

Specifying GAMs & GAMMs with mgcv

David Lawrence Miller

CREEM, University of St Andrews

SPOILER ALERT

your model is probably some kind of (fancy) GLM

General setup of GAMs in `mgcv` (and my brain)

General setup

$$y = X\beta + \epsilon \quad \text{with penalty } \beta^T S \beta$$

I think of this as “linear”, in the sense that $X\beta$ is linear

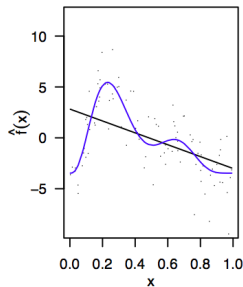
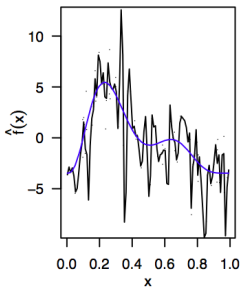
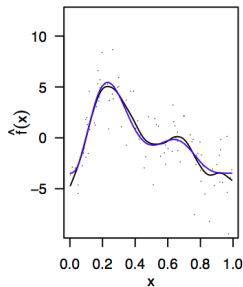
X includes a column for each parametric covariate, plus one for each basis evaluation (at knots or pseudo-knots).

what about this penalty thing?

$$\beta^T S \beta = \int_{\Omega} (D^m f(\mathbf{x}))^2 d\mathbf{x}$$

where D^m is some differential operator, commonly for univariate:

$$D^m f(x) = \frac{\partial^2 f(x)}{\partial x^2}$$



A quick tour of spline bases

How many different bases?

Currently ~17 (some bases are v. similar or inter-related) in mgcv:

"ad", "sf", "cc", "so", "cp", "sos", "cr", "sw", "cs",
"t2", "ds", "tensor", "mrf", "tp", "ps", "ts", "re"

?smooth.terms

?smooth.construct.*.smooth.spec

?gam.models

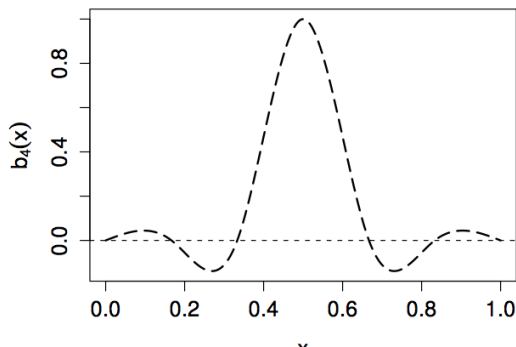
cubic splines

- ▶ simple basis construction
- ▶ orthogonal (Hermite) polynomials defined by their knots

`s(x, bs="cr", k=10, knots=NULL)`

without `knots`, knots are placed evenly over x

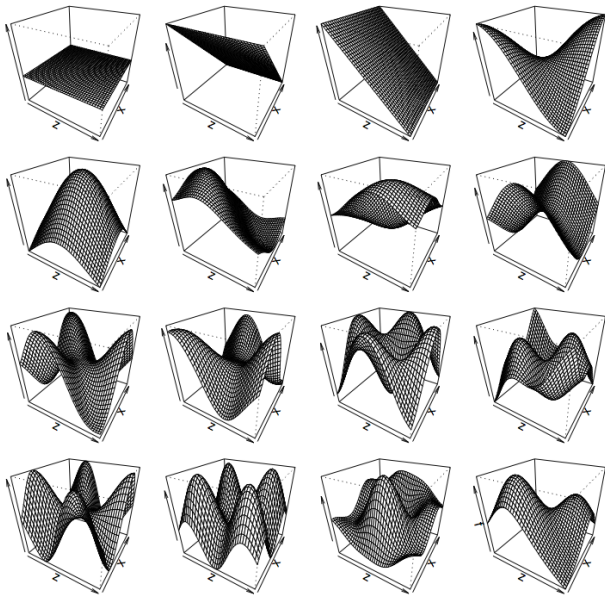
cp basis “more optimal” (see `tprs`)



thin plate splines

- ▶ multi-dimensional basis
- ▶ 2-part basis
 - ▶ global bits (orthogonal polynomial terms)
 - ▶ local bits (radial basis functions)
- ▶ requires 1 radial function per datum
- ▶ knots?

tprs basis



thin plate regression splines – Wood (2003)

- ▶ instead of knots, use all data *but*
- ▶ take eigendecomposition $X = UDU^T$
- ▶ truncate to 1st k columns (D is in “eigen-order”)
- ▶ “more optimal” than knot-based approaches

