Using multidimensional scaling with Duchon splines for reliable finite area smoothing

David Lawrence Miller

Mathematical Sciences
University of Bath

Recent advances in spatial statistics in ecology
5 May 2011
St Andrews

Outline

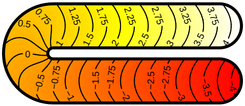
Outline

Smoothing in 2 dimensions

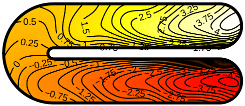
- Have some geographical region and wish to find out something about the biological population in it.
- Response is eg. animal distribution, wish to predict based on (x, y) and other covariates eg. habitat, size, sex, etc.
- This problem is relatively easy if the domain is simple.

Leakage (or, “whales don’t live in glaciers”)

- Smoothing of complex domains makes this a lot more difficult.
- Problem of leakage.
- Wrong metric.
- Euclidean distance doesn’t always make sense.



(modified) Ramsay test function



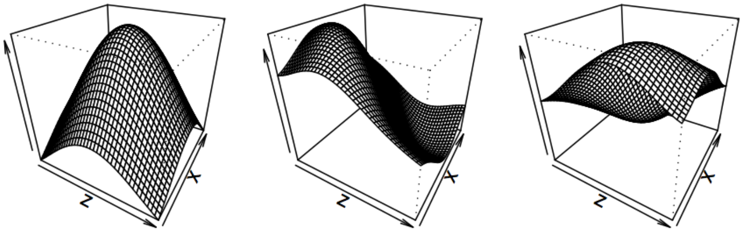
Thin plate spline fit

Smoothing with penalties

- Do smoothing with splines in an additive model framework.
- Take some linear combination of (known) basis functions.
- Penalize based on integral of second squared derivative:

$$\int\int_{\mathbf{R}^2}\left(\left(\frac{\partial^2 f(x_1,x_2)}{\partial x_1^2}\right)^2+\left(\frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2}\right)^2+\left(\frac{\partial^2 f(x_1,x_2)}{\partial x_2^2}\right)^2\right)\mathrm{d}x_1\mathrm{d}x_2.$$

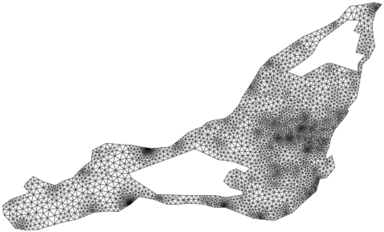
- Using thin plate regression splines (eg Wood (2003)).

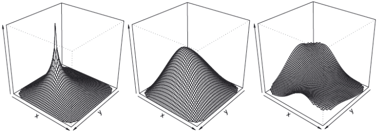


Proposed solutions to leakage problems

Two categories: PDE boundary condition based, within-area distance based.

- PDEs:
  - FELSPLINE (Ramsay, 2002).
  - Soap film smoothers (Wood *et al*, 2008).
- Within-area distance
  - Geodesic Low-Rank Thin Plate Splines (GLTPS) (Wang and Ranalli, 2007).



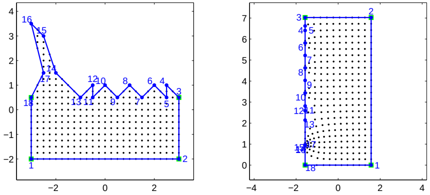


“The Third Way”

- Domain morphing - think of domain as silly putty.
- Takes into account within-area distance.
- Potentially less computationally intensive.
- Eilers (2006) proposes Schwarz-Christoffel transform.
- Gives a known domain - easier to smooth over.

However:

- Crowding - numerical singularities.
- Not clear what this does to the smoothness penalty - artefacts, penalty issues - but NOT GOOD!



