

Spatial Models: A *Quick* Overview

Instituto Tecnológico Autónomo de México

Murali Haran

Department of Statistics
Penn State University

November 29, 2013

What this tutorial will cover

- I will explain why spatial models may be useful to scientists in many disciplines.
- I will outline types of spatial data and some basic concepts.
- The idea is to give you enough information so you know when you might have a spatial data problem and where you could look to find help.

What are spatial data?

- Data that have locations associated with them.
- Assumption: their locations are important in how we interpret and analyze the data. Locations may also be central to the scientific questions of interest.

Some reasons to use spatial models

- Fitting an inappropriate model for the data, say by ignoring dependence, may lead to incorrect conclusions. e.g. underestimated variances
- Can lead to superior estimators (e.g. low mean squared error).
- Sometimes learning about spatial dependence is central to the scientific interprise. e.g. when finding spatial clusters, regions of influence/dependence.

Types of Spatial Data

There are three main categories of spatial data (though it is not always obvious how to classify data into these categories):

- Spatial point processes: When a spatial process is observed at points and the locations themselves are of interest. Typical research questions are: Is the pattern random or does it exhibit clustering?
- Geostatistical data: When a spatial process that varies continuously is observed only at a few points.
- Lattice data: When a spatial process is observed at countably many (often finitely many) locations. Usually this arises due to aggregation of some sort, e.g. averages over a pixel in an image

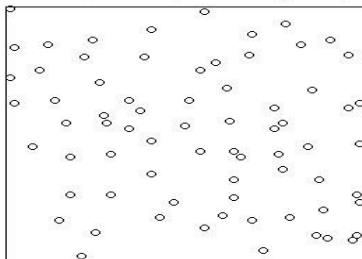
Spatial Point Process Data: Example

Locations of pine saplings in a Swedish forest.

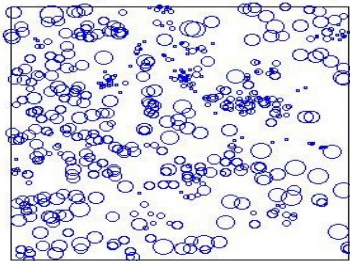
Location and diameter of Longleaf pines (marked point process).

Are they randomly scattered or are they clustered?

Point pattern (Swedish pines)



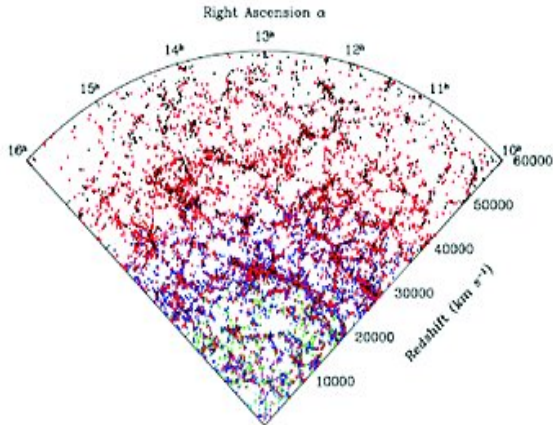
Marked point pattern (Longleafs)



(from Baddeley and Turner \mathcal{R} package, 2006)

The galaxy distribution: 3D spatial point process

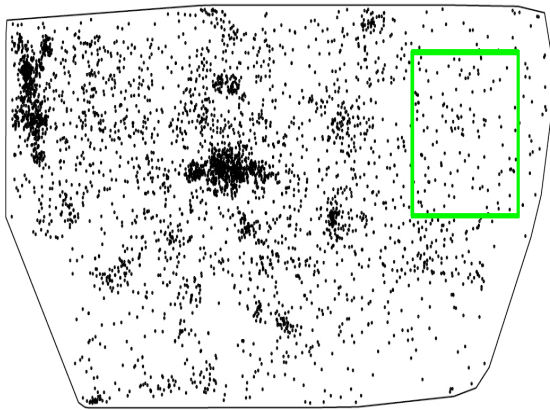
2d location in sky, 1d from redshift as a surrogate for distance.



