A toolbox of smooths

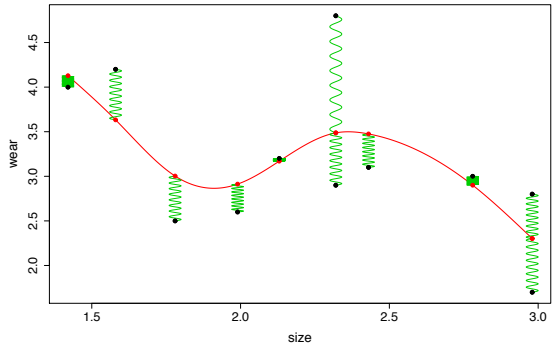
Simon Wood
Mathematical Sciences, University of Bath, U.K.

Smooths for semi-parametric GLMs

- The piecewise linear smoother is not bad, but we can find better and more general basis-penalty smoothers for a variety of modelling purposes.
- In one dimension there are several alternatives, and not alot to choose between them.
- In 2 or more dimensions there is a major choice to make.
  - If the arguments of the smooth function are variables which all have the same units (e.g. spatial location variables) then an *isotropic* smooth may be appropriate. This will tend to exhibit the same degree of flexibility in all directions.
  - If the relative scaling of the covariates of the smooth is essentially arbitrary (e.g. they are measured in different units), then *scale invariant* smooths should be used, which do not depend on this relative scaling.

Splines

- All the smooths covered here are based on *splines*. Here's the basic idea ...

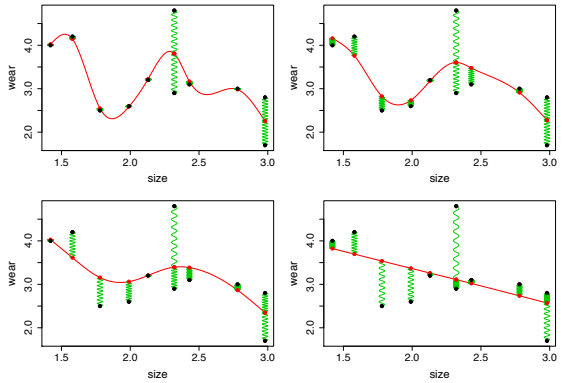


- Mathematically the red curve is the *function* minimizing

$$\sum_i (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx.$$

Splines have variable stiffness

- Varying the flexibility of the strip (i.e. varying λ) changes the *spline function* curve.



- But irrespective of λ the spline functions always have the same basis.

Why splines are special

- We can produce splines for a variety of penalties, including for functions of several variables. e.g.

$$\int f'''(x)^2 dx \quad \text{or} \quad \int\int f_{xx}(x,z)^2 + 2f_{xz}(x,z)^2 + f_{zz}(x,z)^2 dxdz$$

- Splines always have an n dimensions basis - quadratic penalty representation.
- If $y_i = g(x_i)$ and f is the cubic spline interpolating x_i, y_i then

$$\max|f - g| \leq \frac{5}{384} \max(x_{i+1} - x_i)^4 \max(g'''')$$

(best possible — end conditions are a bit unusual for this).

- Bases that are optimal for approximating known functions are a good starting point for approximating unknown functions.

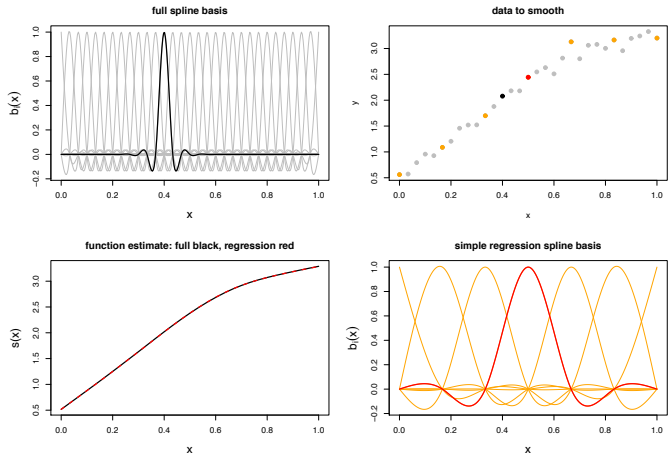
Penalized regression splines

- Full splines have one basis function per data point.
- This is computationally wasteful, when penalization ensures that the *effective* degrees of freedom will be much smaller than this.
- Penalized regression splines simply use fewer spline basis functions. There are two alternatives:
  1. Choose a representative subset of your data (the 'knots'), and create the spline basis as if smoothing only those data. Once you have the basis, use it to smooth all the data.
  2. Choose how many basis functions are to be used and then solve the problem of finding the set of this many basis functions that will optimally approximate a full spline.

