Low-rank smoothing splines for complex domains and manifold recovery

M. Giovanna Ranalli
Department of Statistics, Colorado State University
Joint work with Haonan Wang, CSU
Outline

1. Low-rank smoothing splines for complex domains
   - Motivation
   - Low-rank thin plate splines
   - Geodesic Low-rank thin plate splines
Outline

1. Low-rank smoothing splines for complex domains
   • Motivation
   • Low-rank thin plate splines
   • Geodesic Low-rank thin plate splines

2. Low-rank smoothing splines for manifold recovery
   • Motivation
   • Examples of nonfunctional manifolds
   • How essentially the same idea employed in 1. can be used in these situations
Motivation

New Hampshire Estuary – 97 sites where mercury in sediment concentrations has been surveyed in the years 2000/1 and 2003 (data from NHNCA and NHDES)
Needs and Problems
Needs and Problems

NEEDS –

• Mapping quantities of interest, such as pollutants, by making predictions at non-observed locations
• Simple way to introduce covariates other than geographical coordinates
Needs and Problems

NEEDS –

- Mapping quantities of interest, such as pollutants, by making predictions at non-observed locations
- Simple way to introduce covariates other than geographical coordinates

Bivariate Smoothers like thin plate splines and kriging would obtain a map by employing covariance functions between locations that depend on their Euclidean distance
Needs and Problems

NEEDS –

- Mapping quantities of interest, such as pollutants, by making predictions at non-observed locations
- Simple way to introduce covariates other than geographical coordinates

Bivariate Smoothers like thin plate splines and kriging would obtain a map by employing covariance functions between locations that depend on their Euclidean distance

PROBLEMS –

- Irregularly shaped non-convex domains
- Euclidean distance might not be a good way to measure similarity between data points
- Using different distance metrics in kriging does not guarantee a positive definite covariance matrix (Rathbun, 1998; Gardner et al., 2003; Ver Hoef et al., 2004)
Low-rank thin plate splines – LTPS

Data from the example of the estuary are of the form \((x_i, y_i)\), for \(i = 1, \ldots, N\), with \(x_i\) geographical locations and \(y_i\) measurements of a variable of interest. Bivariate smoothing assumes that

\[
y = f(x) + \varepsilon,
\]

where \(y = (y_1, \ldots, y_N)^T\), \(f(\cdot)\) is some unspecified smooth function of \(x\) and the distribution of the errors is given by \(\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)\).
Low-rank thin plate splines – LTPS

Data from the example of the estuary are of the form \((x_i, y_i)\), for \(i = 1, \ldots, N\), with \(x_i\) geographical locations and \(y_i\) measurements of a variable of interest. Bivariate smoothing assumes that

\[ y = f(x) + \varepsilon, \quad (1) \]

where \(y = (y_1, \ldots, y_N)^T\), \(f(\cdot)\) is some unspecified smooth function of \(x\) and the distribution of the errors is given by \(\varepsilon \sim \mathcal{N}(0, \sigma^2 \varepsilon I)\).

Ruppert et al. (2003) advocate the use of a mixed models–low rank representation of the problem to

1. speed computation
2. make computation easy through mixed models software
3. insert other covariates in the fixed part (parametric continuous or factors) or in the random part (nonparametric continuous and random effects)
LTPS: the model and the predictions

The mixed model representation of model (1) is

\[ y = X\beta + Zu + \varepsilon, \]  

(2)
LTPS: the model and the predictions

The mixed model representation of model (1) is

\[ y = X\beta + Zu + \varepsilon, \tag{2} \]

where

- \( X = [1 \ x_i]_{1 \leq i \leq N} \)
- \( Z \) contains \( T \leq N \) radial basis functions for the estimation of the non-linear structure of \( f(\cdot) \)
- \( u \sim \mathcal{N}(0, \sigma_u^2 I) \) are random effects independent of \( \varepsilon \)