(Tegmark et al. 2004, The three-dimensional power spectrum of galaxies from the Sloan Digital Sky Survey,

Shapley concentration: 2D spatial point process

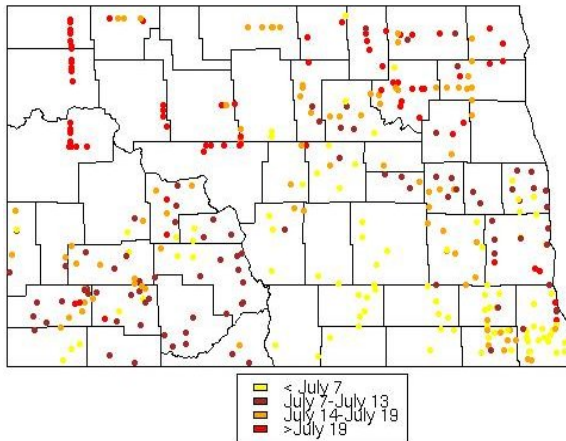
Nearby rich supercluster of galaxies. Several thousand galaxy redshift measurements. Galaxies show statistically significant clustering on small scales (Baddeley, 2008)



Geostatistical (point-referenced) data: Example

Spatial analogue to continuous-time time series data.

Wheat flowering dates by location (below):



Lattice Data: Example

Spatial analogue to discrete-time data, e.g. images



<http://www.scantips.com>

Spatial Point Processes: Introduction

- **Spatial point process:** The *locations* where the process is observed are random variables, process itself may not be defined; if defined, it is a **marked spatial point process**.
- **Observation window:** the area where points of the pattern can possibly be observed. The observation window specification is vitally important since absence of points in a region where they could potentially occur is also valuable information whereas absence of points outside of an observation window does not tell us anything.

Questions related to spatial randomness of process

Some examples:

- Is there regular spacing between locations where process was observed or do locations show a tendency to cluster together? Need to fit clustering models and perhaps do some hypothesis testing.
- Does the probability of observing the event vary according to some factors? (Need to relate predictors to observations in a regression type setting.)
- Can we estimate the overall count from only partial observations? Need to fit a model to observations and make estimates/predictions based on fitted model.

Questions related to spatial randomness (contd)

Assume multiple (sometimes competing) models for the process. For instance, when studying point patterns of observations:

- Perhaps non-homogenous environmental conditions (associated with locations) are related to the presence/absence.
- Maybe the pattern arose by virtue of how the process spreads (e.g. clustering of 'offspring' near 'parents')?
- Note that **hypothesis testing alone is inadequate** for most of these questions. Can try to resolve these by fitting appropriate models where intensity of the process is modeled according to one of the models above.

Spatial Point Processes: Notes

- There appear to be many important problems where spatial point process modeling may be the most appropriate approach.
- However, the complexity of the theory along with computational difficulties have made it much less ‘friendly’ to applications than geostatistical models or lattice models.
- Recent methodological developments and software such as the R library `spatstat` (A.Baddeley and Turner) are slowly opening up greater possibilities for practical modeling and analyses.

Some definitions for spatial point processes

- A spatial point process is a stochastic process, a realization of which consists of a countable set of points $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ in a bounded region $S \in \mathbb{R}^2$
- The points \mathbf{s}_i are called **events**.
- For a region $A \in S$, $N(A) = \#(\mathbf{s}_i \in A)$.
- The **intensity measure** $\Lambda(A) = E(N(A))$ for any $A \in S$.
- If measure $\Lambda(A)$ has a density with respect to Lebesgue measure (we will typically assume this holds), then it can be written as:

$$\Lambda(A) = \int_A \lambda(\mathbf{s}) d\mathbf{s} \quad \text{for all } A \in S.$$

$\lambda(\mathbf{s})$ is called the **intensity function**.

