# Statistics 260: Spatial Statistics

Cari Kaufman cgk@stat.berkeley.edu



What are spatial data?

Three types:

- *Point-referenced or geostatistical*
  $Y(s), s \in \Re^d$ and $s$ varies continuously
- *Areal or lattice or discrete*
  Finite number of areal units, e.g. counties or elements of a grid
  Observations are typically sums or averages
- *Point patterns or point processes*
  Locations themselves are random

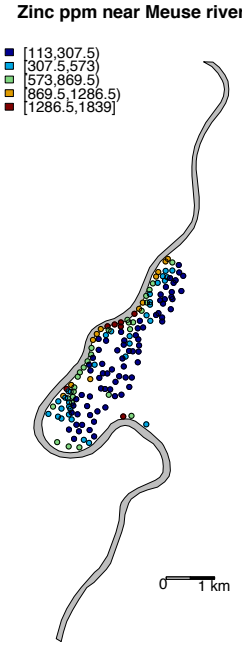


Example: point-referenced

How are heavy metal concentrations related to other spatial variables, such as elevation and distance to the river? (estimation)

What are the concentrations at unmeasured locations? (prediction)

Can we model the processes that govern the concentrations?






Possible model for the covariance:

$$Cov(Y(s_i), Y(s_j)) = \begin{cases} \sigma^2 e^{-||s_i - s_j||/\rho} & i \neq j \\ \sigma^2 + \tau^2 & i = j \end{cases}$$

This is an *isotropic* covariance function, meaning it is only a function of distance between locations.

We will examine models of the form

$$Y \sim MVN(\mu, \Sigma(\theta))$$

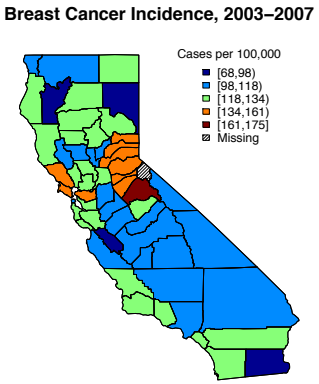
and generalize to *Gaussian process* models.

Example: areal data

Is there a spatial pattern? (testing)

What is the local risk for disease? (possible smoothing)

Are changes in risk related to other variables, such as exposure to contaminants?






We often specify spatial dependence for such data

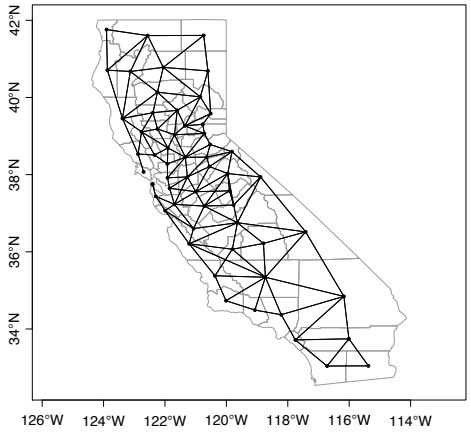
- Conditionally, for example with a conditional auto-regressive (CAR) model $Y_i \overset{indep}{\sim} N(\phi_i, \sigma^2)$ and

$$\phi_i|\phi_{-i} \sim N(\mu + \sum_{j=1}^{n} a_{ij}(\phi_j - \mu), \tau_i^2)$$

- Using a graphical model structure to impose conditional independence constraints, such as $x_i \perp x_j | x_{-ij}$ if $x_i, x_j$ do not share an edge (are not neighbors).

