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

Introduction to VGLMs and VGAMs (1)

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,

$$\eta_j(\boldsymbol{x}) = \boldsymbol{\beta}_j^T \boldsymbol{x} = \sum_{k=1}^{p} \beta_{(j)k}\, x_k, \qquad j = 1, \ldots, M,$$

- **VGAMs** extend VGLMs to

$$\eta_j(\boldsymbol{x}) = \sum_{k=1}^{p} f_{(j)k}(x_k), \qquad j = 1, \ldots, M.$$

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.

P-spline VGAMs (1)

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.

P-spline VGAMs (2)

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.

P-spline VGAMs (3)

- 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 wiggliness penalty** is applied to the model fitting objective,

$$\sum_{s=1}^{S} \left(\Delta^{[d]} \, a_s \right)^2 \;=\; \boldsymbol{a}^T \, \mathbf{P}_{[d]} \, \boldsymbol{a}, \tag{1}$$

where $\mathbf{P}_{[d]} = \mathbf{D}_{[d]}^T \, \mathbf{D}_{[d]}$, $\mathbf{D}_{[d]}$ is the matrix consisting of d^{th} order difference of the coefficients, and $\boldsymbol{a}$ is a parameter vector.

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

$$f(x_i) \;=\; \sum_{s=1}^{S} a_s \, B_s(x_i). \tag{2}$$

P-spline VGAMs (4)

The basic structure of P-spline VGAMs is

$$g_j(\theta_j) = \eta_j(\boldsymbol{x}) = f_{(j)1}(x_1) + \cdots + f_{(j)p}(x_p), \qquad j = 1, \ldots, M, \tag{3}$$

- 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 η_j**, or
- **to have no effect for others**.
- For example

$$\begin{aligned} \eta_1 &= \beta_{(1)1} + f_{(1)2}(x_2) + f_{(1)3}(x_3) \\ \eta_2 &= \beta_{(2)1} + f_{(1)2}(x_2). \end{aligned} \tag{4}$$

P-spline VGAMs and constraint matrices (2)

- **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.

Setting up P-spline VGAMs as penalized VGLMs (1)

In general for P-spline VGAMs, we represent the models as

$$\begin{aligned} \boldsymbol{\eta} &= \boldsymbol{f}_1(x_1) + \cdots + \boldsymbol{f}_p(x_p) \\ &= \mathbf{H}_1\, \boldsymbol{f}_1^*(x_1) + \cdots + \mathbf{H}_p\, \boldsymbol{f}_p^*(x_p), \end{aligned} \tag{5}$$

where

- $\mathbf{H}_1, \ldots, \mathbf{H}_p$ are known full column-rank 'constraint matrices',
- $\boldsymbol{f}_k^* = \left(f_{(1)k}(x_k), \ldots, f_{(\mathrm{R}_k)k}(x_k)\right)^T$ is a vector consisting of a possibly reduced set of smooth functions.
- Each smooth term is centered for identifiability.
- No constraints at all, $\mathbf{H}_k = \mathbf{I}_M$.

Setting up P-spline VGAMs as penalized VGLMs (2)

The smooth predictor vector $\boldsymbol{\eta}$ can be now written as

$$\boldsymbol{\eta}(\boldsymbol{x}_i) \;=\; \sum_{k=1}^{p} \mathbf{H}_k \, \mathbf{X}^*_{ik} \, \boldsymbol{\beta}^*_k, \qquad (6)$$

where

- $\mathbf{X}^*_{ik} \;=\; \boldsymbol{x}^{*\,T}_{ik} \otimes \mathbf{I}_{\mathrm{R}_k}, \; \boldsymbol{x}^*_{ik} \;=\; \left(B_{k:1}(x_{ik}), \ldots, B_{k:\mathrm{S}_k}(x_{ik})\right)^T$ is a vector of B-splines generated at the values of x_k and ith observation,
- $\boldsymbol{\beta}^*_k \;=\; \mathrm{vec}\begin{pmatrix} \boldsymbol{\beta}^{*\,T}_{(1)k} \\ \vdots \\ \boldsymbol{\beta}^{*\,T}_{(\mathrm{R}_k)k} \end{pmatrix}$, where $\boldsymbol{\beta}^*_{(r_k)k} = \left(a_{(r_k)k:1}, \ldots, a_{(r_k)k:\mathrm{S}_k}\right)^T$, for $r_k = 1, \ldots, \mathrm{R}_k$.