I'll refer to 1 as *knot based* and 2 as *eigen based*.

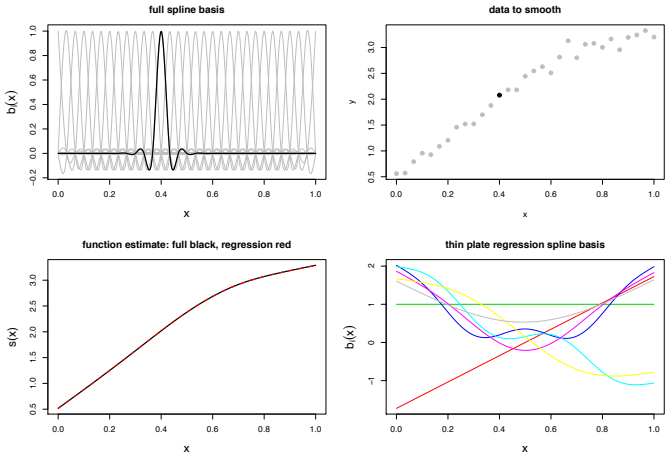
Knot based example: `"cr"`

- In `mgcv` the `"cr"` basis is a knot based approximation to the minimizer of $\sum_i (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx$ — a cubic spline. `"cc"` is a cyclic version.



Eigen based example: `"tp"`

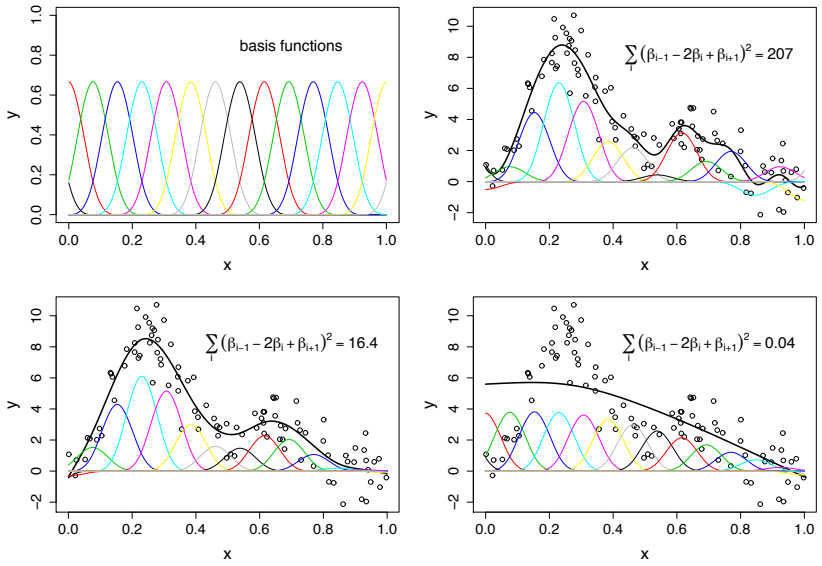
- The `"tp"`, *thin plate regression spline* basis is an eigen approximation to a thin plate spline (including cubic spline in 1 dimension).



Asymptotic considerations

- Is this reduced rank approach ok? Consider cubic *regression* spline with k equally spaced (h) knots.
- Average variance is $n^{-1}\mathrm{tr}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\sigma^2 = \sigma^2 k/n$.
- Squared bias is $O(h^8) = O(k^{-8})$, from approximation error of spline.
- Need squared bias to have same order as variance for neither to dominate as $n \to \infty$.
- i.e. $k \propto n^{-1/9}$, so that MSE is $O(n^{-8/9})$.
- For consistency merely require k to grow with n at less than $O(n)$.
- Penalization designed to improve MSE at any particular n.

P-spline illustration



P-splines: "ps" & "cp"

- There are many equivalent spline bases.
- With bases for which all the basis functions are translations of each other, it is sometimes possible to penalize the coefficients of the spline directly, rather than penalizing something like $\int f''(x)^2 dx$.
- Eilers and Marx coined the term 'P-splines' for this combination of spline bases with direct discrete penalties on the basis coefficients.
- P-splines allow a good deal of flexibility in the way that bases and penalties are combined.
- However splines with derivative based penalties have good approximation theoretic properties bound up with the use of derivative based penalties, and as a result tend to slightly out perform P-splines for routine use.