`s(x,y,..., bs="tp", k=M+k.def, knots=NULL)`

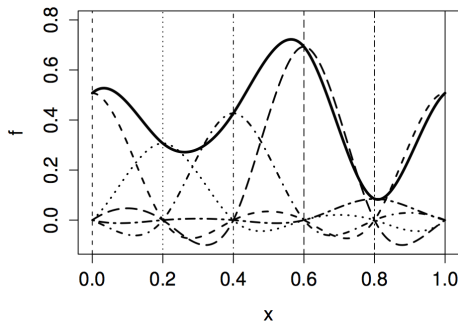
where M is nullspace size and `k.def` is 8 (1D), 27 (2D), 100 (3D+)

cyclic smoothers

- ▶ seasonality?
- ▶ temporal periodicity?
- ▶ angles?

```
s(x, bs="cc", k=10, knots=NULL)
```

wrap at range(x), unless knots specified



random effects

- ▶ IID normal random effects
- ▶ multivariate ($s(x,y,z,bs="re")$ is $\sim x:y:z-1$ interaction)
- ▶ exploits equivalence of random effects and splines
- ▶ useful when you just have a “few” random effects

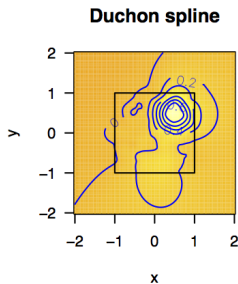
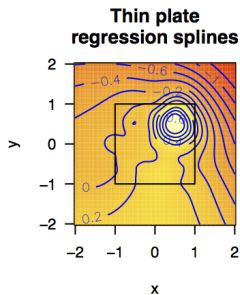
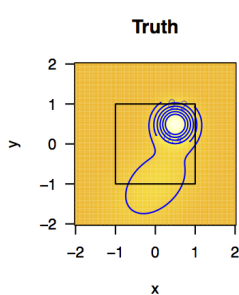
`s(x,bs="re")`

`?gam.vcomp`

Duchon splines

- ▶ sometimes spatial smoothers curl up at the edges
- ▶ Duchon splines limit nullspace in 2D+

`s(x,y,..., bs="ds", k=M+k.def, m=c(1,.5) knots=NULL)`

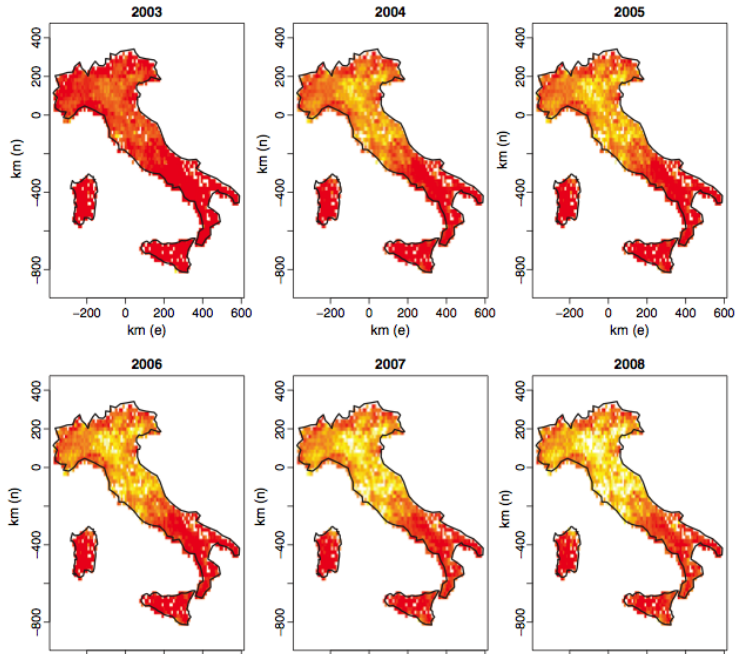


tensor products

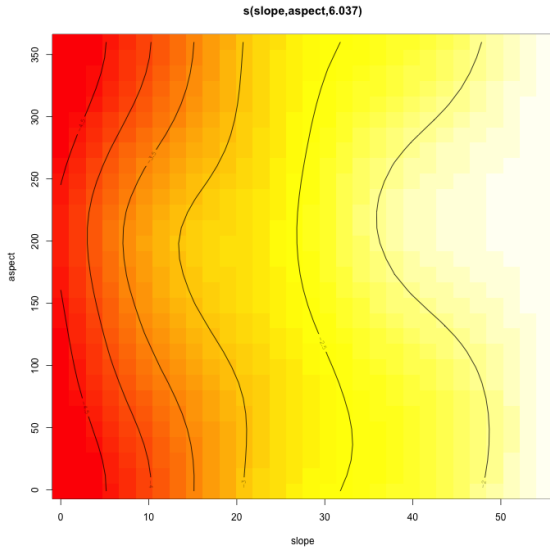
- ▶ tprs multivariate but assumes isotropy
- ▶ are space and time the same? (hint: NO)
- ▶ different smoothing parameters
- ▶ “push” 2D spatial smoother through time
- ▶ Marra et al (2011) give an example

```
te(x,y,t, bs=c("tp","cr"), d=c(2,1), k=c(100,10))
```

tensor products (space-time)



tensor products (slope-aspect)



by=

- ▶ what if you only have a couple of years?
- ▶ for factors: multiple smooths
- ▶ for numerics: “parametric” tensor
- ▶ need to add parameteric term
- ▶ can use `id=` to “link” smooths to have same (estimated) parameters

```
s(x,y,bs="tp",by=as.factor(year)) + as.factor(year)
```