This type of models can be fitted using PROC MIXED in SAS or the \texttt{lme()} function in Splus
The $Z$ matrix

\[
Z = \left[ C(\|x_i, \kappa_t \|_E) \right]_{1 \leq i \leq N} \left[ C(\|\kappa_t, \kappa_{t'} \|_E) \right]^{-1/2}_{1 \leq t, t' \leq T},
\]
The **Z** matrix

\[
Z = \left[ C(\|x_i, \kappa_t\|_E) \right]_{1 \leq i \leq N}^{1 \leq i \leq N} \left[ C'(\|\kappa_t, \kappa_{t'}\|_E) \right]_{1 \leq t \leq T}^{1 \leq t \leq T}^{-1/2}
\]

where

- \( \kappa_1, \ldots, \kappa_T \) is a set of knot locations (next slide)
- \( \| \cdot \|_E \) denotes Euclidean distance
- the function \( C \) is given by \( C(r) = r^{2m-d} \log r \), if \( d \) is even and \( C(r) = r^{2m-d} \), if \( d \) is odd, with \( d \) dimension of the predictors’ space and \( m > 1 \) controlling the smoothness of \( f \). If \( m = 2 \) we are penalizing the second derivative of \( f \).
The Z matrix

\[
Z = \left[ C(||x_i, \kappa_t||_E) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C'(||\kappa_t, \kappa_{t'}||_E) \right]_{1 \leq t, t' \leq T}^{-1/2}
\]  

where

- \( \kappa_1, \ldots, \kappa_T \) is a set of knot locations (next slide)
- \( || \cdot ||_E \) denotes Euclidean distance
- the function \( C \) is given by \( C(r) = r^{2m-d} \log r \), if \( d \) is even and \( C(r) = r^{2m-d} \), if \( d \) is odd, with \( d \) dimension of the predictors’ space and \( m > 1 \) controlling the smoothness of \( f \). If \( m = 2 \) we are penalizing the second derivative of \( f \).

IF \( T = N \) \( \sim \) knots ≡ observations and FULL-RANK case (Thin plate splines)
The Z matrix

\[
Z = \left[ C(||x_i, \kappa_t||_E) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C(||\kappa_t, \kappa_{t'}||_E) \right]^{-1/2}_{1 \leq t, t' \leq T},
\]

where

- \( \kappa_1, \ldots, \kappa_T \) is a set of knot locations (next slide)
- \( || \cdot ||_E \) denotes Euclidean distance
- the function \( C \) is given by \( C(r) = r^{2m-d} \log r \), if \( d \) is even and \( C(r) = r^{2m-d} \), if \( d \) is odd, with \( d \) dimension of the predictors’ space and \( m > 1 \) controlling the smoothness of \( f \). If \( m = 2 \) we are penalizing the second derivative of \( f \).

IF \( T = N \) \( \Rightarrow \) knots\( \equiv \) observations and FULL-RANK case (Thin plate splines)

IF \( T = N \) & \( C(r) \) some Matérn, exponential, gaussian corr functions \( \Rightarrow \) FULL-RANK Kriging
Knots – 2 issues: how many & where

**HOW MANY** rule of thumb: 1 every 3-4 observations, never more than 100.

**WHERE** rectangular lattices, regular grids on the domain or space filling designs
(FUNFITS in Splus and FIELDS in R do space filling)
Knots – 2 issues: how many & where

HOW MANY rule of thumb: 1 every 3-4 observations, never more than 100.