Setting up P-spline VGAMs as penalized VGLMs (3)

- Let $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_1^{*T}, \ldots, \boldsymbol{\beta}_p^{*T}\right)^T$ be a vector containing all of the possibly reduced sets of B-spline coefficients in the models,
- $\mathbf{X}_{\text{vam}} = \left((\mathbf{X}_{\text{am}}\, \widetilde{\boldsymbol{e}}_1) \otimes \mathbf{H}_1 \,\middle|\, (\mathbf{X}_{\text{am}}\, \widetilde{\boldsymbol{e}}_2) \otimes \mathbf{H}_2 \,\middle|\, \cdots \,\middle|\, (\mathbf{X}_{\text{am}}\, \widetilde{\boldsymbol{e}}_p) \otimes \mathbf{H}_p \right),$

  where
  - $\mathbf{X}_{\text{vam}}$ is the model matrix for P-spline VGAMs,
  - $\mathbf{X}_{\text{am}}$ is the 'additive model' model matrix for one η_j, and
  - $\widetilde{\boldsymbol{e}}_k = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_{(s_k \times s_k)} \\ \mathbf{0} \end{pmatrix}.$
- Then,

$$\boldsymbol{\eta} = \mathbf{X}_{\text{vam}}\, \boldsymbol{\beta}^*. \tag{7}$$

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

$$\ell(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} w_i\, \ell\{\eta_1(\boldsymbol{x}_i), \ldots, \eta_M(\boldsymbol{x}_i)\}. \tag{8}$$

Now, it is a good chance to control the model's smoothness by adding **a wiggliness penalty** to the log-likelihood objective of $\ell\left(\boldsymbol{\beta}^{*}\right)$.

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

$$
\begin{aligned}
J(\boldsymbol{\lambda}) &= \sum_{k=1}^{p} \sum_{j=1}^{\mathrm{R}_k} \lambda_{(j)k}\, \boldsymbol{\beta}_{(j)k}^{T}\, \mathbf{D}_{[d]k}^{T}\, \mathbf{D}_{[d]k}\, \boldsymbol{\beta}_{(j)k} \\
&= \sum_{k=1}^{p} \boldsymbol{\beta}_{k}^{*\,T} \left\{ \left(\mathbf{D}_{[d]k}^{T}\, \mathbf{D}_{[d]k} \right) \otimes \mathrm{diag}\left(\lambda_{(1)k}, \ldots, \lambda_{(\mathrm{R}_k)k} \right) \right\} \boldsymbol{\beta}_{k}^{*} \\
&= \sum_{k=1}^{p} \boldsymbol{\beta}_{k}^{*\,T}\, \mathbf{P}_{\lambda k}^{*}\, \boldsymbol{\beta}_{k}^{*},
\end{aligned}
$$

where $\mathbf{P}_{\lambda k}^{*} = \left(\mathbf{D}_{[d]k}^{T}\, \mathbf{D}_{[d]k} \right) \otimes \mathrm{diag}\left(\lambda_{(1)k}, \ldots, \lambda_{(\mathrm{R}_k)k} \right).$

Setting up P-spline VGAMs as penalized VGLMs (5)

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

$$J(\boldsymbol{\lambda}) \; = \; \boldsymbol{\beta}^{*T} \, \mathbf{P}^*_{\lambda} \, \boldsymbol{\beta}^*,$$

  - where $\mathbf{P}^*_{\lambda} = \text{blockdiag}\left(\mathbf{P}^*_{\lambda 1}, \ldots, \mathbf{P}^*_{\lambda p}\right)$.