Outline

Multidimensional scaling and within-area distances

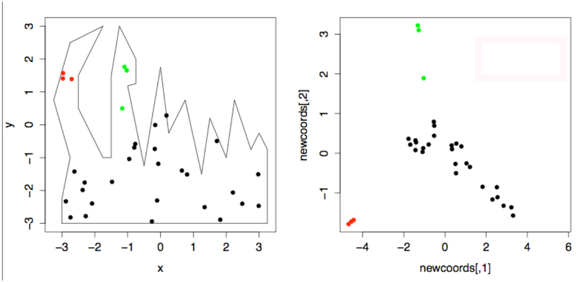
- Idea: use MDS to arrange points in the domain according to their distance within the domain.

Scheme:

- First need to find the within-area distances.
- Perform MDS on the matrix of within-area distances.
- Find the new configuration of the points.
- Smooth over the new points.

Multidimensional scaling refresher

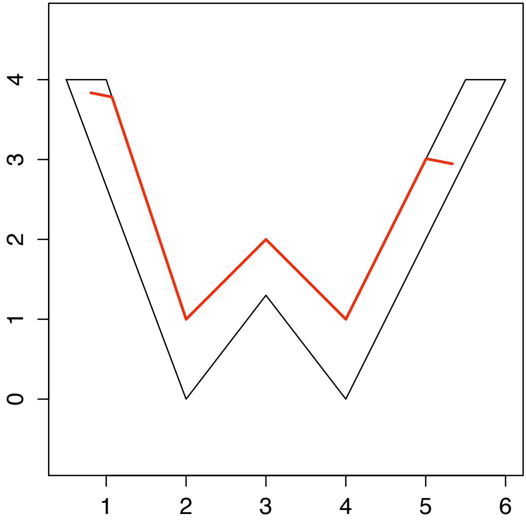
- Use double centred matrix of between point distances, D, using the eigen-decomposition of DD^T we can find new points.
- Finds a projection of points such that Euclidean distance between points in new arrangement is approximately the same as the distances in D.
- New points can be added into the MDS configuration via Gower's interpolation (Gower, 1968)



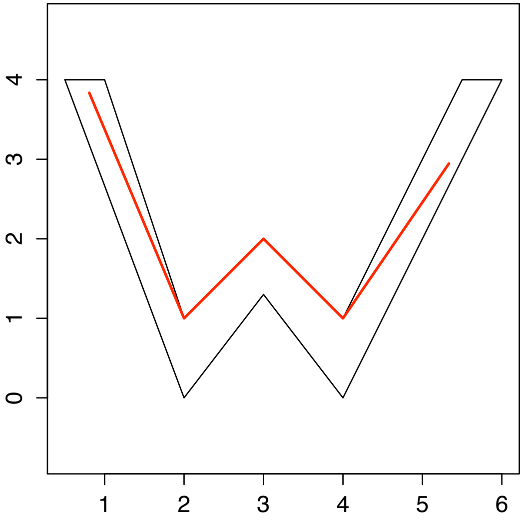
A “new” algorithm for finding within-area distances

So, how do we find the distances to put in D?