Spatial point process modeling

Spatial point process models can be specified by :

- A deterministic intensity function (analogous to generalized linear model framework)
- A random intensity function (analogous to random effects models)
- Two classes of models:
 - Poisson Processes \approx provide models for no interaction patterns.
 - Cox processes \approx provide models for aggregated point patterns.
- Poisson process: Fundamental point process model — basis for exploratory tools and constructing more advanced point process models.

Homogeneous Poisson Process

Poisson process on \mathbf{X} defined on S with intensity measure Λ and intensity function λ , satisfies for any bounded region $B \in S$ with $\Lambda(B) > 0$:

- ❶ $N(B) \sim \text{Poisson}(\Lambda(B))$.
- ❷ Conditional on $N(B)$, the points (event locations) $\mathbf{X}_B = \{X_1, \dots, X_{N(B)}\}$ in the bounded region are (i.i.d.) and each uniformly distributed in the region B .
- **Homogeneous Poisson process:** The intensity function, $\lambda(\mathbf{s})$ is constant for all $\mathbf{s} \in S$.
- Poisson process is a model for complete spatial randomness since \mathbf{X}_A and \mathbf{X}_B are independent for all $A, B \in S$ that are disjoint.

Poisson Process (contd.)

- The intensity $\lambda(\mathbf{s})$ specifies the mean number of events per unit area as a function of location \mathbf{s} .
- Intensity is sometimes called the 'density' in other fields such as ecology (this term would be confused with a probability density, which is why it is not used in statistics).
- It is important as a null model and as a simple model from which to build other models.
- Homogeneous Poisson process is model for complete spatial randomness against which spatial point patterns are compared.

Poisson Process (contd)

Some notes:

- Independence properties of Poisson process makes it unrealistic for most applications. However, it is mathematically tractable and hence easy to use/study. Almost always a good starting point.
- For modeling, usually consider log model of intensity function (to preserve non-negativity of intensity):
$$\log \lambda(\mathbf{s}) = \mathbf{z}(\mathbf{s})\boldsymbol{\beta}^T$$

Intensity of Poisson point process

Let $d\mathbf{s}$ denote a small region containing location \mathbf{s} .

- First-order intensity function of a spatial point process:

$$\lambda(\mathbf{s}) = \lim_{d\mathbf{s} \rightarrow 0} \frac{E(N(d\mathbf{s}))}{|d\mathbf{s}|}.$$

- Second-order intensity function of a spatial point process:

$$\lambda^{(2)}(\mathbf{s}_1, \mathbf{s}_2) = \lim_{d\mathbf{s}_1 \rightarrow 0} \lim_{d\mathbf{s}_2 \rightarrow 0} \frac{E\{N(d\mathbf{s}_1)N(d\mathbf{s}_2)\}}{|d\mathbf{s}_1||d\mathbf{s}_2|}.$$

- Covariance density of a spatial point process

$$\gamma(\mathbf{s}_1, \mathbf{s}_2) = \lambda^{(2)}(\mathbf{s}_1, \mathbf{s}_2) - \lambda(\mathbf{s}_1)\lambda(\mathbf{s}_2).$$

Intensity of Poisson point process (contd.)

Assume process is stationarity (same intensity everywhere) and isotropic (direction/rotations do not affect the process)

$$K(d) = \frac{1}{\lambda} E(\text{number of events within distance } d \text{ of an arbitrary event}).$$

- If process is clustered: Each event is likely to be surrounded by more events from the same cluster. $K(d)$ will therefore be *relatively large* for small values of d .
- If process is randomly distributed in space: Each event is likely to be surrounded by empty space. For small values of d , $K(d)$ will be *relatively small*.

Can obtain an intuitive estimator for $K(d)$ for a given data set.

Ripley's K Function

Let λ be the intensity of the process.