WHERE rectangular lattices, regular grids on the domain or space filling designs (FUNFITS in Splus and FIELDS in R do space filling)
Predictions

Once estimates from model (2) for $\beta$ and predictions for $u$ are obtained through Maximum Likelihood or REstricted ML, predicted values at observed locations are given by

$$\hat{y} = X\hat{\beta} + Z\hat{u}$$
Predictions

Once estimates from model (2) for $\beta$ and predictions for $u$ are obtained through Maximum Likelihood or REstricted ML, predicted values at observed locations are given by

$$\hat{y} = X\hat{\beta} + Z\hat{u}$$

The Splus commands for this would be simply

```r
fit<-lme(y~-1+X, random=pdIdent(~-1+Z))
beta<-fit$coef$fixed
u<-fit$coef$random
predictions<-X%*%beta+Z%*%u
```
Predictions

Once estimates from model (2) for $\beta$ and predictions for $u$ are obtained through Maximum Likelihood or REstricted ML, predicted values at observed locations are given by

$$\hat{y} = X\hat{\beta} + Z\hat{u}$$

The Splus commands for this would be simply

```splus
fit<-lme(y∼-1+X, random=pdIdent(∼-1+Z))
beta<-fit$coef$fixed
u<-fit$coef$random
predictions<-X%*%beta+Z%*%u
```

Predictions at other locations can be done by adding new rows to $X$ and $Z$ to include the new prediction points.
LTPS fit of the estuary data

Recall Needs and Problems
Do we really need a different distance metric??

Monte Carlo simulation: real function
Simulation results: average prediction error

LTPS error (top) and GLTPS error (bottom)

Recall estuary 2
Geodesic LTPS

Change the Euclidean distance measure in the $Z$ matrix in (3) with the GEODESIC DISTANCE = THE SHORTEST PATH A FISH WOULD SWIM
Geodesic LTPS

Change the Euclidean distance measure in the $Z$ matrix in (3) with the GEODESIC DISTANCE = THE SHORTEST PATH A FISH WOULD SWIM

$$Z = \left[ C(||x_i, \kappa_t||_G) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C(||\kappa_t, \kappa_{t'}||_G) \right]_{1 \leq t, t' \leq T}^{-1/2}$$
Geodesic LTPS

Change the Euclidean distance measure in the $Z$ matrix in (3) with the GEODESIC DISTANCE = THE SHORTEST PATH A FISH WOULD SWIM

$$Z = \left[ C(||x_i, \kappa_t||_G) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C(||\kappa_t, \kappa_{t'}||_G) \right]_{1 \leq t, t' \leq T}^{-1/2}$$

- The geodesic distance is estimated by means of the Floyd Algorithm:
  1. build a graph for which nodes are locations and each node is connected to its $nn$ nearest neighbors;
  2. obtain the shortest path between two locations and get the geodesic distance as the length of such path.
- Small $nn$ → the Graph might not be connected; too big $nn$ → Euclidean distance. ⇒ take the smallest $nn$ for which the Graph is connected.
- The density of the data influences the final estimate.
Floyd Algorithm for the Northern part of the estuary
GLTPS fit of the estuary

<table>
<thead>
<tr>
<th></th>
<th>GLTPS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>p-value</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.164</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>F(Year03)</td>
<td>0.090</td>
<td>0.011</td>
</tr>
<tr>
<td>(\sigma_u)</td>
<td>0.0016</td>
<td>0.011</td>
</tr>
</tbody>
</table>
# AIC of models and standard deviations of predictions

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X=1,\text{lat,lon,year}$</td>
<td>-55.258</td>
</tr>
<tr>
<td>$X=1,\text{lat,lon}$</td>
<td>-55.122</td>
</tr>
<tr>
<td>$X=1,\text{year}$</td>
<td>-72.477</td>
</tr>
</tbody>
</table>