An adaptive smoother

- Can let the p-spline penalty vary with the predictor. e.g.

$$\mathcal{P}_a = \sum_{k=2}^{K-1} \omega_k(\beta_{k-1} - 2\beta_k + \beta_{k+1})^2 = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\omega})\mathbf{D}\boldsymbol{\beta}$$

where $\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdot \\ 0 & 1 & -2 & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$.

- Now let ω_k vary smoothly with k, using a B-spline basis, so that $\boldsymbol{\omega} = \mathbf{B}\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is the vector of basis coefficients.
- So, writing $\mathbf{B}_{.k}$ for the k^{th} column of $\mathbf{B}$ we have

$$\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\omega})\mathbf{D}\boldsymbol{\beta} = \sum_k \lambda_k \boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathrm{diag}(\mathbf{B}_{.k})\mathbf{D}\boldsymbol{\beta} = \sum_k \lambda_k \boldsymbol{\beta}^{\mathrm{T}}\mathbf{S}_k\boldsymbol{\beta}.$$

1 dimensional smoothing in `mgcv`

- Smooth functions are specified by terms like `s(x,bs="ps")`, on the rhs of the model formula.
- The `bs` argument of `s` specifies the class of basis...

  `"cr"` knot based cubic regression spline.

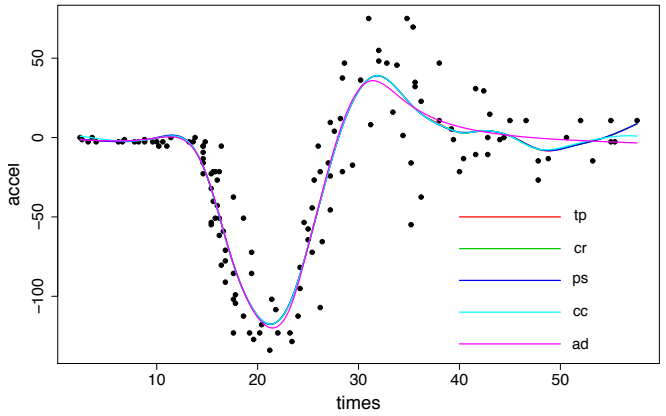
  `"cc"` cyclic version of above.

  `"ps"` Eilers and Marx style p-splines, with flexibility as to order of penalties and basis functions.

  `"ad"` adaptive smoother in which strength of penalty varies with covariate.

  `"tp"` thin plate regression spline. Optimal low rank eigen approx. to a full spline: flexible order penalty derivative.
- Smooth classes can be added (`?smooth.construct`).

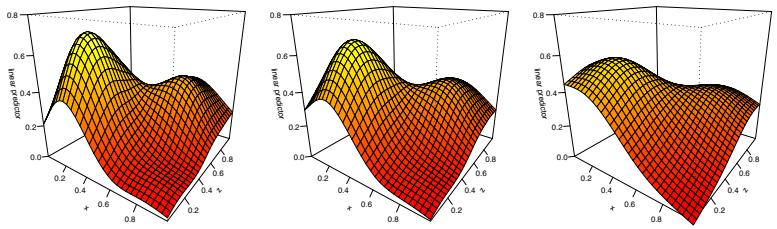
1D smooths compared



- So cubic regression splines, P-splines and thin plate regression splines give very similar results.
- A cyclic smoother is a little different, of course.
- An adaptive smoother can look very different.

Isotropic smooths

- One way of generalizing splines from 1D to several D is to turn the flexible strip into a flexible sheet (hyper sheet).
- This results in a *thin plate spline*. It is an *isotropic* smooth.
- Isotropy may be appropriate when different covariates are naturally on the same scale.
- In `mgcv` terms like `s(x,z)` generate such smooths.



Thin plate spline details

- In 2 dimensions a thin plate spline is the function minimizing

$$\sum_i \{y_i - f(x_i, z_i)\}^2 + \lambda \int f_{xx}^2 + 2f_{xz}^2 + f_{zz}^2 dxdz$$

- This generalizes to any number of dimensions, d, and any order of differential, m, such that $2m > d + 1$.
- *Any* thin plate spline is computed as

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_i(\mathbf{x}) + \sum_{i=1}^{M} \alpha_i \phi_i(\mathbf{x})$$

where η_i and ϕ_i are basis functions of known form and $\boldsymbol{\alpha}, \boldsymbol{\delta}$ minimize $\|\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}\boldsymbol{\alpha}\|^2 + \boldsymbol{\delta}^{\mathrm{T}}\mathbf{E}\boldsymbol{\delta}$ s.t. $\mathbf{T}^{\mathrm{T}}\boldsymbol{\delta} = \mathbf{0}$, where $\mathbf{E}$ and $\mathbf{T}$ are computed using the η_i and ϕ_i.