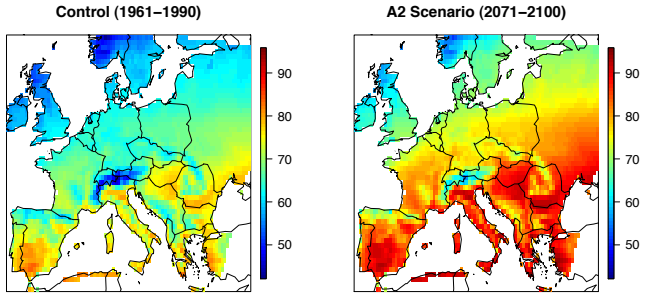


Graph structure for CA counties based on shared borders






Average summer temperatures from the Hadley Centre (UK) regional climate model predictions



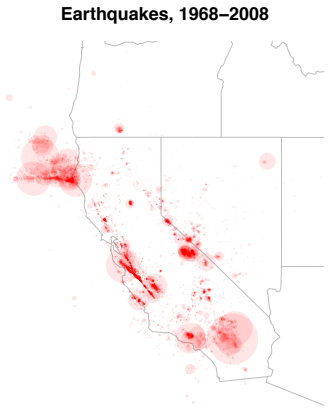

Should we use a geostatistical or areal data model?

Example: point process data

This is an example of a *marked point process*. Each earthquake location is associated with a magnitude.

Is there spatial randomness? (testing)

Can we distinguish clustering from a non-constant rate function?






A *homogeneous Poisson process* defines the number of points in any region $A$ to be Poisson with

$$E[\text{points in A}] = \lambda \text{area}(A)$$

This model is also known as *complete spatial randomness (CSR)*. Deviations from it can be detected using Ripley's K function:

$$K(d) = \frac{1}{\lambda} E[\text{points within distance d of an aribtrary point}]$$

Under CSR, we have $K(d) = \pi d^2$. Clustering corresponds to $K(d) > \pi d^2$ and spatial repulsion to $K(d) < \pi d^2$. Now we need to estimate K.



# Geostatistical Models



A *spatial stochastic process* is a spatially indexed collection of random variables

$$\{Y(s) : s \in D \subset \Re^d\}$$

For us, $d$ will typically be 2 or 3.

A realization from a stochastic process is sometimes referred to as a *sample path*.

Typically we observe a vector

$$Y \equiv (Y(s_1), \ldots, Y(s_n))^T$$

Under some consistency criteria, we can define a stochastic process using its *finite-dimensional distributions*

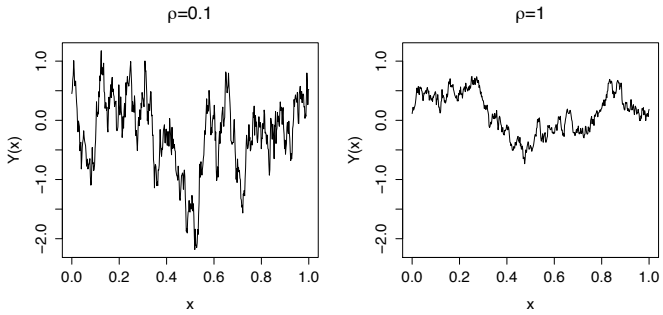
$$F(y_1, \ldots, y_n; s_1, \ldots, s_n) = P(Y(s_1) \leq y_1, \ldots, Y(s_n) \leq y_n)$$

A *Gaussian process (GP)* has finite-dimensional distributions that are all multivariate normal distributions. We define a GP using its *mean and covariance functions*

$$\begin{aligned} \mu(s) &= E[Y(s)] \\ K(s_1, s_2) &= Cov(Y(s_1), Y(s_2)) \end{aligned}$$



Example: Returning to the covariance function $K(s_1, s_2) = \sigma^2 \exp\{||s_i - s_j||/\rho\}$ (taking no nugget effect this time), we can simulate a "realization" from a mean zero GP by first constructing the covariance matrix on a fine grid, then sampling a multivariate normal random vector.






Returning to stochastic processes in general, some useful properties are

- *Strict stationarity*
  $F(y_1, \ldots, y_n; s_1 + h, \ldots, s_n + h) = F(y_1, \ldots, y_n; s_1, \ldots, s_n)$
- *Second-order stationarity*
  $E[Y(s)] = E[Y(s + h)] = \mu$
  $Cov(Y(s), Y(s + h)) = Cov(Y(0), Y(h)) = C(h)$

  Note that for a GP, second-order stationarity and strict stationarity are equivalent.