- Effective method for seeing whether the process is completely random in space.

$$K(d) = \frac{\text{Mean number of events within distance } d \text{ of an event}}{\lambda}$$

- This can be estimated by

$$\hat{K}(d) = \frac{\sum_{i \neq j} w_{ij} I(d_{ij} \leq d)}{\hat{\lambda}}$$

where $\hat{\lambda} = N/|A|$ with $|A|$ as the total area of the observation window and N is the observed count.

- Note: K can also be viewed as an integral of the two point correlation function as used by astronomers (cf. Martinez and Saar, 2002).

Inhomogeneous Poisson processes

Useful for modeling spatial process that varies in intensity over space. An inhomogeneous Poisson process with intensity λ satisfies:

- Number of events $N(A)$ in an observation window A is Poisson with mean

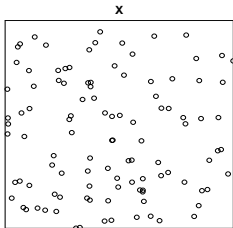
$$\Lambda(A) = \int_A \lambda(\mathbf{s}) d\mathbf{s},$$

equivalently, $P(N(A) = N) = \frac{1}{N!} e^{-\Lambda(A)} (\Lambda(A))^N$.

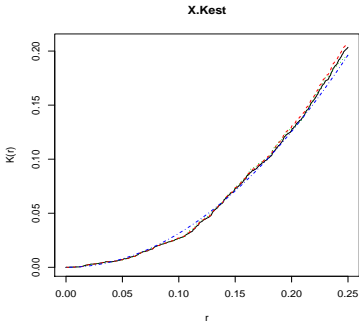
- Conditional on $N(A)$, event locations are independently sampled from a probability density function proportional to $\lambda(\mathbf{s})$.

Ripley's K for homogeneous Poisson Process

Process was simulated with intensity function $\lambda(x, y) = 100$.
homogeneous Poisson Process



Ripley's K



blue= K function under complete spatial randomness

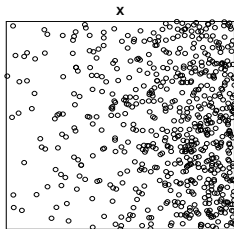
black (and red and green) are various versions of estimates of the K function

Ripley's K for inhomogeneous Poisson Process (Eg.1)

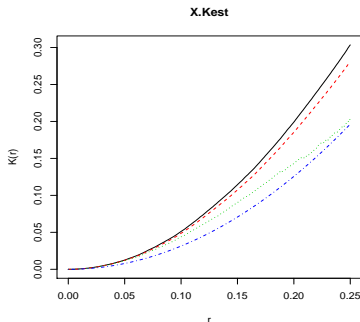
Process was simulated with intensity function

$$\lambda(x, y) = 100 \exp(3x).$$

Inhomogeneous Poisson Process



Ripley's K



blue=K function under complete spatial randomness

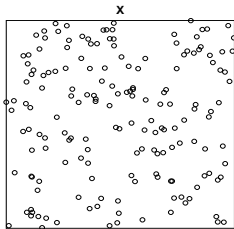
black (and red and green) are various versions of estimates of the K function

Ripley's K for inhomogeneous Poisson Process (Eg.2)

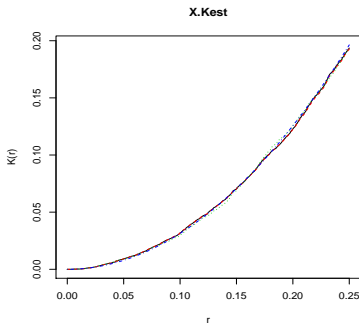
Process was simulated with intensity function

$$\lambda(x, y) = 100 \exp(y).$$

Inhomogeneous Poisson Process



Ripley's K



blue= K function under complete spatial randomness

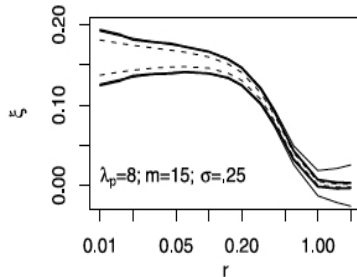
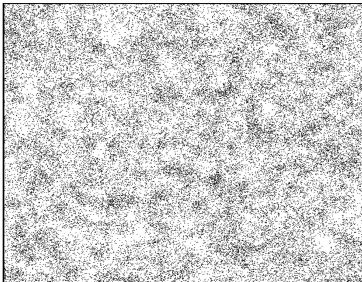
black (and red and green) are various versions of estimates of the K function