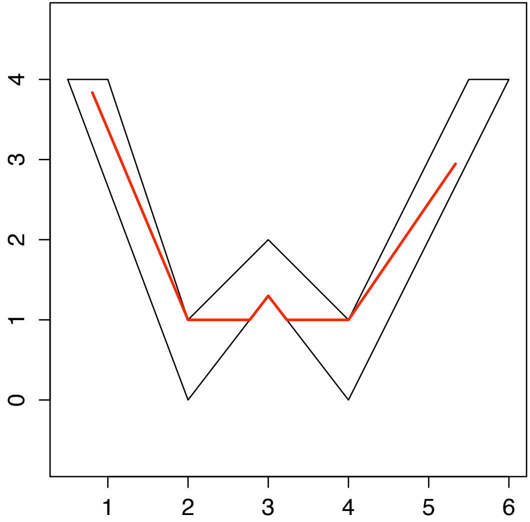
Finding the path - 1



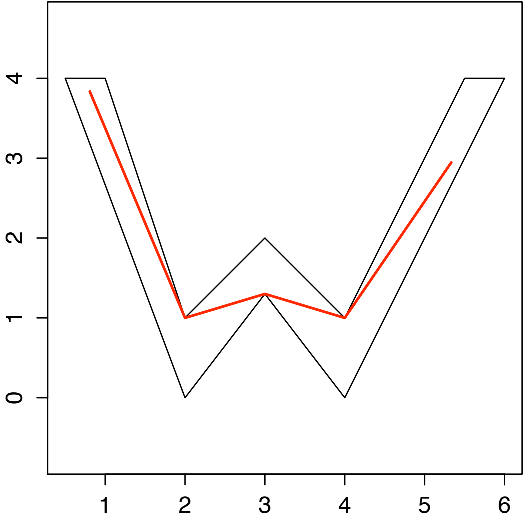
Finding the path - 2



Finding the path - 3



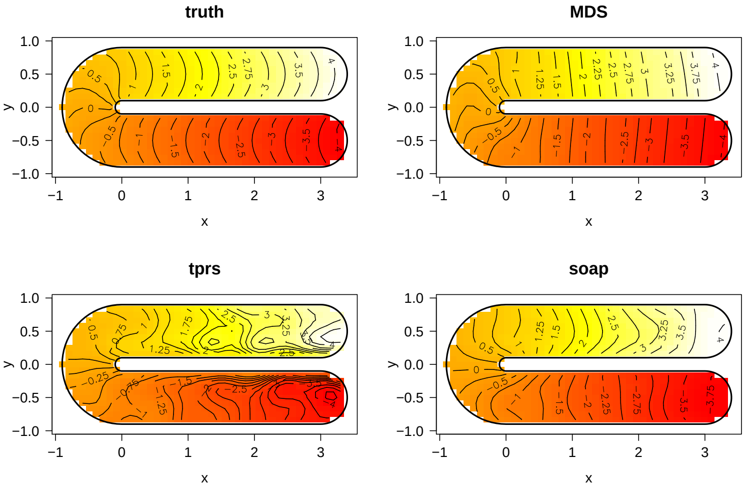
Finding the path - 4



Simulations

But how does it perform?

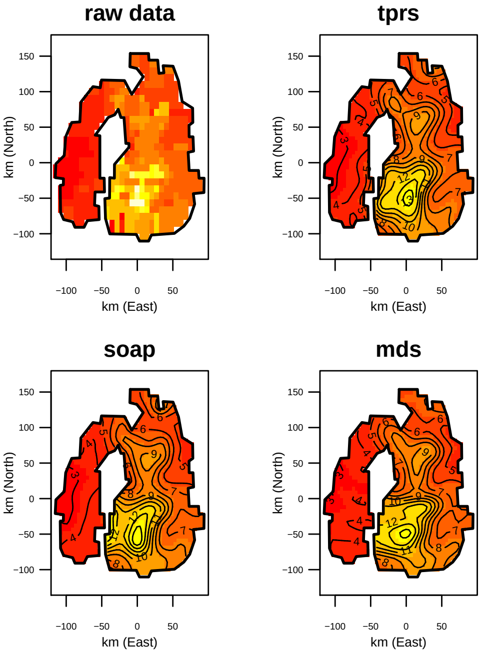
Ramsay simulations



The Aral sea

ARAL SEA PICTURE

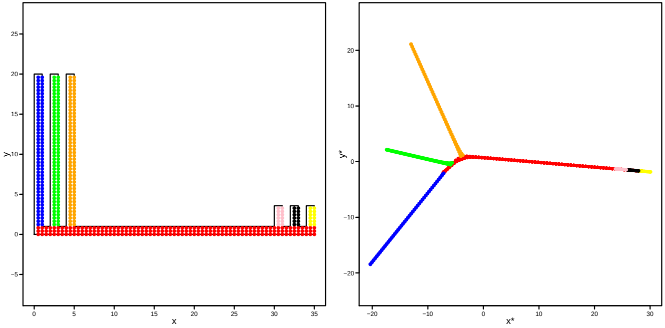
The Aral sea



But...

