Overview of Spatial Statistics with Applications to fMRI

Stefano Castruccio

School of Mathematics & Statistics
Newcastle University

April 8<sup>th</sup>, 2016
Outline

- Why spatial statistics?
- Basic results
- Nonstationary models
- Inference for large data sets
- An example of application
‘The best way to account for spatial dependence is to avoid doing it.’ (Michael Stein)
Why Spatial Statistics for neurological data?

Figure: Time series of an fMRI scan in one voxel
Figure: fMRI time series for neighboring voxels
Figure: single voxel vs averaging
However...

**Figure:** fMRI time series for neighboring voxels, a discontinuous case
Spatial statistics is different from temporal statistics because the questions we want to answer are different.

- In temporal statistics you want to forecast the future given the past. We want to know the distribution of $Y_{t+k}$ given \{$Y_t, \ldots, Y_1$\}.
- In spatial statistics, there is no future.
Some theory

Let $D \subset \mathbb{R}^2$, then $Y_s$ where $s \in \mathbb{R}^2$ is a **Gaussian random field** if $\{Y_{s_1}, \ldots, Y_{s_n}\}$ is a multivariate (scalar) Gaussian random vector for every $\{s_1, \ldots, s_n\} \subset D$.

Here $Y_s$ is a point value, but there is also a theory for areal values (e.g. values for different counties).
Main assumption \( Y_s = \sum_{i=1}^{p} f_i(s) \beta_i + \varepsilon_s \), or in matrix form
\[
Y = X\beta + \varepsilon
\]
\( \varepsilon \sim \mathcal{N}(0, K) \), Gaussian process.
Let us assume, for now, that \( X\beta = \mu 1 \): constant mean.
A **Gaussian random field** is stationary if and only if

\[ E(Y_s) = E(Y_{s+u}) = \mu, \]

and in particular

\[ \text{cov}(Y_s, Y_{s'}) = \text{cov}(Y_{s+u}, Y_{s'+u}) = K(u), \]

and in particular

\[ \text{var}(Y_s) = \text{var}(Y_{s+u}) = K(0). \]

\( K \) is called the **covariance function**.
- So $K(u)$ is defined for all locations in $D \subset \mathbb{R}^2$, not fixed lags.
- If $K(u) = K(\|u\|)$ the random field is called **isotropic**: no preferred direction in $\mathbb{R}^2$.
- Stationarity/isotropy almost never occurs, but it can be used as a reference for more complex dependence structures.
Not every function $K(u)$ could be a covariance function for a random field.

$K(u)$ must be positive definite, i.e. for every $\{s_1, \ldots, s_n\}$ and every $\{w_1, \ldots, w_n\}$ we have that

$$\text{var} \left( \sum_{i=1}^{n} w_i Y_{s_i} \right) = \sum_{i=1}^{n} \sum_{i' = 1}^{n} w_i w_i' K(s_i - s_{i'}) \geq 0.$$
Valid covariance, examples

Here are some widely used covariance functions

- squared exponential (not Gaussian): $K(u) = \sigma^2 e^{-\left(\frac{\|u\|}{\phi}\right)^2}$. $\sigma^2$ is the variance and $\phi$ is the range.
- more generally, $\alpha$-exponential: $K(u) = \sigma^2 e^{-\left(\frac{\|u\|}{\phi}\right)^\alpha}$ for $0 < \alpha \leq 2$.
- spherical

$$K(u) = \begin{cases} 
\sigma^2 \left(1 - \frac{3}{2} \left(\frac{\|u\|}{\phi}\right) + \frac{1}{2} \left(\frac{\|u\|}{\phi}\right)^3\right), & \|u\| \leq \phi, \\
0, & \|u\| > \phi.
\end{cases}$$
More valid covariances

- Rational quadratic: \( K(\mathbf{u}) = \sigma^2 \left( 1 + \frac{\|\mathbf{u}\|^2}{2\alpha\phi^2} \right)^{-\alpha} \)

- Matérn:

\[
K(\mathbf{u}) = \sigma^2 \left\{ 2^{\nu-1} \Gamma(\nu) \right\}^{-1} \left( \frac{\|\mathbf{u}\|}{\phi} \right)^\nu K_\nu \left( \frac{\|\mathbf{u}\|}{\phi} \right).
\]

\( \sigma^2 \) is the **variance**, \( \phi \) is the **range** and \( \nu \) is the **smoothness**. Special cases

1. \( \nu = 1/2 \rightarrow K(\mathbf{u}) = \sigma^2 e^{-\|\mathbf{u}\|/\phi}, \)
2. \( \nu = 3/2 \rightarrow K(\mathbf{u}) = \sigma^2 \left( 1 + \frac{\|\mathbf{u}\|}{\phi} \right) e^{-\|\mathbf{u}\|/\phi}, \)
3. \( \ldots \)
4. \( \nu \rightarrow \infty \) squared exponential.
Some examples

Some examples of covariance functions, all of them are isotropic.
A realization

Figure: Realization of an isotropic model
Non-isotropic models

- A simple generalization: a non-isotropic model, but still stationary: $K(u)$ is not a function of $||u||$.