Exploring non-homogeneity

- A common way to study spatial point processes is to compare the realization of the process (observations) to a homogeneous Poisson process. This kind of exploratory data analysis or hypothesis test-based approach can be a useful first step.
- Estimating errors on the 2-point correlation or K function has been derived for a random Poisson process (Ripley 1988; Landy & Szalay 1993) or, if a model for the underlying process is known, from a parametric bootstrap (Eisenstein et al. 2005).
- But Poisson errors may be too small for spatially correlated samples.

Exploring non-homogeneity: recent developments

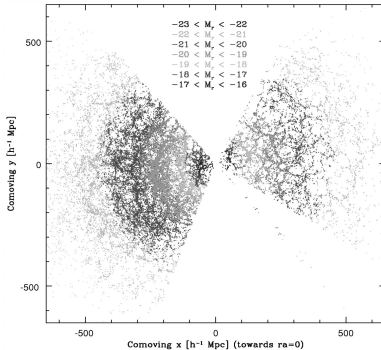
Loh (2008) recommends a ‘marked point bootstrap’ resampling procedure. Figures below show a simulated clustered process, and the resulting 2-point correlation function with Poisson (dashed) and bootstrap (solid) 95% confidence error bands respectively.



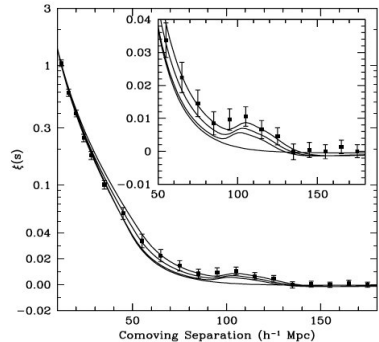
Loh 2008, A valid and fast spatial bootstrap for correlation functions, *Astrophys. J.* 681, 726-734

Example: Galaxy clustering (Sloan Digital Sky Survey)

Galaxy distribution



Two-point correlation function



Distribution of 67,676 galaxies in two slices of the sky showing strong anisotropic clustering (Tegmark et al. 2004).

Bottom: Two-point correlation function showing the faint feature around 100 megaparsec scales revealing cosmological

Baryonic Acoustic Oscillations (Eisenstein et al. 2005).

Exploratory data analysis and inference

- So far we have been discussing exploratory data analysis techniques.
- These can be enormously useful for analyses. Fitting a model to data can, however, provide a richer approach for analyses, particularly when there are competing data generating mechanisms
- Recently developed algorithms and software (`spatstat` in R) make it easier to fit at least relatively simple or standard point process models. More advanced models need more sophisticated techniques like Markov chain Monte Carlo.

Cox Process

The Cox process or the *doubly stochastic Poisson process* (Cox, 1955) is a more flexible and realistic class of models than the Poisson process model.

- Natural extension of a Poisson process: Consider the intensity function of the Poisson process as a realization of a random field. We assume $\Lambda(A) = \int_A \lambda(\mathbf{s}) d\mathbf{s}$.
 - Stage 1: $N(A) | \Lambda \sim \text{Poisson}(\Lambda(A))$.
 - Stage 2: $\lambda(\mathbf{s}) | \Theta \sim f(\cdot; \Theta)$ so that λ is stochastic, a nonnegative random field parametrized by Θ .
- Simple case: If $\lambda(\mathbf{s})$ is deterministic, \mathbf{X} is a Poisson process with intensity $\lambda(\mathbf{s})$.