Thin plate regression splines

- Full thin plate splines have n parameters and $O(n^3)$ computational cost.
- This drops to $O(k^3)$ if we replace $\mathbf{E}$ by its rank k eigen approximation, $\mathbf{E}_k$, at cost $O(n^2k)$. Big saving if $k \ll n$
- Out of all rank k approximations this one minimizes

$$\max_{\delta \neq \mathbf{0}} \frac{\|(\mathbf{E} - \mathbf{E}_k)\delta\|}{\|\delta\|} \quad \text{and} \quad \max_{\delta \neq \mathbf{0}} \frac{\delta^{\mathrm{T}}(\mathbf{E} - \mathbf{E}_k)\delta}{\|\delta\|^2}$$

  i.e. the approximation is somewhat optimal, and avoids choosing 'knot locations'.
- For very large datasets, randomly subsample the data the data and work out the truncated basis from the subsample, to avoid $O(n^2k)$ eigen-decomposition costs being too high.

TPRS illustration

- As the theory suggests, the eigen approximation is quite effective. The following figure compares reconstructions of of the true function on the left, using and eigen based thin plate regression spline (middle), and one based on choosing knots. Both are rank 16 approximations.

Duchon Splines `s(x,z,bs="ds")`

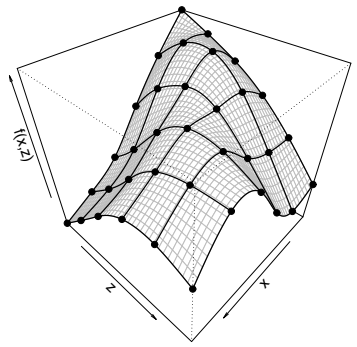
- The $m > d/2$ requirement causes thin plate splines in more than a few dimensions to be impractical, as the null space of the penalty rapidly becomes too high dimensional.
- But thin plate splines are only one special case of the splines introduced by Duchon (1977). He also considered penalties

$$\int_{\mathbb{R}^d} \|\tau\|^{2s} \sum_{\nu_1+\cdots+\nu_d=m} \frac{m!}{\nu_1! \ldots \nu_d!} \left(\mathfrak{F} \frac{\partial^m f}{\partial x_1^{\nu_1} \ldots \partial x_d^{\nu_d}} (\tau) \right)^2 \mathrm{d}\tau$$

  where $\mathfrak{F}$ denotes Fourier transform and τ is frequency.
- With $s = 0$ this is a thin plate spline penalty, but with $s > 0$ higher frequencies of the derivative field are penalized more heavily.
- Smoothers using this penalty exist if $m + s > d/2$, and have the form of a TPS, with a reduced dimensional null space. e.g. $m = 2$, $s = d/2 - 1$ gives null space dimension $d + 1$.
- Eigen-approximation is as for TPS.

Scale invariant smoothing: tensor product smooths

- Isotropic smooths assume that a unit change in one variable is equivalent to a unit change in another variable, in terms of function variability.
- When this is not the case, isotropic smooths can be poor.
- *Tensor product smooths* generalize from 1D to several D using a lattice of bendy strips, *with different flexibility in different directions*.



Tensor product smooths

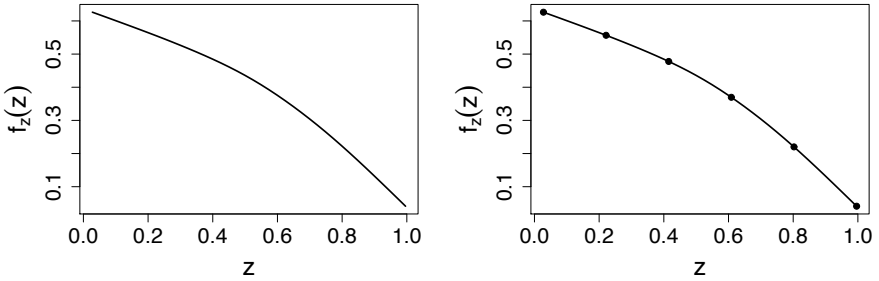
- Carefully constructed tensor product smooths are scale invariant.
- Consider constructing a smooth of x, z.
- Start by choosing *marginal* bases and penalties, as if constructing 1-D smooths of x and z. e.g.

$$f_x(x) = \sum \alpha_i a_i(x), \quad f_z(z) = \sum \beta_j b_j(z),$$

$$J_x(f_x) = \int f_x''(x)^2 dx = \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{S}_x \boldsymbol{\alpha} \;\; \& \;\; J_z(f_z) = \mathcal{B}^{\mathrm{T}} \mathbf{S}_z \mathcal{B}$$