- *Isotropy*
  $C(h)$ depends only on $||h||$ ; we write $C(h) = \varphi(||h||)$

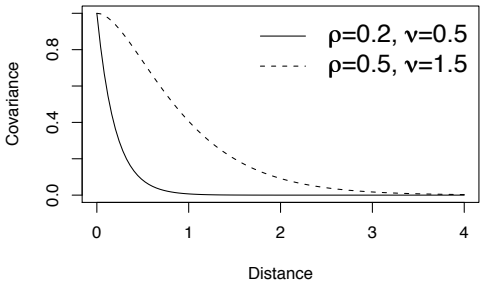
Example: the Matern class of covariance functions

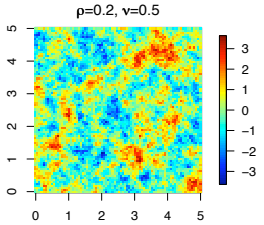
$$\varphi(t) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(t/\rho)^\nu \mathcal{K}_\nu(d/\rho)$$

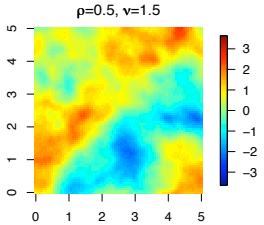
The function $\mathcal{K}_\nu$ is a modified Bessel function of order $\nu$. $\nu$ is known as the *smoothness parameter*. In a sense we will make precise later, $\nu$ controls the differentiability of a GP with this covariance.

The exponential covariance is a special case with $\nu = 1/2$.

$$\varphi(t) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(t/\rho)^\nu \mathcal{K}_\nu(d/\rho)$$






Another form of stationarity, even weaker than second-order stationarity, is

- *Intrinsic stationarity*
  $Var[Y(s+h) - Y(s)]$ depends only on $h$

When this is true, we call the function

$$\gamma(h) = \frac{1}{2}Var[Y(s+h) - Y(s)]$$

the *semivariogram* (and $2\gamma(h)$ the *variogram*).



Second-order stationarity implies intrinsic stationarity, and in this case $\gamma(h) = C(0) - C(h)$

However, the converse is not true. For example, 1D Brownian motion is intrinsically stationary but not second-order stationary.

For this reason, the variogram is a more general way of describing second-order structure of the process, and it is often estimated as a diagnostic tool.

Exercise: calculate the variogram for the exponential model on page 4 and plot it.



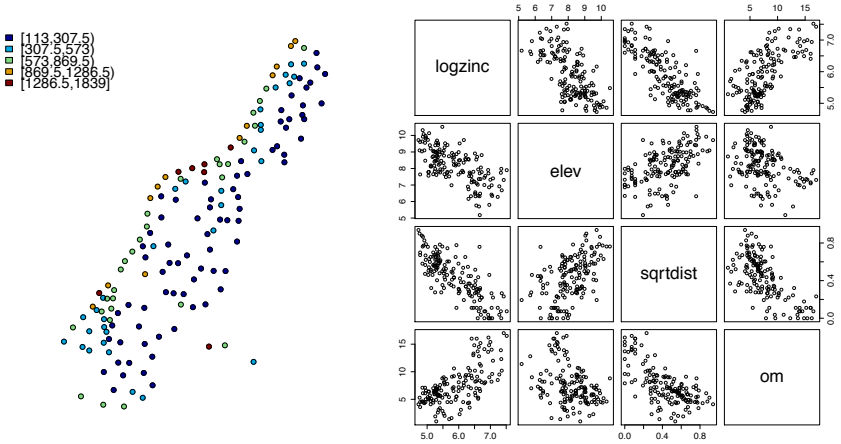
In "classical geostatistics" we observe

$$Y \equiv (Y(s_1), \ldots, Y(s_n))^T$$

and then

- Make inference about the data generating process, i.e. specify a model and estimate its parameters
- Predict $Y(s_0)$ at a new location $s_0$ that we did not observe (or multiple locations). Other terms for this are interpolation or kriging.

As an example, I'll consider the zinc concentration data, with n = 155.