Markov Point Processes

- Point patterns may require a flexible description that allows for the points to interact.
- Markov point processes are models for point processes with interacting points (attractive or repulsive behavior can be modeled).
- ‘Markovian’ in that intensity of an event at some location \mathbf{s} , given the realization of the process in the remainder of the region, depends only on information about events within some distance of \mathbf{s} .
- Origins in statistical physics, used for modeling large interacting particle systems.

Inference for spatial point process models

- Maximum likelihood for all but the simplest spatial point process model is analytically intractable. Maximum pseudolikelihood (MPL) is a useful approximation to maximum likelihood.
- For some models, can use Newton-Raphson or some variant but often need (Markov chain) Monte Carlo maximum likelihood (MCML), also referred to as simulated maximum likelihood (SML).
- No 'automatic' methods exist for fitting such models.
- Simulating from a point process model is often easy but inference (estimating a point process model based on observations) is usually more difficult. Challenging to fit flexible new models.

Spatial point processes: computing

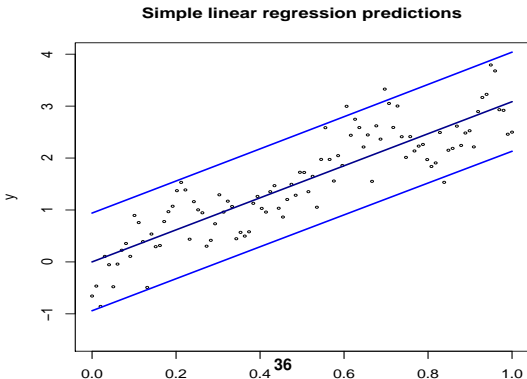
- **R command:** `spatstat` function `ppm` fits models that include spatial trend, interpoint interaction, and dependence on covariates, generally using MPL.
- MPL often works well in practice (Baddeley, 2005). Caveat: MPL can work very poorly in some cases, particularly when there is strong dependence.
- MPLE can be used to get a guess for MLE before doing something more elaborate like Markov chain Maximum Likelihood (cf. C.J.Geyer's chapter in "MCMC in Practice", 1996 for a gentle introduction.)
- There is not much in the way of computing resources for fitting Bayesian models, even though they are becoming increasingly common.

Continuous-domain/geostatistical data

We will now talk briefly about continuous-domain spatial data. This is often useful for interpolation and for adjusting for spatial dependence when running regressions, for instance.

The importance of dependence (contd.)

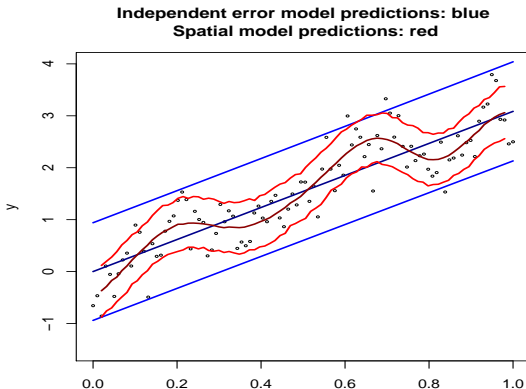
Toy example: simple linear regression with the correct mean but assuming iid error structure. $Z(s_i) = \beta s_i + \epsilon_i$, where ϵ_i s are iid. Does not capture the data/data generating process well even though trend (β) is estimated correctly.



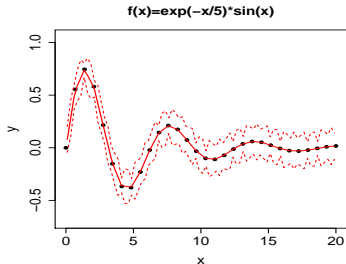
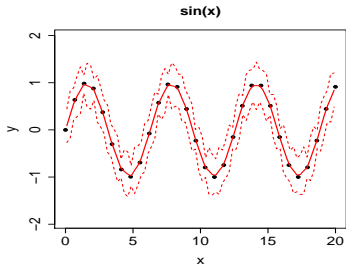
The importance of dependence (contd.)