- Computationally costly to find the within-area distances. (Esp. prediction.)
- Comparative in MSE terms to the soap film smoother.
- Artefacts – ordering?
- Taking a m-D projection of n-D space.

[[3D PLOT]]



Ordering

- So, where are the artefacts coming from?
- Taking projections of n-dimensional space.
- Points can lose ordering.
- 3-D projection does better, but can we keep going? (Nullspace gets BIG!)

Outline

High dimensional smoothing with Duchon splines

How can we fix this?

Nullspaces

- TPRS functions of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_{md}(\mathbf{x} - \mathbf{x_i}) + \sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x})$$

$$= \text{radial basis functions} + \text{“planes”}$$

- $\phi_j(\mathbf{x})$ are linearly independent polynomials of degree $< m$.

WHAT IS m, d etc!!!

When nullspaces get big...(I)

- For TPRS, nullspace size is

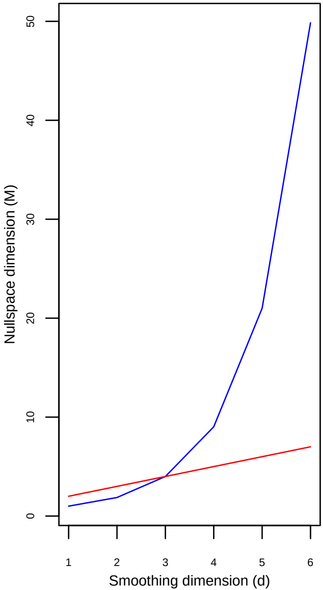
$$M = \binom{m+d-1}{d}$$

  Where

$$m = \text{the derivative order in the penalty}$$
$$d = \text{smoothing dimension}$$

- **AND** must obey $2m > d$
  smoothing dimension++ $\Rightarrow$ derivative order++++.
- M gets very big, very fast.

When nullspaces get big... (II)



A brief digression into Fourier transforms (I)

- Given some function, we can take the Fourier transform:

$$\mathfrak{F}g(\boldsymbol{\tau}) = \int \dots \int_{\mathbb{R}^n} e^{2\pi\sqrt{-1}\mathbf{x}^\intercal \boldsymbol{\tau}} g(\mathbf{x})\mathsf{d}\mathbf{x}.$$

- Can think of this as a change of basis **OR** as decomposing the function into frequency components.

THINK ABOUT THIS!!!

Enter Jean Duchon...

SPLINES MINIMIZING ROTATION-INVARIANT

SEMI-NORMS IN SOBOLEV SPACES

Jean DUCHON

Université Scientifique et Médicale

Laboratoire de Mathématiques Appliquées

B.P. 53 - 38041 Grenoble - France

ABSTRACT

We define a family of semi-norms $\|u\|_{m,s} = \left(\int_{\mathbb{R}^n} |\tau|^{2s} |\mathcal{F}\, D^m u(\tau)|^2 \, d\tau \right)^{1/2}$. Minimizing such semi-norms, subject to some interpolating conditions, leads to functions of very simple forms, providing interpolation methods that : 1°) preserve polynomials of degree $\leq m-1$; 2°) commute with similarities as well as translations and rotations of $\mathbb{R}^n$; and 3°) converge in Sobolev spaces $H^{m+s}(\Omega)$.