Marginal reparameterization

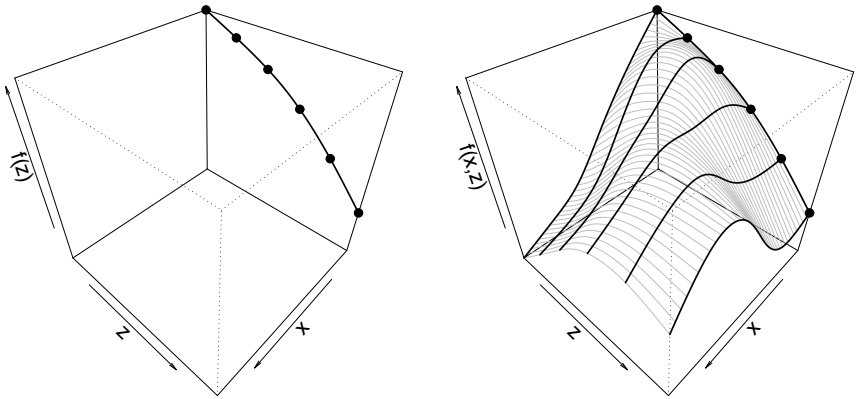
- Suppose we start with $f_z(z) = \sum_{i=1}^{6} \beta_j b_j(z)$, on the left.



- We can always re-parameterize so that its coefficients are functions heights, at knots (right). Do same for f_x.

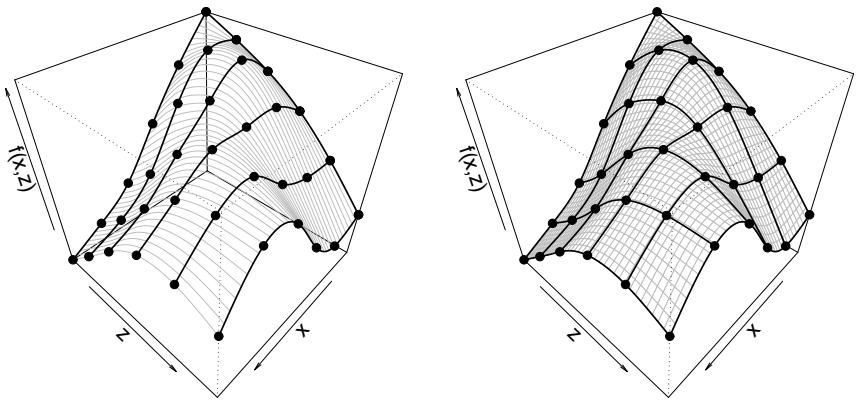
Making f_z depend on x

- Can make f_z a function of x by letting its coefficients vary smoothly with x



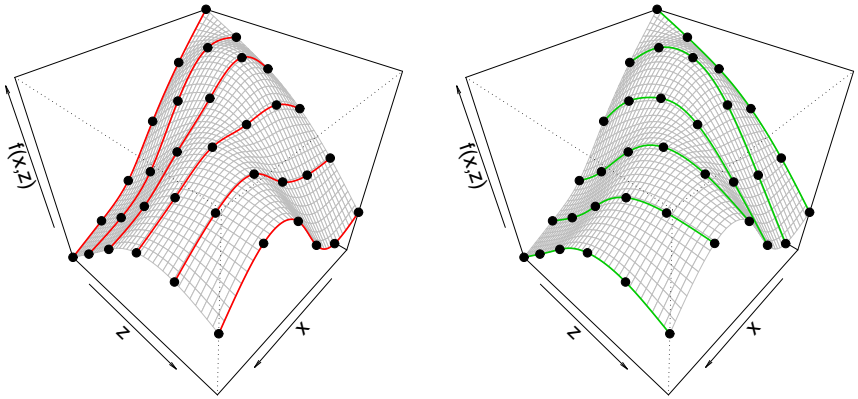
The complete tensor product smooth

- Use f_x basis to let f_z coefficients vary smoothly (left).
- Construct in symmetric (see right).



Tensor product penalties - one per dimension

- x-wiggliness: sum marginal x penalties over red curves.
- z-wiggliness: sum marginal z penalties over green curves.