Model: linear regression with correct mean, now assuming dependent error structure. This picks up the 'wiggles'.

Independent error model: blue. Dependent error model: red.



Fitting complicated mean structures



Functions: $f(x) = \sin(x)$ and $f(x) = \exp(-x/5) \sin(x)$.

Same model used both times: $f(x) = \alpha + \epsilon(x)$, where $\{\epsilon(x), x \in (0, 20)\}$ is a Gaussian process, α is a constant.

Note: the dependence is being introduced to indirectly capture the non-linear structure, not to model dependence per se.

Spatial (linear) model for geostatistics and lattice data

Although geostatistical models and lattice data models are usually talked about separately, they can be viewed in a unified framework.

- Spatial process at location \mathbf{s} is $Z(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s})$ where:
 - $\mu(\mathbf{s})$ is the mean. Often $\mu(\mathbf{s}) = X(\mathbf{s})\beta$, $X(\mathbf{s})$ are covariates at \mathbf{s} and β is a vector of coefficients.
- Model dependence among spatial random variables by imposing it on the errors (the $w(\mathbf{s})$'s).
- For n locations, $\mathbf{s}_1, \dots, \mathbf{s}_n$, $\mathbf{w} = (w(\mathbf{s}_1), \dots, w(\mathbf{s}_n))^T$ can be jointly modeled via a zero mean Gaussian process (GP), for geostatistics, or Gaussian Markov random field (GMRF), for lattice data.

Gaussian Processes

- Gaussian Process (GP): Let Θ be the parameters for covariance matrix $\Sigma(\Theta)$. Then:

$$\mathbf{w}|\Theta \sim N(0, \Sigma(\Theta)).$$

This implies:

$$\mathbf{Z}|\Theta, \beta \sim N(\mathbf{X}\beta, \Sigma(\Theta))$$

- We have used the simplest multivariate distribution (the multivariate normal). We will specify $\Sigma(\Theta)$ so it reflects spatial dependence.
- Need to ensure that $\Sigma(\Theta)$ is positive definite for this distribution to be valid, so we assume some valid parametric forms for specifying the covariance.

Gaussian Processes: Example

- Consider the popular **exponential** covariance function.
- Let $\Sigma(\Theta) = \kappa I + \psi H(\phi)$ where I is the $N \times N$ identity matrix. Note that $\Theta = (\kappa, \psi, \phi)$ and $\kappa, \psi, \phi > 0$.
- The i, j th element of the matrix H ,
$$H(\|\mathbf{s}_i - \mathbf{s}_j\|; \phi)_{ij} = \exp(-\phi \|\mathbf{s}_i - \mathbf{s}_j\|).$$
- Note: covariance between i, j th random variables depends only on distance between \mathbf{s}_i and \mathbf{s}_j , and does not depend on the locations themselves (implying *stationarity*) and only depends on the magnitude of the distance, not on direction (implying *isotropy*).
- Extremely flexible models, relaxing these conditions, can be easily obtained though fitting them can be more difficult.

Gaussian Processes: Inference

- The model completely specifies the likelihood, $\mathcal{L}(\mathbf{Z}|\Theta, \beta)$.
- This means we can do likelihood-based inference:
 - If we observe \mathbf{Z} , can find maximum likelihood estimates of Θ, β by maximizing $\mathcal{L}(\mathbf{Z}; \Theta, \beta)$ with respect to Θ, β .
 - Using the MLEs of Θ, β , and conditioning on the observed values \mathbf{Z} , we can easily estimate the value of this process at other locations ('kriging' with Gaussian processes.)
- If we place priors on Θ, β , we can do Bayesian inference:
 - Simulate from the posterior distribution, $\pi(\Theta, \beta | \mathbf{Z})$ via Markov chain Monte Carlo (tutorial tomorrow!)
 - Using sampled values of Θ, β , conditioning on \mathbf{Z} , can easily simulate value of this process at other locations.
 - Bayesian version incorporates variability due to uncertainty about Θ, β .