High dimensional smoothing with Duchon splines

Wood (2003) type penalty:

$$J_{m,d} = \int \dots \int_{\mathbb{R}^d} \sum_{\nu_1+\dots+\nu_d=m} \frac{m!}{\nu_1! \dots \nu_d!} \left(\frac{\partial^m f(x_1, \dots, x_d)}{\partial x_1^{\nu_1} \dots \partial x_d^{\nu_d}} \right)^2 \mathrm{d}x_1 \dots \mathrm{d}x_d$$

Why not take the Fourier transform? (Equivalent by Plancherel theorem)

$$\breve{J}_{m,d} = \int \dots \int_{\mathbb{R}^d} \sum_{\nu_1+\dots+\nu_d=m} \frac{m!}{\nu_1! \dots \nu_d!} \left(\mathfrak{F} \frac{\partial^m f}{\partial x_1^{\nu_1} \dots \partial x_d^{\nu_d}} (\boldsymbol{\tau}) \right)^2 \mathrm{d}\boldsymbol{\tau}$$

Putting in some arbitrary weighting, $w(\boldsymbol{\tau})$:

$$\breve{J}_{m,d} = \int \dots \int_{\mathbb{R}^d} w(\boldsymbol{\tau}) \sum_{\nu_1+\dots+\nu_d=m} \frac{m!}{\nu_1! \dots \nu_d!} \left(\mathfrak{F} \frac{\partial^m f}{\partial x_1^{\nu_1} \dots \partial x_d^{\nu_d}} (\boldsymbol{\tau}) \right)^2 \mathrm{d}\boldsymbol{\tau}$$

Duchon splines penalty

Let $w(\boldsymbol{\tau}) = |\boldsymbol{\tau}|^{2s}$ (Duchon 1977)

$$\breve{J}_{m,d} = \int \dots \int_{\mathbb{R}^n} |\boldsymbol{\tau}|^{2s} \sum_{\nu_1+\dots+\nu_d=m} \frac{m!}{\nu_1! \dots \nu_d!} \left(\mathfrak{F} \frac{\partial^m f}{\partial x_1^{\nu_1} \dots \partial x_d^{\nu_d}} (\boldsymbol{\tau}) \right)^2 \mathrm{d}\boldsymbol{\tau}$$

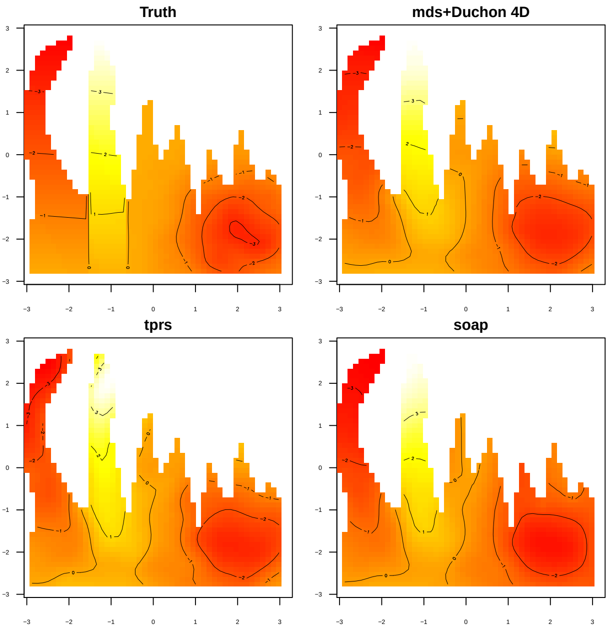
- This penalises high frequency components of the derivatives.
- Now the condition relating d and m is $d/2 - m \leq s$.
- Given we want to use $m = 2$ (some pseudo-theoretical reasons for that) and we have a set d, determine s by