To begin specifying a model, we may assume

$E[Y(s)]$ (nonrandom)

zero mean stationary process

$$\begin{aligned} Y(s) &= \mu(s) + e(s) \\ &= \mu(s) + \eta(s) + \epsilon(s) \end{aligned}$$

spatially correlated process

white noise process (measurement error)

Now we will estimate the components.



The most common parametric form for the mean is linear, with

$$\mu(s;\beta) = X(s)^T\beta$$

Things we might include in X are

- An intercept
- A trend surface model (polynomial terms in x and y)
- Other spatial covariates

Classical geostatistical practice is to form a preliminary estimate of $\beta$, then estimate the variogram, then re-estimate $\beta$ taking the dependence into account. This procedure isn't optimal in any statistical sense, but it can be useful for exploratory data analysis.



The ordinary least squares (OLS) estimate of $\beta$ minimizes

$$\begin{aligned} SSE(\beta) &= \sum_{i=1}^{n}[Y(s_i) - X(s_i)^T\beta] \\ &= (Y - X\beta)^T(Y - X\beta) \end{aligned}$$

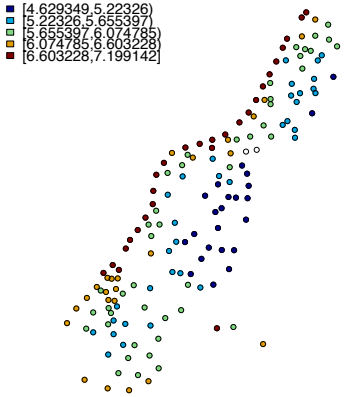
where $X$ is an $n \times p$ matrix whose $i^{th}$ row is $X(s_i)^T$.

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^TY$$

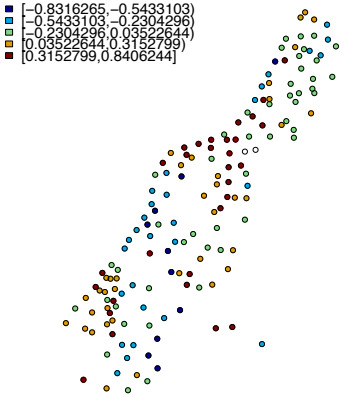
and now we will estimate the variogram for the residual vector

$$\hat{e} = Y - X\hat{\beta}_{OLS}$$

Fitted




Residuals






Note that since $E[e(s) = 0]$,

$$\begin{aligned} 2\gamma(h) &= Var[e(s+h) - e(s)] \\ &= E[(e(s+h) - e(s))^2] \end{aligned}$$

For each pair of points, we have both the vector $h$ (the difference in their coordinates) and a squared difference.

The *method of moments* finds an estimator by equating a distributional moment with its sample equivalent. We typically have to bin the data to estimate the variogram, since we don't have replication for each $h$.



Let $H_1, \ldots, H_k$ be a partition of the space of possible lags (differences in coordinates), with $h_u$ giving a representative member of $H_u$. Let

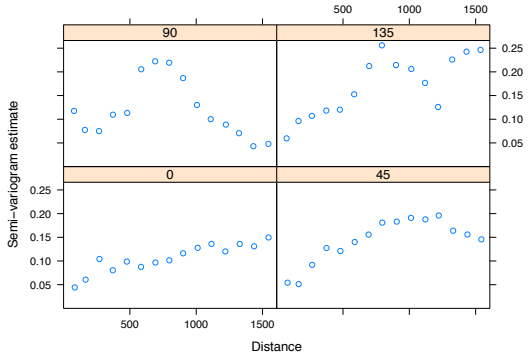
$$\hat{\gamma}(h_u) = \frac{1}{2\#\{s_i - s_j \in H_u\}} \sum_{\{s_i - s_j \in H_u\}} [\hat{e}(s_i) - \hat{e}(s_j)]^2$$

Under the assumption of isotropy, the partition would only take into account the distance $||h||$.

More generally, we can examine plots against distance, stratified by angle.



Our residuals show some evidence of anisotropy.




Looking back at the map of the residuals, it does seem that the strength of spatial correlation varies with orientation relative to the river.