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

$$\ell^*(\boldsymbol{\beta}^*) \; = \; \ell(\boldsymbol{\beta}^*) - \frac{1}{2} \, \boldsymbol{\beta}^{*T} \, \mathbf{P}^*_{\lambda} \, \boldsymbol{\beta}^*. \tag{9}$$

P-IRLS formulation (1)

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

$$\boldsymbol{\beta}^{*(t+1)} = \boldsymbol{\beta}^{*(t)} + \boldsymbol{\mathcal{I}}\left(\boldsymbol{\beta}^{*(t)}\right)^{-1} \boldsymbol{U}\left(\boldsymbol{\beta}^{*(t)}\right). \qquad (10)$$

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

$$\boldsymbol{\beta}^{*(t+1)} = \left(\mathbf{X}_{\text{vam}}^T \mathbf{W}^{(t)} \mathbf{X}_{\text{vam}} + \mathbf{P}_\lambda^*\right)^{-1} \left(\mathbf{X}_{\text{vam}}^T \mathbf{W}^{(t)} \boldsymbol{z}^{(t)}\right). \qquad (11)$$

P-IRLS formulation (2)

- **Given values for $\boldsymbol{\lambda}$**, $\widehat{\boldsymbol{\beta}}^{*(t+1)}$ is the solution to

$$\min_{\boldsymbol{\beta}^*} \quad \left(\boldsymbol{z} - \mathbf{X}_{\text{vam}}\,\boldsymbol{\beta}^*\right)^T \mathbf{W} \left(\boldsymbol{z} - \mathbf{X}_{\text{vam}}\,\boldsymbol{\beta}^*\right) + \boldsymbol{\beta}^{*T}\,\mathbf{P}^*_{\lambda}\,\boldsymbol{\beta}^*, \quad (12)$$

where

- $\mathbf{W} = \text{blockdiag}\left(\mathbf{W}_1, \ldots, \mathbf{W}_n\right)$,

  - Fisher scoring: $\left(\mathbf{W}_i\right)_{jk} \;=\; -w_i\,E\left(\dfrac{\partial^2 \ell_i}{\partial\eta_j\,\partial\eta_k}\right)$,

- $\boldsymbol{z} \;=\; \left(\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T\right)^T$ and $\boldsymbol{u} \;=\; \left(\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_n^T\right)^T$, where $\left(\boldsymbol{u}_i\right)_j \;=\; w_i\,\dfrac{\partial \ell_i}{\partial \eta_j}$, and $\boldsymbol{z}_i \;=\; \boldsymbol{\eta}_i + \left(\mathbf{W}_i\right)^{-1}\,\boldsymbol{u}_i$.

P-IRLS formulation (3)

- **Data augmentation** is applied to the adjusted dependent vector, regressors, and weights:

$$\boldsymbol{z}^{'} = \begin{pmatrix} \boldsymbol{z} \\ \vartheta \end{pmatrix}, \qquad \mathbf{X}_{\text{vam}}^{'} = \begin{pmatrix} \mathbf{X}_{\text{vam}} \\ \widetilde{\mathbf{P}}_{\lambda}^{*} \end{pmatrix}, \qquad \mathbf{W}^{'} = \text{blockdiag}\left(\mathbf{W}, \mathbf{I}_{\vartheta}\right),$$

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

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

$$\min_{\boldsymbol{\beta}^*} \quad \left(\boldsymbol{z}^{'} - \mathbf{X}_{\text{vam}}^{'}\,\boldsymbol{\beta}^*\right)^{T} \mathbf{W}^{'} \left(\boldsymbol{z}^{'} - \mathbf{X}_{\text{vam}}^{'}\,\boldsymbol{\beta}^*\right). \tag{13}$$

- We convert the GLS system of equations to OLS:

$$\boldsymbol{z}^{''(t)} = \mathbf{X}_{\text{vam}}^{''(t)}\,\boldsymbol{\beta}^* + \varepsilon^{''(t)}. \tag{14}$$

- **In P-IRLS**, we fit OLS model above using the data augmentation of $\boldsymbol{z}$, $\mathbf{W}$, and $\mathbf{X}_{\text{vam}}$ until the convergence is achieved.

Estimating smoothing parameters (1)

- Given an estimate for β^*, 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:

$$\nu_u(\boldsymbol{\lambda}) = \frac{1}{nM} \left\| \sqrt{\mathbf{W}} \left(\boldsymbol{z} - \mathbf{X}_{\text{VAM}} \boldsymbol{\beta}^* \right) \right\|^2 - 1 + \frac{2}{nM} \operatorname{tr}(\mathbf{A}_\lambda).$$

  - $\mathbf{A}_\lambda = \mathbf{X}_{\text{vam}}^T \mathbf{W} \mathbf{X}_{\text{vam}} \left(\mathbf{X}_{\text{vam}}^T \mathbf{W} \mathbf{X}_{\text{vam}} + \mathbf{P}_\lambda \right)^{-1}$ is the hat matrix.
  - $\boldsymbol{\lambda}$ enters the UBRE score through $\mathbf{A}_\lambda$.
  - $\operatorname{tr}(\mathbf{A}_\lambda)$ represents the estimated effective degree of freedom (EDF).