$$s = d/2 - 2 \tag{1}$$

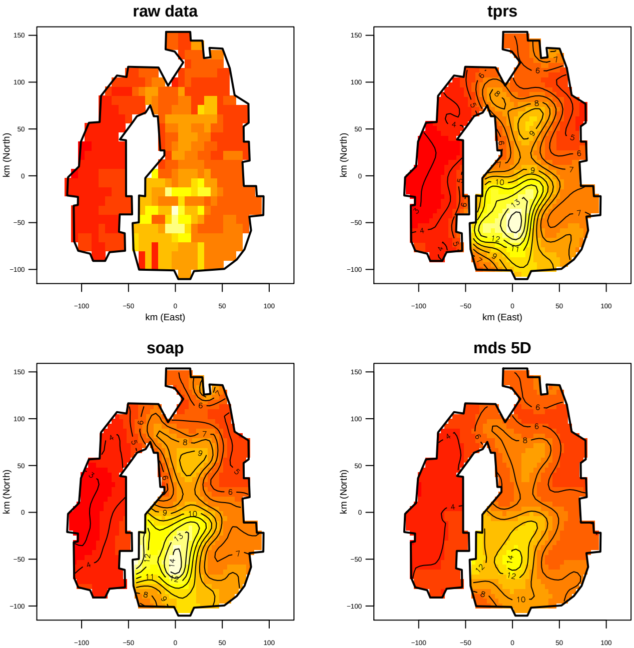
Using ML/GCV score to select projection dimension

- So, we can now smooth in higher dimensions, **but how many?**
- Select MDS dimension using GCV or ML score.
- Fit model in a range of dimensions, pick best score.
- No monotonicity guarantees.

Pretend coastline



Aral sea



raw data
tprs
soap
mds 5D
km (North)
km (East)
150
100
50
0
−50
−100
−100
−50
0
50
100

Outline

General distance smoothing

- Given a general distance matrix, D.
- Project the points in D using MDS.
- Smooth over those points.

Examples:

- Voting data
- Microarrays
- Socio-economic variables

Outline

Final comments

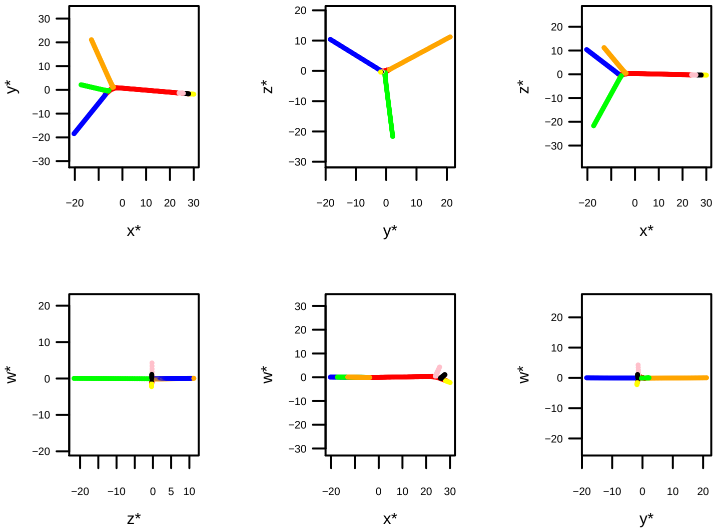
- Something
- Something
- Something

References

- T. Ramsay. *Spline smoothing over difficult regions. JRSSB, 2002*
- P.H.C. Eilers. *P-spline smoothing on difficult domains. University of Munich seminar, 2006*
- J.C. Gower. *Adding a point to vector diagrams in multivariate analysis. Biometrika, 1968.*
- Duchon, J. (1977). *Splines minimizing rotation-invariant semi-norms in Sobolev spaces. Constructive theory of functions of several variables.* Springer.

Slides available at `http://people.bath.ac.uk/dlm27`

Appendix - Ordering problems in 4-D



Appendix - A “new” algorithm for finding within-area distances

- Algorithm “bounces” around inside the polygon.
- Initial path is just the path around the edge, then iterate over two steps: *delete* and *alter*.
- Delete (iterating over all nodes):
  - If we can shorten the path by simply deleting a node, do that.
- Alter (iterating over all nodes):
  - If we can find a shorter sub-path by bouncing off the other “side” of the polygon, do that.