Geometric anisotropy means that the variogram is isotropic under a linear transformation of the coordinate space. That is, $\gamma(h) = \gamma^0(||Ah||)$.

We might parameterize $A$ using a rotation and scaling

$$A = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix} \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}$$

setting $s_x$ or $s_y$ to 1 if $\gamma^0$ has a range parameter.

Classical geostatistical methods are better suited to isotropic covariance functions, so we'll defer fitting this model for now.



Traditionally, the non-parametric estimate of the isotropic variogram is then used to estimate a parametric model for $\gamma(h)$ using weighted least squares. $\hat{\theta}_{WLS}$ minimizes

$$\sum_u \frac{n_u}{\gamma(h_u;\theta)^2}[\hat{\gamma}(h_u) - \gamma(h_u;\theta)]^2$$

where $n_u = \#\{s_i - s_j \in H_u\}$.

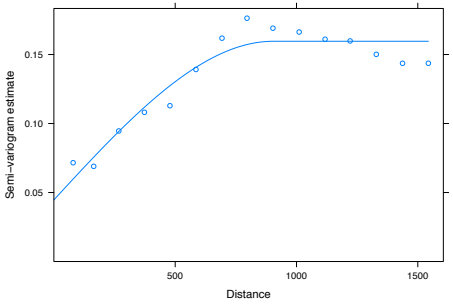
Note that the criterion tends to downweight variogram estimates with small sample sizes or large lags.



We choose a parametric form for $\gamma$ and estimate its parameters. Let's use the spherical variogram

$$\gamma(h;\theta) = \begin{cases} 0 & h = 0 \\ \theta_3 + \theta_1 \left( \frac{3h}{2\theta_2} - \frac{h^3}{2\theta_2^3} \right) & 0 < h \leq \theta_2 \\ \theta_3 + \theta_1 & h > \theta_2 \end{cases}$$

(Exercise: Calculate the covariance function corresponding to the spherical variogram, under second-order stationarity.)






Finally, we may re-estimate $\beta$ using generalized least squares (GLS), plugging in estimates of the covariance parameters from the last stage.

$$\hat{\beta}_{GLS} = (X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}Y$$

where $\hat{\Sigma}_{ij} = C(s_i - s_j; \hat{\theta}_{WLS})$.

If the parameters were actually known, the GLS estimate would be the best (smallest variance) linear unbiased estimator for $\beta$.

Properties of the multivariate normal distribution

Suppose $Y|\mu, \Sigma \sim MVN(\mu, \Sigma)$ where $Y$ and $\mu$ are vectors of length $n$, and $\Sigma$ is an $n \times n$ positive definite matrix. $Y$ has density

$$p(Y; \mu, \Sigma) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\{(Y-\mu)^T \Sigma^{-1}(Y-\mu)\}$$

Computing $|\Sigma|$ and solving $\Sigma^{-1}(Y-\mu)$ are computationally expensive when $n$ is large, which will motivate some approximations and alternate models we will see.



If we partition Y and its parameters as

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} | \mu, \Sigma \sim MVN\left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

we have

- *Marginal distribution* $Y_1 \sim MVN(\mu_1, \Sigma_{11})$
- *Conditional distribution* $Y_2|Y_1 \sim MVN(\mu_{2|1}, \Sigma_{2|1})$

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Y_1 - \mu_1)$$
$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$



These equations are closely related to *kriging*, a more general problem for spatial processes.

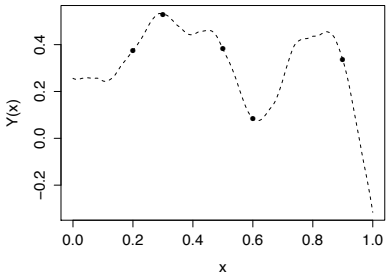
Suppose we observe vector $Y \equiv (Y(s_1), \ldots, Y(s_n))^T$ and we want to predict $Y(s_0)$ for some new location $s_0$. The kriging predictor is the *best linear unbiased predictor (BLUP)*. That is, among all predictors satisfying

- *Linearity:* $\hat{Y}(s_0) = \lambda_0 + \lambda^T Y$
- *Unbiasedness:* $E[\hat{Y}(s_0) - Y(s_0)] = 0$