Estimating smoothing parameters (2)

- each working penalized linear model of the P-IRLS iteration, $\nu_u(\boldsymbol{\lambda})$ is minimized w.r.t. $\boldsymbol{\lambda}$ (**performance iteration**).
- The two steps:

  Step 1 Obtain an estimate of $\boldsymbol{\beta}^*$ via:

  $$\boldsymbol{\beta}^{*(t+1)} = \underset{\boldsymbol{\beta}^*}{\operatorname{argmax}} \ \ell^*(\boldsymbol{\beta}^*).$$

  Step 2 Obtain an estimate of $\boldsymbol{\lambda}$ via:

  $$\boldsymbol{\lambda}^{(t+1)} = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \ \nu_u(\boldsymbol{\lambda}).$$

  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.

Simulation Study (1): a semiparametric bivariate probit model

- The simulation study was based on the model:

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

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

$$
\begin{cases} y_{ij} = 1 & \text{if} \quad y_{ij}^* > 0 \\ y_{ij} = 0 & \text{if} \quad y_{ij}^* \leq 0 \end{cases} ; j = 1, 2.
$$

- 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 $f_{(1)4}(x_{i4}) = f_{(2)4}(x_{i4}) = 0x_{i4}$.
- 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

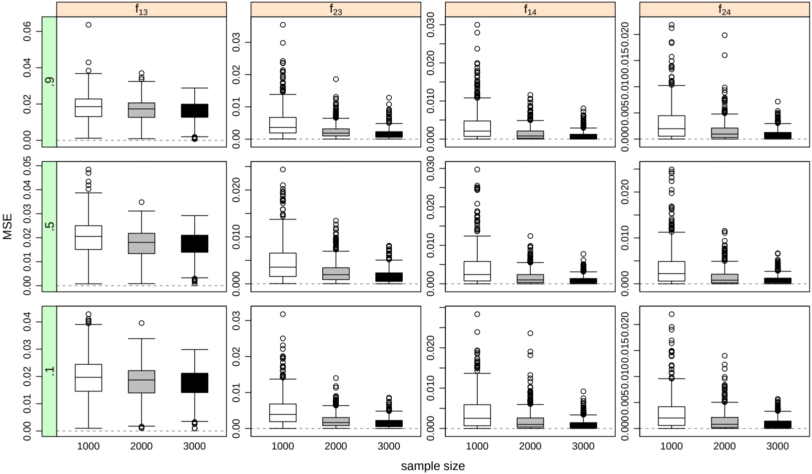


Figure: Boxplots of the MSE of $\widehat{f}_{(1)3}$, $\widehat{f}_{(2)3}$, $\widehat{f}_{(1)4}$, and $\widehat{f}_{(2)4}$, when employing the P-spline VGAM approach. The numbers .1, .5, .9 in the y-axis captions denote the three different correlations, ρ.

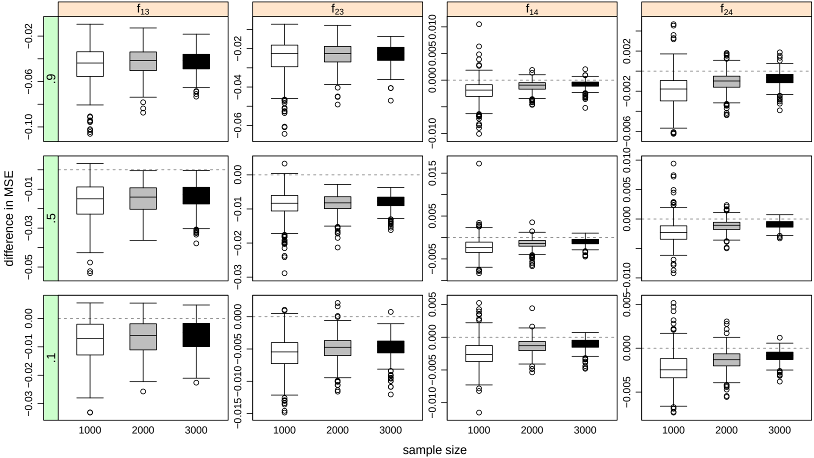


Figure: Boxplots of the difference in MSE between the P-spline VGAM approach and the VGAM approach of $\widehat{f}_{(1)3}$, $\widehat{f}_{(2)3}$, $\widehat{f}_{(1)4}$, and $\widehat{f}_{(2)4}$. Negative values indicate that the new approach performs better than the alternative.