Tensor product expressions

- So the tensor product basis construction gives:

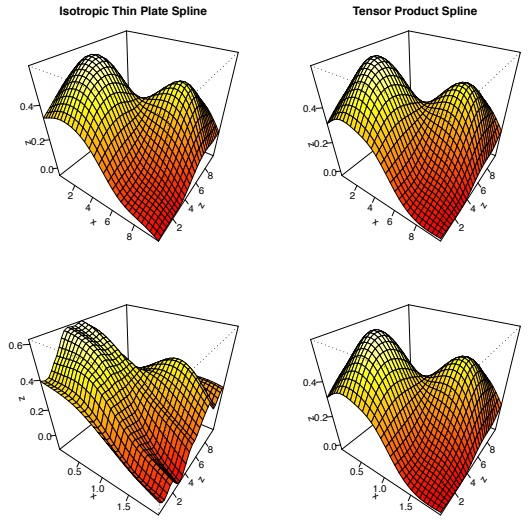
$$f(x,z) = \sum\sum \beta_{ij} b_j(z) a_i(x)$$

- With double penalties

$$J_z^*(f) = \beta^{\mathrm{T}} \mathbf{I}_I \otimes \mathbf{S}_z \beta \text{ and } J_x^*(f) = \beta^{\mathrm{T}} \mathbf{S}_x \otimes \mathbf{I}_J \beta$$

- The construction generalizes to any number of marginals and multi-dimensional marginals.
- Can start from any marginal bases & penalties (including mixtures of types).
- Note that the penalties maintain the basic meaning inherited from the marginals.

Isotropic vs. tensor product comparison



. . . each figure smooths the same data. The only modification is that x has been divided by 5 in the bottom row.

Tensor product smoothing in `mgcv`

- Tensor product smooths are constructed automatically from *marginal* smooths of lower dimension. The resulting smooth has a penalty for each marginal basis.
- `mgcv` can construct tensor product smooths from any *single penalty* smooths useable with `s` terms.
- `te` terms within the model formula invoke this construction. For example:
  - `te(x,z,v,bs="ps",k=5)` creates a tensor product smooth of x, z and v using rank 5 P-spline marginals: the resulting smooth has 3 penalties and basis dimension 125.
  - `te(x,z,t,bs=c("tp","cr"),d=c(2,1),k=c(20,5))` creates a tensor product of an isotropic 2-D TPS with a 1-D smooth in time. The result is isotropic in x,z, has 2 penalties and a basis dimension of 100. This sort of smooth would be appropriate for a location-time interaction.
- `te` terms are invariant to linear rescaling of covariates.

`ti` terms and functional ANOVA

- The basis for a `te` tensor product smooth, $f(x, z)$, contains a subspace of functions of the form $f(x) + f(z)$ (similar applies in higher dimensions).
- `ti` terms have this additive space removed, so that functional ANOVA models of the form $f(x) + f(z) + f(x, z)$ can be fitted via `ti(x) + ti(z) + ti(x,z)`, in a stably interpretable manner.
- `mgcv` also *allows* specifications like `s(x) + s(z) + s(x,z)` and `te(x) + te(z) + te(x,z)`, but the confounding of the main effects and interactions tends to lead to unstable effect estimates.

`t2` alternative tensor products

An alternative construction, due to Fabian Scheipl, and closely related to smoothing spline ANOVA, starts from a different marginal reparameterization

- Reparameterize each marginal smooth into unpenalized components and a component with an identity penalty.
- Form tensor product bases from each combination of unpenalized and penalized components, picking one from each margin.
- Each of the resulting bases is subject to a separate identity penalty, except for the basis made up only from unpenalized marginal components.
- The basis for the whole smooth is the sum of all these bases.
- `t2` in `mgcv` implements this using same syntax as `te`.

Other interactions with smooths `s(...,by=z)`

- Suppose we want a term of the form $f(x)z$, where z is metric.
- `s(x,by=z)` achieves this.
- An interaction with a factor variable, *a*, is also possible.
- `s(x,by=a)` produces a smooth of x for each level of *a*.
- `s(x,by=a,id="foo")` forces all these smooths to have the same smoothing parameter.
- `s(x,a,bs="fs")` is similar, but efficient with `gamm` and `gamm4` when *a* has many levels.
- `te/2` terms also accept `by` variables.

The basis dimension

- You have to choose the number of basis functions to use for each smooth, using the `k` argument of `s` or `te`.
- The default is essentially arbitrary.
- Provided `k` is not too small its exact value is not critical, as the smoothing parameters control the actual model complexity. However
  1. if `k` is too small then you will oversmooth.
  2. if `k` is much too large then computation will be very slow.
- Checking that k is not too small will be covered in a later segment.

Miscellanea

- Most smooths will require an identifiability condition to avoid confounding with the model intercept: `gam` handles this by automatic reparameterization.
- `gam` will also handle the side conditions required for nested smooths. e.g. `gam(y~s(x)+s(z)+s(x,z))` will work.
- However, such nested models are not always easy to interpret.
- `te`, `t2`, `s(...,bs="tp")` and `s(...,bs="ds")` can, in principle, handle any number of covariates.
- The `"ad"` basis can handle 1 or 2 covariates, but no more.

