$$

where

- $X_{vam}$ is the model matrix for P-spline VGAMs,
- $X_{am}$ is the ‘additive model’ model matrix for one $\eta_j$, and
- $\tilde{e}_k = \begin{pmatrix} 0 \\ I(S_k \times S_k) \\ 0 \end{pmatrix}$.

Then,

$$\eta = X_{vam} \beta^*.$$  \hfill (7)

Equation (7) is just the form of VGLMs, therefore

$$\ell(\beta^*) = \sum_{i=1}^{n} w_i \ell\{\eta_1(x_i), \ldots, \eta_M(x_i)\}.$$  \hfill (8)
Now, it is a good chance to control the model’s smoothness by adding a \textbf{wiggliness penalty} to the log-likelihood objective of $\ell(\beta^*)$.

- The penalty term for P-spline VGAMs is given by

\[
J(\lambda) = \sum_{k=1}^{p} \sum_{j=1}^{R_k} \lambda(j)_k \beta^T_{(j)k} D_{[d]k}^T D_{[d]k} \beta_{(j)k} \\
= \sum_{k=1}^{p} \beta^*_k \left\{ \left( D_{[d]k}^T D_{[d]k} \right) \otimes \text{diag} \left( \lambda(1)_k, \ldots, \lambda(R_k)_k \right) \right\} \beta^*_k \\
= \sum_{k=1}^{p} \beta^*_k P^*_k \beta^*_k,
\]

where $P^*_k = \left( D_{[d]k}^T D_{[d]k} \right) \otimes \text{diag} \left( \lambda(1)_k, \ldots, \lambda(R_k)_k \right)$. 


Then, the quadratic penalty on the parameter vector $\beta^*$ for P-spline VGAMs is given by

$$J(\lambda) = \beta^*^T P_\lambda^* \beta^*,$$

where $P_\lambda^* = \text{blockdiag}(P_{\lambda_1}, \ldots, P_{\lambda_p})$.

The penalized log-likelihood for P-spline VGAMs is then

$$\ell^*(\beta^*) = \ell(\beta^*) - \frac{1}{2} \beta^*^T P_\lambda^* \beta^*.$$  

(9)
Newton-Raphson algorithm is applied for maximizing the log-likelihood (9),

\[ \beta^*(t+1) = \beta^*(t) + \mathcal{I}(\beta^*(t))^{-1} U(\beta^*(t)) . \]  

Maximizing \( \ell^*(\beta^*) \) using (10) leads to an iterative solution for penalized iteratively reweighted least squares (P-IRLS) of

\[ \beta^*(t+1) = \left( X_{vam}^T W^{(t)} X_{vam} + P_\lambda^* \right)^{-1} \left( X_{vam}^T W^{(t)} z^{(t)} \right) . \]  

(11)
Given values for $\lambda$, $\hat{\beta}^{(t+1)}$ is the solution to

$$\min_{\beta^*} \left( z - X_{vam} \beta^* \right)^T W \left( z - X_{vam} \beta^* \right) + \beta^*^T P_\lambda^* \beta^*,$$  

(12)

where

- $W = \text{blockdiag}(W_1, \ldots, W_n)$,
  - Fisher scoring: $(W_i)_{jk} = -w_i E \left( \frac{\partial^2 \ell_i}{\partial \eta_j \partial \eta_k} \right)$,

- $z = (z_1^T, \ldots, z_n^T)^T$ and $u = (u_1^T, \ldots, u_n^T)^T$, where
  - $(u_i)_j = w_i \frac{\partial \ell_i}{\partial \eta_j}$, and $z_i = \eta_i + (W_i)^{-1} u_i$. 

Chanatda Somchit  (UoA)  
P-Spline VGAMs  
1 December 2015  17 / 32
Data augmentation is applied to the adjusted dependent vector, regressors, and weights:

\[
    z' = \begin{pmatrix} z \\ \vartheta \end{pmatrix}, \quad X'_\text{vam} = \begin{pmatrix} X_{\text{vam}} \\ \tilde{P}^* \lambda \end{pmatrix}, \quad W' = \text{blockdiag}(W, I_{\vartheta}),
\]

where \( P^*_\lambda = \tilde{P}^*_\lambda \tilde{P}^*_\lambda \), \( \vartheta = \sum_{k=1}^{p} (S_k - d) \cdot M \).

Therefore, (12) can be replaced by the equivalent:

\[
    \min_{\beta^*} \left( z' - X'_\text{vam} \beta^* \right)^T W' \left( z' - X'_\text{vam} \beta^* \right). \tag{13}
\]

We convert the GLS system of equations to OLS:

\[
    z''(t) = X''(t) \beta^* + \varepsilon''(t). \tag{14}
\]

In P-IRLS, we fit OLS model above using the data augmentation of \( z, W, \) and \( X_{\text{vam}} \) until the convergence is achieved.
Given an estimate for $\beta^*$, the multiple smoothing parameter selection for penalized least squares above can be solved by the minimization of the GCV or the ‘unbiased risk estimator’ (UBRE) w.r.t. the multiple smoothing parameters.

The UBRE score for the P-spline VGAM approach:

$$\nu_u(\lambda) = \frac{1}{nM} \left\| \sqrt{W} (z - X_{VAM} \beta^*) \right\|^2 - 1 + \frac{2}{nM} \operatorname{tr}(A_\lambda).$$

- $A_\lambda = X_{vam}^T W X_{vam} \left( X_{vam}^T W X_{vam} + P_\lambda \right)^{-1}$ is the hat matrix.
- $\lambda$ enters the UBRE score through $A_\lambda$.
- $\operatorname{tr}(A_\lambda)$ represents the estimated effective degree of freedom (EDF).
Estimating smoothing parameters (2)

- each working penalized linear model of the P-IRLS iteration, $\nu_u(\lambda)$ is minimized w.r.t. $\lambda$ (performance iteration).
- The two steps:
  
  **Step 1** Obtain an estimate of $\beta^*$ via:
  
  $$\beta^{*(t+1)} = \arg\max_{\beta^*} \ell^*(\beta^*).$$

  **Step 2** Obtain an estimate of $\lambda$ via:
  
  $$\lambda^{(t+1)} = \arg\min_{\lambda} \nu_u(\lambda).$$

  The two steps of estimations above are iterated until convergence is met.