Gaussian Processes: Computing

- For likelihood based inference: R's `geoR` package by Ribeiro and Diggle.
- For Bayesian inference:
 - R's `spBayes` package by Finley, Banerjee and Carlin.
 - WINBUGS software by Spiegelhalter, Thomas and Best.
- Very flexible packages: can fit many versions of the linear Gaussian spatial model. Also reasonably well documented.
- Warning: With large datasets (>1000 data points), matrix operations (of order $O(N^3)$) become very slow. Either need to be clever with coding or modeling. Above software will not work.

Useful ideas for non-spatial data

Some spatial modeling techniques may be useful in non-spatial scenarios:

- Gaussian processes: Useful for modeling complex relationships of various kinds. Examples: flexible nonparametric regression, classification.
- Fast approximations for complex computer models.
- Ideas for modeling time series, particularly multivariate time series.

Summary: spatial data types and associated models

General spatial process: $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, D is set of locations.

- **Geostatistics:** D is a fixed subset of \mathbb{R}^2 (or \mathbb{R}^3 in 3D case).

$Z(\mathbf{s})$ is a random variable at each location $\mathbf{s} \in D$.

Usual (basic) model: [Gaussian process](#).

- **Lattice data:** $D = \{\mathbf{s}_1, \dots, \mathbf{s}_N\}$ is a fixed regular or irregular lattice, on \mathbb{R}^2 (or \mathbb{R}^3).

$Z(\mathbf{s})$ is a random variable at each location $\mathbf{s} \in D$.

Usual (basic) model: [Gaussian Markov random field](#).

- **Spatial point process:** $D = \{\mathbf{s}_1, \dots, \mathbf{s}_N\}$ is a random collection of points on the plane.

Ordinary point process: $Z(\mathbf{s})$ does not exist. For marked point process, $Z(\mathbf{s})$ is a random variable as well. Usual (basic) models: [Poisson process](#), [Cox process](#).

References: Geostatistics and Lattice Processes

Geostatistics and Lattice Data:

- Schabenberger and Gotway (2005) "Statistical Methods for Spatial Data Analysis". A fairly comprehensive easy to read book on spatial models for (in order of emphasis): geostatistics, lattice data and point processes.
- Cressie (1994) "Statistics for Spatial Data". This is a comprehensive guide to classical spatial statistics, but it is considerably more technical than the other two references listed here.
- S. Banerjee, B.P. Carlin and A.E. Gelfand (2004) "Hierarchical Modeling and Analysis for Spatial Data". This is a textbook on Bayesian models for spatial data.

References: Spatial Point Processes

Spatial Point Processes:

- Møller and Waagepetersen review article “Modern spatial point processes modelling and inference (with discussion)” (Scandinavian Journal of Statistics, 2007) or online chapter: <http://people.math.aau.dk/~jm/spatialhandbook.pdf>
- Baddeley and Turner’s R `spatstat` package.
- Baddeley et al. “Case Studies in Spatial Point Process Modeling” (2005).
- P.J.Diggle’s online lecture notes:

<http://www.maths.lancs.ac.uk/~diggle/spatialepi/notes.ps>

References: Spatial Point Processes

- P.J.Diggle “Statistical Analysis of Spatial Point Patterns” (2003)
- “Modern statistics for spatial point processes” by J.Møller and R.P.Waagepetersen (2004).
- V.J.Martinez and E.Sarr “Statistics of the Galaxy Distribution.”

Notes about the references:

- ❶ Several of these references also cover spatiotemporal (space-time) process, that may also be of significant interest.
- ❷ Acknowledgement: A lot of the material and examples in this tutorial were drawn from several of the listed references.