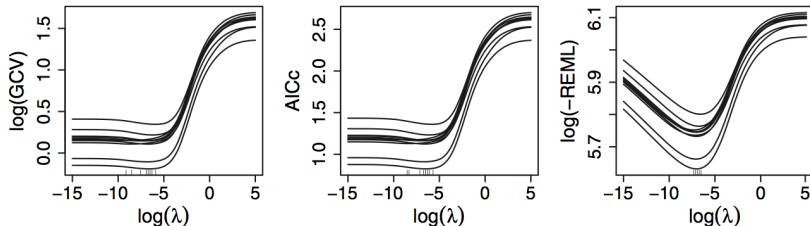
Model checking

SECOND SPOILER

the default options are (almost definitely) wrong
(for you)

Quick note on fitting

- ▶ by default `gam` uses GCV for smoothing parameter selection
- ▶ GCV prone to overfitting (Wood, 2011)
- ▶ GCV also problematic w. correlated covariates (Wood, 2006; pers. obsn.)
- ▶ REML better – BUT can only compare models w. same parametric terms (ML?)



how do I best control flexibility?

- ▶ `k` parameter controls “basis size”
- ▶ look at output of `summary` and `gam.check`
- ▶ `?choose.k`
- ▶ double `k`, see what happens?
- ▶ watch out, larger basis gives more, weirder functions

```
> gam.check(b)
```

Method: GCV Optimizer: magic

Smoothing parameter selection converged after 8 iterations.

The RMS GCV score gradient at convergence was 1.072609e-05 .

The Hessian was positive definite.

The estimated model rank was 37 (maximum possible: 37)

Model rank = 37 / 37

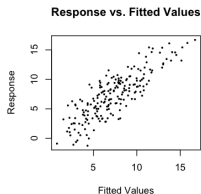
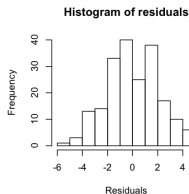
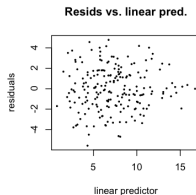
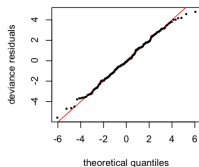
Basis dimension (k) checking results. Low p-value (k-index<1) may indicate that k is too low, especially if edf is close to k'.

	k'	edf	k-index	p-value
s(x0)	9.000	2.318	0.996	0.45
s(x1)	9.000	2.306	0.969	0.35
s(x2)	9.000	7.655	0.961	0.25
s(x3)	9.000	1.233	1.037	0.68



how do I know when I've got it right?

- ▶ plot the gam object – over/under-fitting?
- ▶ looking at `gam.check` (brain scan example in Wood 2006)
 - ▶ left column – response distribution correct?
 - ▶ right column – non-constant variance?
- ▶ plot residuals vs. covariates



I've probably talked for too long already...

Other stuff

- ▶ randomised quantile residuals (Dunn and Smyth, 1996)
- ▶ `bam` for big additive models (Wood et al, 2014)
 - ▶ can do AR1 correlation structures (in order)
- ▶ `gamm` when you have “many” random effects or correlation
 - ▶ correlation specified as in `lme`
 - ▶ useful link: <http://glmm.wikidot.com/faq#modelspec>
 - ▶ e.g. `correlation=corAR1(form=~segment|tr.su)`
 - ▶ smooth \leftrightarrow random effect relation
 - ▶ numerically unstable? (pers. exp.)
 - ▶ autocorrelogram can save you some stress :)
- ▶ use `nb` for `negbin` and `tw` for Tweedie if you want their parameters estimated with smoothing pars
- ▶ `select=TRUE` adds extra smoothing to *every* term, meaning smooths can be estimated as 0 effect

References (for later)

Dunn, P K, and G K Smyth. Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics* 5(3) (1996): 236–244.

Marra, G, DL Miller, and L Zanin. Modelling the Spatiotemporal Distribution of the Incidence of Resident Foreign Population. *Statistica Neerlandica* 66(2) (2011): 133–160.

Wood, SN. Thin Plate Regression Splines. *Journal of the Royal Statistical Society. Series B* 65(1) (2003): 95–114.

Wood, SN. *Generalized Additive Models: an Introduction with R*, Chapman & Hall/CRC, 2006.

Wood, SN. Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society. Series B* 73(1) (2011): 3–36.

Wood, SN, Y Goude, and S Shaw. Generalized Additive Models for Large Data Sets. *Journal of the Royal Statistical Society Series C* (2014).

Almost all figures stolen from Wood (2006) or (2011)

Thanks!

Talk available at:

converged.yt/talks/creemcrackers-splines/talk.pdf