AIC: smaller is better

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st.Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd.Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed HG</td>
<td>0.010</td>
<td>0.110</td>
<td>0.210</td>
<td>0.232</td>
<td>0.300</td>
<td>0.766</td>
</tr>
<tr>
<td>pred at obs locations</td>
<td>0.163</td>
<td>0.172</td>
<td>0.186</td>
<td>0.232</td>
<td>0.267</td>
<td>0.399</td>
</tr>
<tr>
<td>stdev at obs locations</td>
<td>0.006</td>
<td>0.011</td>
<td>0.020</td>
<td>0.018</td>
<td>0.021</td>
<td>0.031</td>
</tr>
<tr>
<td>pred at nonobs locations</td>
<td>0.250</td>
<td>0.260</td>
<td>0.262</td>
<td>0.271</td>
<td>0.268</td>
<td>0.436</td>
</tr>
<tr>
<td>stdev at nonobs locations</td>
<td>0.148</td>
<td>0.148</td>
<td>0.148</td>
<td>0.148</td>
<td>0.148</td>
<td>0.152</td>
</tr>
</tbody>
</table>
LTPS vs GLTPS fit of the estuary data

- Airborne deposition vs different patterns
- Great Bay and Cocheco River
- Recall estuary 2
The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;
The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;

2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)
The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;

2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)

3. Estimate the matrix of geodesic distances from this grid with the Floyd Algorithm starting with \( nn = 3 \) and then increasing \( nn \) until the graph is connected
The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;

2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)

3. Estimate the matrix of geodesic distances from this grid with the Floyd Algorithm starting with \( nn = 3 \) and then increasing \( nn \) until the graph is connected

4. Calculate the \( X_P \) and the \( Z_P \) matrices for all the grid points
The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;

2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)

3. Estimate the matrix of geodesic distances from this grid with the Floyd Algorithm starting with $nn = 3$ and then increasing $nn$ until the graph is connected

4. Calculate the $X_P$ and the $Z_P$ matrices for all the grid points

5. Obtain the $X$ and the $Z$ matrices for the observed data locations as a subset of rows of $X_P$ and $Z_P$ (time saver and more accurate)
The steps to obtain GLTPS

1. Determine number and location of the knots from the observed data locations through a space filling design;

2. Lay down a reasonably dense grid of locations on the domain to get the matrix of geodesic distances (and eventually predictions)

3. Estimate the matrix of geodesic distances from this grid with the Floyd Algorithm starting with \( nn = 3 \) and then increasing \( nn \) until the graph is connected

4. Calculate the \( X_P \) and the \( Z_P \) matrices for all the grid points

5. Obtain the \( X \) and the \( Z \) matrices for the observed data locations as a subset of rows of \( X_P \) and \( Z_P \) (time saver and more accurate)

6. Fiddle with \( lme \) models. Issues:
   - other covariates: size of \( X_P \);
   - tests for the significance of the covariates are carried in the usual way;
   - tests for the significance of the random components (i.e. the spatial component) if done within the mixed models framework can be really conservative (alternative: Crainiceanu & Ruppert, 2004, for one variance component or bootstrap for more than one).
Outline

1. Low-rank smoothing splines for complex domains
   • Motivation
   • Low-rank thin plate splines
   • Geodesic Low-rank thin plate splines

2. Low-rank smoothing splines for manifold recovery
   • Motivation
   • Examples of nonfunctional manifolds
   • How essentially the same idea employed in 1. can be used in these situations
Motivation: data mining framework

- Understand the structure of large high dimensional data
- Handle nonlinear structures (nonlinear manifolds)
- Visualization in lower dimensional spaces
- Classification and clustering – Image recognition
Manifold recovery: Swiss-roll and S examples
NEW Needs and Problems
NEW Needs and Problems

NEEDS –

- Recover the manifold shape from noisy data
NEW Needs and Problems

NEEDS –

• Recover the manifold shape from noisy data

PROBLEMS –

• Bivariate smoothers run into difficulties when the manifold is not a functional relationship of the locations, see model (1).
• Both Euclidean and geodesic distance on the domain are not of use for shape recovery
Proposed approach: 3LTPS and 3GLTPS

Recall the key modification employed in GLTPS: 12.
Proposed approach: 3LTPS and 3GLTPS

Recall the key modification employed in GLTPS: 12.