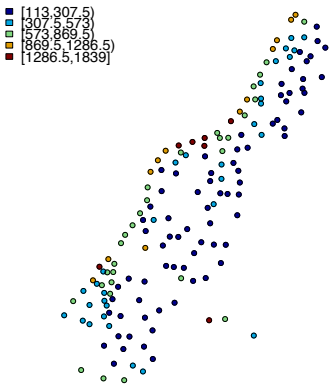
the BLUP minimizes $Var[\hat{Y}(s_0) - Y(s_0)]$



## Toy example




## Real example



First, let's be precise with notation for multivariate statistics. If $X \in \Re^k$ is a random vector, denote $E[X] = [E(X_1), E(X_2), \ldots, E(X_k)]^T \equiv \mu$ and

$$\begin{aligned} Var(X) &= E[(X-\mu)(X-\mu)^T] \\ &= \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_k) \\ Cov(X_1,X_2) & Var(X_2) & \cdots & Cov(X_2,X_k) \\ \vdots & & \ddots & \vdots \\ Cov(X_1,X_k) & \cdots & \cdots & Var(X_k) \end{bmatrix} \\ &\equiv \Sigma \end{aligned}$$

The matrix $\Sigma$ must be symmetric and non-negative definite, i.e. $z^T \Sigma z \geq 0 \quad \forall z$.

If $Y \in \Re^r$ has mean $\eta$, $Cov(X,Y) = E[(X-\mu)(Y-\eta)^T]$.



Some facts about linear transformations (assuming conformable dimensions):

$$\begin{aligned} E[AX+b] &= A\,E[X] + b \\ Var(AX+b) &= A\,Var(X)\,A^T \\ Cov(AX+b, CY+d) &= A\,Cov(X,Y)\,C^T \\ Var(X+Y) &= Var(X) + Cov(X,Y) + Cov(Y,X) + Var(Y) \end{aligned}$$

We will also need the derivative for scalar $y$ with respect to vector $\lambda$, defined as

$$\frac{\partial y}{\partial \lambda} = \left[\frac{\partial y}{\partial \lambda_1}, \frac{\partial y}{\partial \lambda_2}, \ldots, \frac{\partial y}{\partial \lambda_k}\right]^T$$



Back to kriging, our basic problem is to observe $Y$ and use it to form a prediction of $Y(s_0)$. Denote their moments by

$$\begin{pmatrix} Y \\ Y(s_0) \end{pmatrix} \sim \left( \begin{pmatrix} m \\ m_0 \end{pmatrix}, \begin{bmatrix} \Sigma & k \\ k^T & \sigma^2 \end{bmatrix} \right)$$

*Setup #1: Simple kriging*

Assume $E[Y(s)] = \mu(s)$ and $Cov(Y(s), Y(t)) = C(s-t;\theta)$, where *both $\mu$ and $\theta$ are known.*

Considering estimators of the form $\hat{Y}(s_0) = \lambda_0 + \lambda^T Y$, note that

$$E[\hat{Y}(s_0) - Y(s)] = 0 \Rightarrow \lambda_0 = m_0 - \lambda^T m$$



We now seek $\lambda$ to minimize

$$Var[\hat{Y}(s_0) - Y(s_0)] = \lambda^T \Sigma \lambda - 2\lambda^T k + \sigma^2$$

$$\begin{aligned} &\frac{\partial}{\partial \lambda} Var[\hat{Y}(s_0) - Y(s_0)] = 2\Sigma\lambda - 2k \equiv 0 \\ \Rightarrow\ &\lambda = \Sigma^{-1} k \\ \Rightarrow\ &\boxed{\hat{Y}(s_0) = m_0 + k^T \Sigma^{-1}(Y - m)} \end{aligned}$$

The mean squared error of this predictor is

$$\boxed{Var[\hat{Y}(s_0) - Y(s_0)] = \sigma^2 - k^T \Sigma^{-1} k}$$

That this predictor is the BLUP does not require a specific distribution, but note the correspondence with the conditional distributions of a MVN.