- **Geometrically anisotropic**: $K(u) = K(||Au||)$, where $A$ is an affine matrix. $A$ is controlled by distortion over the two axes $(\lambda_1, \lambda_2)$ and rotation $\theta$.

- Does not require any new theory: just modify the existing covariance functions

- Applications: Diffusion Tensor Imaging (DTI), water molecules parallel to the fiber tract
A realization

Figure: Realization of a geometrically anisotropic model
In many applications, \( \text{cov}(Y_s, Y_{s'}) \neq K(s - s') \): nonstationary.

DTI: water molecules do not move at the same rate everywhere.

Possibly the simplest construction (Fuentes, 2001)

\[
Y_s = \sum_{i=1}^{n} w_i(s) \tilde{Y}_s^i
\]

- \( w_i \): weight function, decaying from a centroid \( c_i \).
- \( \tilde{Y}_s^i \): iid isotropic/anisotropic random fields.
Figure: From an isotropic model to a locally anisotropic model
Figure: From an isotropic model to a locally isotropic model
A generalization (Fuentes and Smith, 2002): convolution

\[ Y_s = \int_{\mathbb{R}^2} w(s - u) \tilde{Y}_s^{\theta(u)} du \]

- \( w \): kernel function, \( \theta(u) \) slowly varying function.
- \( \tilde{Y}_s^{\theta(u)} \) iid isotropic/anisotropic random fields.
• $Y_s$ is not isotropic, however we can assume $Y_{f(s)}$ (Sampson and Guttorp, 1992).

• Example: gravitational lensing in the cosmic microwave background radiation.

• $f(s) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a smooth invertible transformation

• So

\[
\text{cov}(Y_s, Y_{s'}) = K(\| f(s) - f(s') \|)
\]

• In practice two problems:
  1. estimate $f$ (from the physics)
  2. estimate $K$ (with statistics)
Figure: From an isotropic model to a deformed model
What about interpolation?

- Why do you need a spatial model?
- Typical problem in spatial statistics: **interpolate** on different locations. We want to know the distribution of $Y_s$ given \{ $Y_{s_1}, Y_{s_2}, \ldots, Y_{s_n}$ \}. **Kriging.**
- For fMRI data, interpolation is not the primary concern: we are measuring all the voxels at the same time.

$$Y = X\beta + \varepsilon$$

environmental statistics is focused on interpolation, neuroimaging on inference on $\beta$.

... so why should we care?
Simulation study

Simulation for every ROI

- **empirical covariance** for $\varepsilon(t)$
- assume a **constant mean** across ROI
- **inactive mean**
- 100 simulations, test for activation for different models
Table: False Positives (nominal 5%) for three regions and the mean across ROI.

<table>
<thead>
<tr>
<th>ROI</th>
<th>ind</th>
<th>iso</th>
<th>aniso</th>
<th>l-aniso</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOG, Right</td>
<td>80</td>
<td>26</td>
<td>28</td>
<td>13</td>
</tr>
<tr>
<td>PHG, Right</td>
<td>72</td>
<td>14</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>ORBsup, Left</td>
<td>64</td>
<td>35</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>mean</td>
<td>78.7</td>
<td>31.9</td>
<td>28.3</td>
<td>26.3</td>
</tr>
</tbody>
</table>
Power curve

Figure: Power curve.
Now we have a spatial model, which implies a covariance matrix $K(\theta)$.

We need to do inference!

Likelihood

$$L(\theta \mid Y) = (2\pi)^{-n/2} |K(\theta)|^{-1/2} \exp \left\{ -\frac{1}{2} (Y - X\beta)^\top K(\theta)^{-1} (Y - X\beta) \right\}.$$
The likelihood and log likelihood functions

Loglikelihood:

$$
\ell(\theta \mid Y) = \ln L(\theta \mid Y) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |K(\theta)| - (Y - X\beta)^\top K(\theta)^{-1} (Y - X\beta).
$$

The real trouble is $K(\theta)$. Storing it requires $O(n^2)$ space and $O(n^3)$ flops for computing the log determinant and matrix inversion.
You need to store the matrix first. If you have a 150,000 voxels, the total matrix size would be

\[(8 \times 150,000)^2 / (1024)^4 \approx 1.3 \text{Tb}.\]

We can exploit the structure of \( K(\theta) \) (symmetric and positive definite) to improve the storage, but

- This will not solve the problem (650 Gb)
- Most programming languages do not allow (yet) linear algebra operations with symmetric storage.

However... voxels are on a regular grid. Let us assume our model is stationary.
Whittle approximation, the basic idea
Every row consists of a circular shift of 1 element of the previous row. This type of matrix is called **circulant**, and it has some very convenient properties.
In matrix form, $\tilde{K}(\theta) = D^\top \Lambda(\theta) D$. $D$ is the DFT matrix, $\Lambda(\theta)$ is a diagonal matrix with the Fourier coefficients.

$$
\ell(\theta \mid Y) = \ln L(\theta \mid Y)
\approx -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{j=0}^{n-1} \ln \lambda_j(\theta)
- \sum_{j=0}^{n-1} \text{FFT}(Y - X\beta)_j^2 / \lambda_j(\theta).
$$

This is called the **Whittle approximation**.