Discrete spatial smoothing: Markov random fields

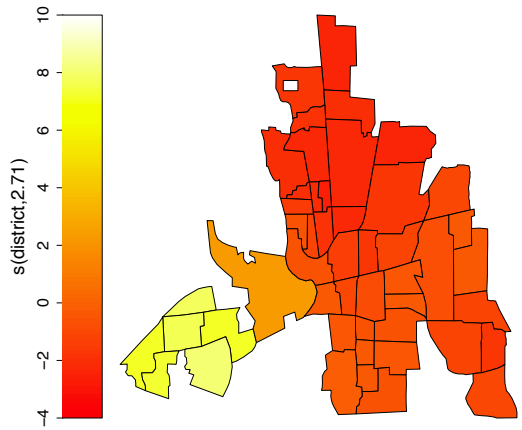
- Sometimes data come allocated to irregular partitions of space (e.g. administrative regions).
- Markov random fields area a popular way of smoothing such data.
- The smooth has a coefficient, γ_i, for each region.
- The neighbouring regions of each region are found, and a quadratic penalty constructed. If N_i is the set of indices of the neighbours of region i, then the simplest penalty is

$$\sum_i (\sum_{j \in N_i} (\gamma_i - \gamma_j))^2$$

- Eigen based rank reduction is effective here.

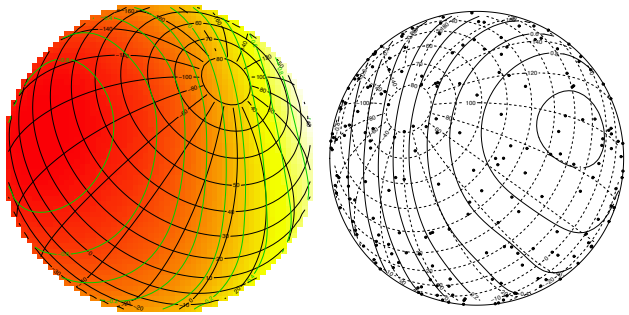
Markov random field illustration

```
data(columb.polys) ## district shapes list
xt <- list(polys=columb.polys)
gam(crime ~ s(district,bs="mrf",xt=xt),data=columb)
```



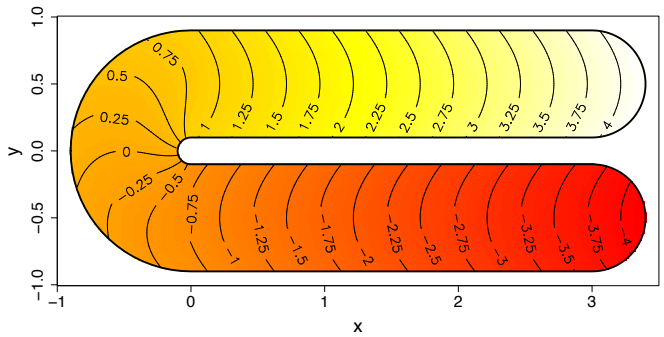
Smoothing on the globe

- Thin plate spline like smoothers can be constructed for the sphere [`s(la,lo,bs="sos")`]...



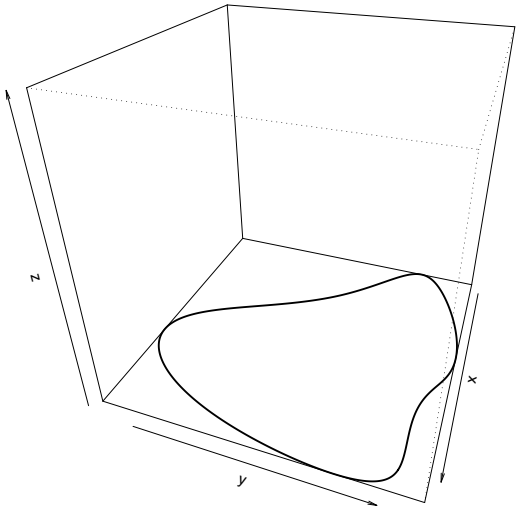
Finite area smoothing

- Suppose now want to smooth samples from this function

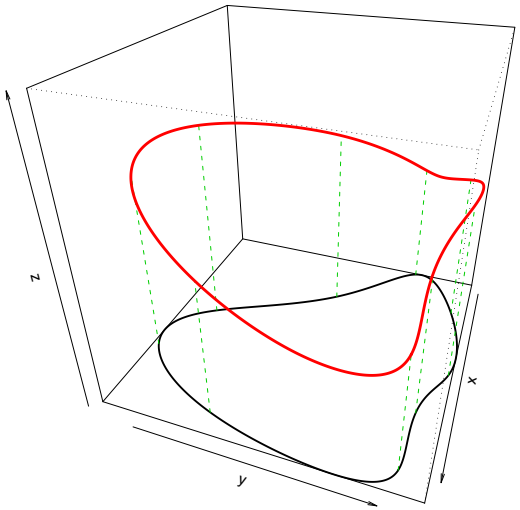


- . . . without ‘smoothing across’ the gap in the middle?
- Let’s use a soap film . . .