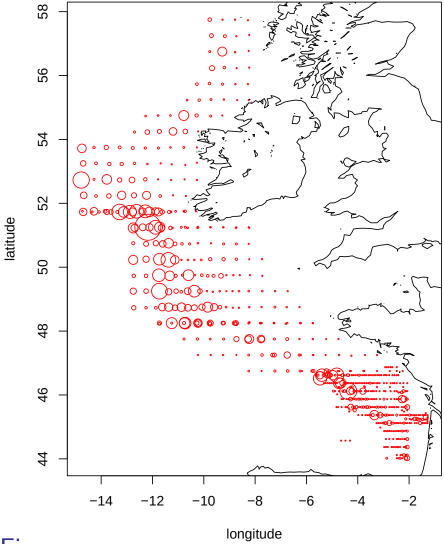


Figure: **Observed mackerel eggs densities per square metre of sea surface as assessed by net samples.**

- 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

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

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

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

where $f_i = \log(g_i)$.

```
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)
```

Examples: Mackerel egg survey (3)

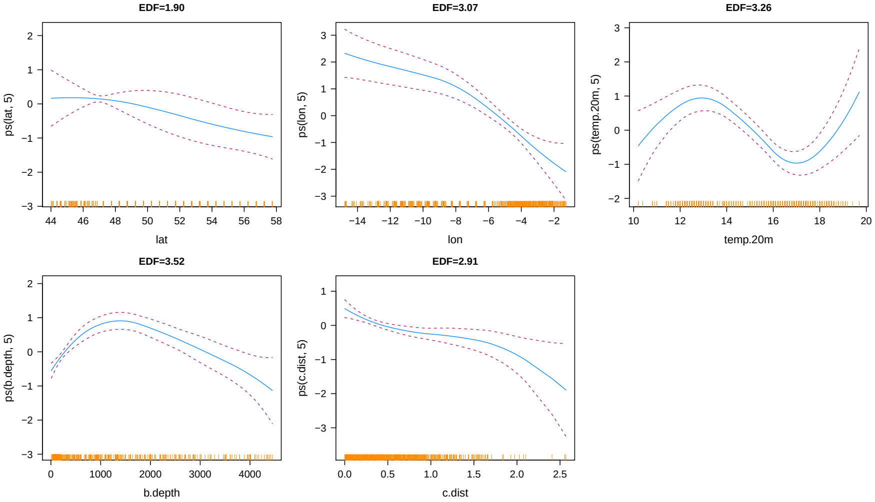
- The dispersion parameter estimate indicates the existence of overdispersion in this data set.