- Here, we employ the computational approach developed by Wood [2004] to minimize the UBRE or GCV scores.
The simulation study was based on the model:

\[
\begin{align*}
y_{i1}^* &= \beta_{(1)1} + \beta_{(1)2} x_{i2} + f_{(1)3}(x_{i3}) + f_{(1)4}(x_{i4}) + \varepsilon_{i1}, \\
y_{i2}^* &= \beta_{(2)1} + \beta_{(2)2} x_{i2} + f_{(2)3}(x_{i3}) + f_{(2)4}(x_{i4}) + \varepsilon_{i2},
\end{align*}
\]

The binary responses \( y_{i1} \) and \( y_{i2} \) are determined according to the rule:

\[
\begin{align*}
y_{ij} &= 1 \quad \text{if} \quad y_{ij}^* > 0 \\
y_{ij} &= 0 \quad \text{if} \quad y_{ij}^* \leq 0
\end{align*}
\]

The three test functions, \( f_{(1)3}(x_{i3}) = \cos(2\pi x_{i3}) \), \( f_{(2)3}(x_{i3}) = 2 \sin(\pi x_{i4}) \) and \( f_{(1)4}(x_{i4}) = f_{(2)4}(x_{i4}) = 0 x_{i4} \).

Three uniform covariates on \((0, 1)\) were simulated.

The error terms \((\varepsilon_{i1}, \varepsilon_{i2})\):

\[
\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \overset{iid}{\sim} N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right).
\]
Simulation Study (2): a semiparametric bivariate probit model

Figure: Boxplots of the MSE of $\hat{f}(1)_3$, $\hat{f}(2)_3$, $\hat{f}(1)_4$, and $\hat{f}(2)_4$, when employing the P-spline VGAM approach. The numbers .1, .5, .9 in the y-axis captions denote the three different correlations, $\rho$. 
Figure: Boxplots of the difference in MSE between the P-spline VGAM approach and the VGAM approach of $\hat{f}_{13}, \hat{f}_{23}, \hat{f}_{14}$, and $\hat{f}_{24}$. Negative values indicate that the new approach performs better than the alternative.
Examples: Mackerel egg survey (1)

- The data from 1992 mackerel egg survey
- The response of interest is the egg counts (egg.count).
- The covariates of interest are:
  - latitude (lat),
  - longitude (lon)
  - water temperature at a depth of 20m (temp.20),
  - the sea bed depth at the sampling location (b.dept),
  - distance from 200m sea bed contour (c.dist).

***Borchers et al. [1997] and Wood [2006]
An additive Poisson model is fitted to the count response (egg.count), with a mean given by

$$E[\text{egg.count}_i] = g_i \times [\text{net area}]_i,$$

where $g_i$ is the density of eggs, per square metre of sea surface, at $i^{th}$ sampling location.

$$\log(E[\text{egg.count}_i]) = f_i + \log([\text{net area}]_i),$$

where $f_i = \log(g_i)$.

```r
fit.ps1 <- psvgam(egg.count ~ ps(lat, 5) + ps(lon, 5) + ps(temp.20m, 5) + ps(b.depth, 5) + ps(c.dist, 5) + offset(log.net.area), data = mack1, family = poissonff)
```
- The dispersion parameter estimate indicates the existence of overdispersion in this data set.

- Trying the negative binomial distribution:

  ```r
  fit.ps2 <- psvgam(egg.count ~ ps(lat, 5) + ps(lon, 5) + 
                    ps(temp.20m, 5) + ps(b.depth, 5) + 
                    ps(c.dist, 5) + offset(log.net.area),
                    data = mack1, family = negbinomial)
  ```

- The negative binomial model is more appropriate than the Poisson model.
Estimated smooth terms for the mackerel model `fit.ps2`.
Figure: (a) Observed mackerel eggs densities per square metre of sea surface as assessed by net samples. (b) Predicted log densities of mackerel eggs over the survey area, according to the model fit.ps1.
Let’s investigate how the probability of having a household cat and a household dog is related to people’s ages.

- We fit a nonparametric bivariate logistic model to the example of cat and dog pet ownership presented by Yee [2015].
- For homogeneity, we restrict the analysis to a subset of 2569 European women and remove any missing values.

\[
\begin{align*}
\eta_1 &= \logit P(Y_1 = 1|x_2) = \beta_{(1)1} + f_{(1)2}(x_2), \\
\eta_2 &= \logit P(Y_2 = 1|x_2) = \beta_{(2)1} + f_{(2)2}(x_2), \\
\eta_3 &= \log \psi = \beta_{(3)1} + f_{(3)2}(x_2),
\end{align*}
\]

where \( \psi \) is the odds ratio.

```
fitps.cd1 <- psvgam(cbind(cat, dog) ~ ps(age, 15),
                    binom2.or(zero = NULL),
                    data = women.eth0.catdog)
```
Examples: Cat and dog data (2)

Plots of fitted component functions, according to the model \texttt{fitps.cd1}.
Figure: Estimated probabilities for all four combinations (both cats and dogs, cats only, dogs only, and no cats or dogs) of a subset of European women using the P-spline VGAM approach.