You don’t need to store any large matrix: just compute the FFT of the data.

Storage $O(n^2) \to O(n)$, flops $O(n^3) \to O(n \log n)$. 
An example of application

One **healthy control** in a clinical study with 15 stroke patients and 12 healthy control subjects.

Each session has 48 consecutive scans (every 2 seconds), task: hand grasping

![Time course](image.png)

Three sessions, $T = 144$ time points. $\approx 150,000$ voxels in a 3D space (2mm×2mm×2mm) for each time.

Total of **22 million** data points.
Response:

Each voxel \( v \) belongs to a Region of Interest (ROI) \( r \).

The fMRI intensity at \( v \) is \( Y_{v;r}(t) \).

\[
Y(t) = \{ Y_{v_1;r_1}(t), \ldots, Y_{v_V;r_V}(t) \}, \quad Y = \{ Y(1), \ldots, Y(T) \}.
\]

Covariates:

\( l_1(t) \): indicator for task. \( l_2(t) = 1 - l_1(t) \): indicator for rest.
Hemodynamic response function \( h(t) \) known and common across voxels

**Figure:** fMRI intensity (blue), below \( l_1(t) \) and \( X_1(t) = (h \ast l_1)(t) \) (red)
The model is:

\[ Y = X\beta + \varepsilon, \]

Mean structure (≠ for every voxel):
- contribution of \( X_1(t) \) and \( X_2(t) \) (\( \beta_1 \) and \( \beta_2 \))
- intercept
- session mean
- time effect
Neuroscience questions

Is a voxel active?: is $\beta_1 - \beta_2 \neq 0$ for some voxel $v$? Activation

This is an hypothesis test, for each voxel.

Standard approach: independent voxels $\rightarrow$ linear regression (general linear model).

Some more advanced: False Discovery Rate to bound false positives under (positive) dependence.

Here: model spatial dependence, to obtain more accurate activation patterns.
A few notes

\[ Y = X\beta + \varepsilon \]

In my model, \( \beta \) is fixed, all spatial information is in \( \varepsilon \).

In a classical **Bayesian model** \( \beta \) is a **Gaussian Markov Random Field**, and \( \varepsilon \) simple noise. All spatial information is in \( \beta \).

Combinations are possible: \( \beta_s \) and a model for \( \varepsilon \).
Figure: The three-step spatial model.
Temporal dependence and multiple scales

Vector AR(2) in time:

\[ \varepsilon(t) = \Phi_1 \varepsilon(t-1) + \Phi_2 \varepsilon(t-2) + S \{ \Omega H_1(t) + (I_V - \Omega)H_2(t) \}. \]

- \( \Phi_i = \{ \phi_{i;v} \} \) \( i = 1, 2 \) diagonal
- \( S = \{ \sigma_v \} \) diagonal
- \( \Omega H_1(t) + (I_V - \Omega)H_2(t) \) unscaled innovations
- \( \Omega \): relative contribution of \( H_1 \) w.r.t. \( H_2 \)
Figure: Fit for four random voxels.
The spatial model

Figure: Scheme of the model of the spatial part of the model
Modeling local dependence

\( \mathbf{H}_1(v; r) \) independent for every ROI.

Sum of geometrically anisotropic processes

\[
\mathbf{H}_1(v; r) = \sum_{i=1}^{k} \mathbf{H}_1^i(v; r) w_i(v),
\]

where \( \mathbf{H}_1^i(v; r) \) are iid Gaussian processes with

\[
\text{cov}\{\mathbf{H}_1^i(v; r), \mathbf{H}_1^i(v'; r)\} = \text{Matérn with anisotropic distance (in 3D!)}.\]
Figure: BIC for all 90 ROIs for the isotropic model and the locally anisotropic model.
Figure: Activation at 0.01% for independent voxels (top) and the model (bottom).
### Results

**Table:** Activated voxels (in %).

<table>
<thead>
<tr>
<th>ROI</th>
<th>Active</th>
<th>Inhibited</th>
<th>Total</th>
<th><em>ind</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Supp Motor Area L</td>
<td>53</td>
<td>14</td>
<td>67</td>
<td>29</td>
</tr>
<tr>
<td>Supp Motor Area R</td>
<td>35</td>
<td>11</td>
<td>46</td>
<td>28</td>
</tr>
<tr>
<td><strong>mean</strong></td>
<td>25</td>
<td>15</td>
<td>40</td>
<td>23</td>
</tr>
</tbody>
</table>
Spatial statistics can help assessing activation, but more work is needed in defining appropriate models for fMRI.

Inference is hard (≈100,000 voxels), but not impossible: we are on a grid and there is a large literature in environmental statistics.

A statistical analysis for a large data set requires a large computer.

Where to inject spatial information? $\beta$ vs $\epsilon$. ‘Someone’s mean function is someone else’s covariance function’