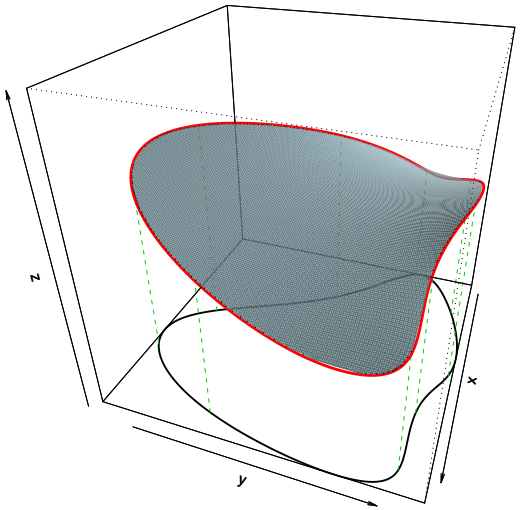
The domain



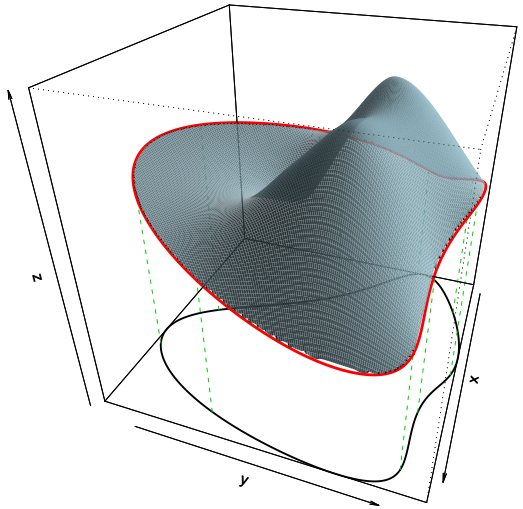
The boundary condition



The boundary interpolating film

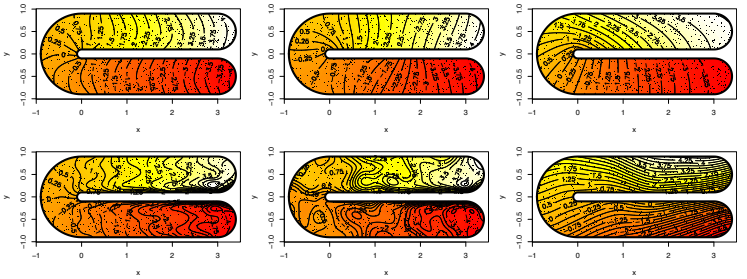


Distorted to approximate data



Soap film smoothers `s(...,bs="so")`

- Mathematically this smoother turns out to have a basis-penalty representation.
- It also turns out to work...



Random effect `s(...,bs="re")`

- Statistically, smooths consist of a basis and a quadratic penalty, where the penalty matrix can be treated as the generalized inverse of a covariance matrix.
- They can therefore be estimated as random effects.
- Reversing this, we can treat simple random effects as (zero dimensional) smooths.
- `s(a,b,bs="re")` creates a terms with model matrix `model.matrix(~a:b-1)` and a scaled identity penalty/covariance matrix.
- Any number of covariates are possible.
- Function `gam.vcomp` helps later interpretation by converting smoothing parameters to variance components.

Summary

- We can treat simple random effects as 0 dimensional smooths.
- In 1 dimension, the choice of basis is not critical. The main decisions are whether it should by cyclic or not and whether or not it should be adaptive.
- In 2 dimensions and above the key decision is whether an isotropic smooth, `s`, or a scale invariant smooth, `ti`/`te`/`t2`, is appropriate. (`te`/`i`/`2` terms may be isotropic in some marginals.)
- Smooths and factors can be made to interact.
- Spatial smoothing may sometimes require more specialized smoothers (Markov random fields, spherical splines, finite area smooths).
- The basis dimension is a modelling decision that should be checked.