- Trying the negative binomial distribution:

```
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`.



Model Predictions

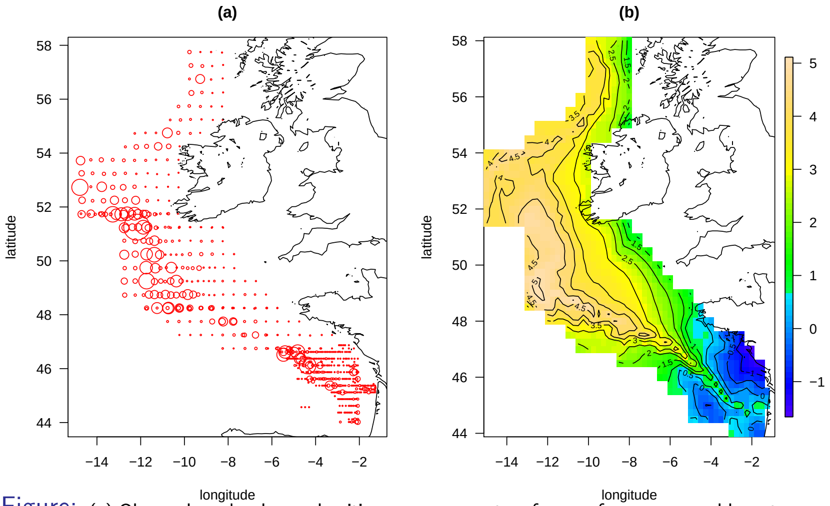


Figure: (a) Observed mackerel eggs densities per square metre of sea surface as assessed by net samples. (b) Predicted log densities of mackerel eggs over the survey area, according to the model `fit.ps1`.

Examples: Cat and dog data (1)

Let's investigate how the probability of having a household cat and a household dog is related to people's ages.

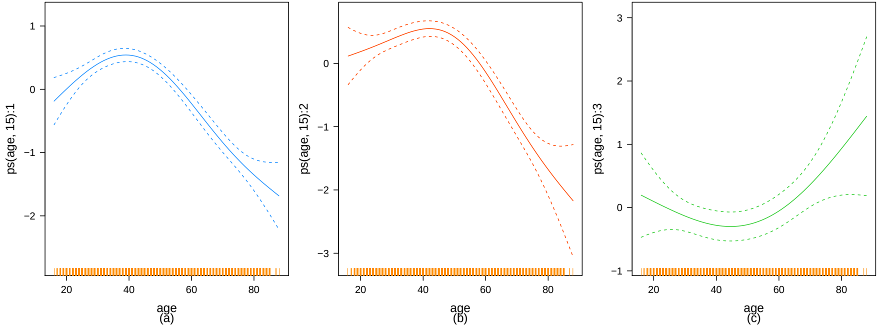
- We fit a nonparametric bivariate logistic model to the example of cat and dog pet ownership presented by Yee [2015].
- For homogeneity, we restrict the analysis to a subset of 2569 European women and remove any missing values.

$$\begin{aligned}
\eta_1 &= \operatorname{logit} P(Y_1 = 1|x_2) = \beta_{(1)1} + f_{(1)2}(x_2),\\
\eta_2 &= \operatorname{logit} P(Y_2 = 1|x_2) = \beta_{(2)1} + f_{(2)2}(x_2),\\
\eta_3 &= \log \psi = \beta_{(3)1} + f_{(3)2}(x_2), \qquad (15)
\end{aligned}$$

where ψ is the odds ratio.

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

Plots of fitted component functions, according to the model `fitps.cd1`.



Examples: Cat and dog data (3)

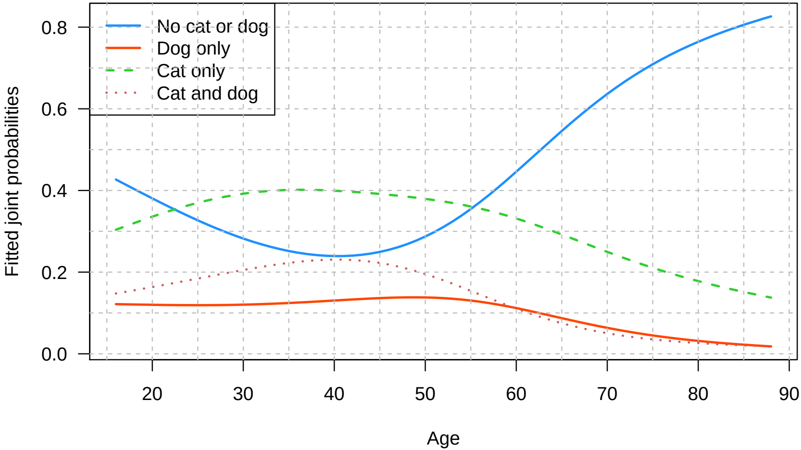


Figure: Estimated probabilities for all four combinations (both cats and dogs, cats only, dogs only, and no cats or dogs) of a subset of European women using the P-spline VGAM approach.

References

D. Borchers, S. Buckland, I. Priede, and S. Ahmadi. Improving the precision of the daily egg production method using generalized additive models. *Canadian Journal of Fisheries and Aquatic Sciences*, 54(12):2727–2742, 1997.

B. D. Marx and P. H. Eilers. Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28(2):193–209, 1998.

S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99:673–686, 2004.

S. N. Wood. *Generalized additive models: An introduction with R*. Chapman & Hall/CRC, Boca Raton, FL, USA, 2006.

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

T. W. Yee and C. J. Wild. Vector generalized additive models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:481–493, 1996.