Use the Euclidean or the geodesic distance in the manifold space, instead of the domain space:

\[
Z = \left[ C'\left( ||z_i, \kappa_t||_{E,G} \right) \right]_{1 \leq i \leq N}^{1 \leq t \leq T} \left[ C'\left( ||\kappa_t, \kappa_{t'}||_{E,G} \right) \right]_{1 \leq t, t' \leq T}^{-1/2}
\]

where \( z_i = (x_i, y_i) \) and knots are chosen on the 3D space. Assume for the moment that we can calculate both the Euclidean and the geodesic distance on the manifold, \( nn=3 \).
Proposed approach: 3LTPS and 3GLTPS

Recall the key modification employed in GLTPS: 12.

Use the Euclidean or the geodesic distance in the manifold space, instead of the domain space:

\[
Z = \left[ C(\|z_i, \kappa_t\|_{E,G}) \right]_{1 \leq i \leq N} \left[ C(\|\kappa_t, \kappa_{t'}\|_{E,G}) \right]_{1 \leq t, t' \leq T}^{-1/2},
\]

where \( z_i = (x_i, y_i) \) and knots are chosen on the 3d space. Assume for the moment that we can calculate both the Euclidean and the geodesic distance on the manifold, \( nn=3 \)
3LTPS and 3GLTPS fit of the Swiss roll
3GLTPS fit of the S manifold

Noisy data

3LTPS – Eucl

3GLTPS – Geo
A key issue: estimating 3d distances from noisy data

HIGH QUALITY DATA  We are still ok, the estimates are not as smooth as before, but still reasonable

HIGH NOISE  Unreliable and wriggly estimates → we are now trying to work out estimates of these distances by smoothing the output of the Floyd algorithm.
Wrap up

GLTPS

1. It is possible to account for non regular domains when mapping quantities of interest

2. The LTPS framework allows inserting other covariates available for all prediction locations in a parametric or nonparametric way easily and guarantees positive definite *covariance* functions

3. Applications other than estuaries include stream networks, domains with holes and irregular boundaries (lakes with islands/land with lakes), response over a nonflat domain (measurements on mountains) ...

4. Other distance measures can be employed
**Wrap up**

**GLTPS**

1. It is possible to account for non regular domains when mapping quantities of interest

2. The LTPS framework allows inserting other covariates available for all prediction locations in a parametric or nonparametric way easily and guarantees positive definite *covariance* functions

3. Applications other than estuaries include stream networks, domains with holes and irregular boundaries (lakes with islands/land with lakes), response over a nonflat domain (measurements on mountains) ...

4. Other distance measures can be employed

**3LTPS and 3GLTPS**: Nonfunctional manifolds can be recovered if observational data is high quality or, if not, if we have a priori information about the manifold.
Current and future work

GLTPS Add remote sensing covariates to the modeling of mercury in NH estuary (land cover) — altern: multiphase sampling for less expensive covariates to be collected on site (grain size)

GLTPS Application on stream networks

GLTPS Fix the Floyd Algorithm to allow only for on-water paths and for flow direction
Current and future work

**GLTPS** Add remote sensing covariates to the modeling of mercury in NH estuary (land cover) — altern: multiphase sampling for less expensive covariates to be collected on site (grain size)

**GLTPS** Application on stream networks

**GLTPS** Fix the Floyd Algorithm to allow only for on-water paths and for flow direction

**3GLTPS** Obtain reliable estimates of distances from high noise level data

**3GLTPS** Application to higher dimensional good quality datasets

**3GLTPS** Manifold recovery under rotation
Essential bibliography


Ver Hoef, J.M., Peterson, E. and Theobald D. (2004), Spatial statistical models that use flow and stream distance *Manuscript*.

ACKNOWLEDGEMENTS: Hal Walker (EPA) and Phil Townbridge (NHDES) for providing the NH data, Jay Breidt for helpful discussion. The work reported here was developed under the STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of the presenter and STARMAP. EPA does not endorse any products or commercial services mentioned in this presentation.