*Setup #2: Universal kriging*

Now assume $E[Y(s)] = X(s)^T\beta$ where $\beta \in \Re^p$ is unknown but $\theta$ is still known. The generalized least squares estimator for $\beta$ is

$$\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

If we plug in $\hat{\beta}_{GLS}$ for $\beta$ in the simple kriging predictor, we obtain

$$\hat{Y}(s_0) = x_0^T\hat{\beta}_{GLS} + k^T\Sigma^{-1}(Y - X\hat{\beta}_{GLS})$$



$$\hat{Y}(s_0) = x_0^T\hat{\beta}_{GLS} + k^T\Sigma^{-1}(Y - X\hat{\beta}_{GLS})$$

The method of Lagrange multipliers can be used to prove that this is exactly the BLUP in this case.

Its mean squared error is

$$V \equiv \sigma^2 - k^T\Sigma^{-1}k + b^T(X^T\Sigma^{-1}X)^{-1}b$$

where $b = x_0 - X^T\Sigma^{-1}k$.



Bayesian kriging (see e.g. Handcock & Stein, 1993)

We can make another link with MVN distributions if we consider the model

$$\begin{pmatrix} Y \\ Y(s_0) \end{pmatrix} | \beta \sim MVN\left(\begin{pmatrix} m \\ m_0 \end{pmatrix}, \begin{bmatrix} \Sigma & k \\ k^T & \sigma^2 \end{bmatrix}\right)$$

with noninformative prior distribution $p(\beta) \propto 1$. Then

$$Y_0|Y, \beta \sim MVN(x_o^T\beta + k^T\Sigma^{-1}(Y - X\beta), \sigma^2 - k^T\Sigma^{-1}k)$$

$$\beta|Y \sim MVN(\hat{\beta}_{GLS}, (X^T\Sigma^{-1}X)^{-1})$$

$$Y_0|Y \sim MVN(\hat{Y}(s_0), V)$$



What do we do if $\theta$, the covariance parameters, are also unknown?

One option is to plug in an estimate, obtained from classical geostatistical techniques or maximum likelihood, for example. This is known as the empirical BLUP (EBLUP), a bit of a misnomer, since it is not even an unbiased linear predictor anymore.

The naive estimate of its mean squared error (plugging in $\hat{\theta}$ here as well) tends to underestimate the true error, as it ignores variability due to estimating $\theta$.

In the Bayesian framework, this additional uncertainty is handled very naturally, by integrating over a prior distribution for $\theta$.

For example, if $Cov(Y(s), Y(t)) = \sigma^2 R(s-t)$ for a known function $R$, specifying the Jeffreys prior $p(\beta, \sigma^2) \propto 1/\sigma^2$ yields a posterior distribution for $Y(s_0)$ that is a t-distribution centered at $\hat{\beta}_{GLS}$ (Handcock & Stein, 1993).

If the parameters of $R$ are also unknown, one can specify a prior for them as well and make inference using MCMC (more on this in Ch 7). Care is needed, since some commonly used improper priors produce improper posteriors (Berger et al., 2001).



Likelihood based methods

Suppose $Y(\cdot)$ is a Gaussian process with mean function $\mu(s) = X(s)^T\beta$ and covariance function

$$Cov[Y(s), Y(t)] = C(s, t; \theta)$$

For observations $Y \equiv (Y(s_1), \ldots, Y(s_n))^T$ we can construct the mean vector $X\beta$ and $n \times n$ covariance matrix $\Sigma(\theta)$; it has entries

$$\Sigma(\theta)_{i,j} = C(s_i, s_j; \theta)$$

The distribution is then $Y \sim MVN(X\beta, \Sigma(\theta))$.



The *likelihood function* for $\beta$ and $\theta$ is the probability density for the observed $Y$, but viewed as a function of $\beta$ and $\theta$.

$$\mathcal{L}(\beta, \theta) = p(Y; \beta, \theta) = (2\pi)^{-n/2}|\Sigma(\theta)|^{-1/2} \exp\left\{-\frac{1}{2}(Y - X\beta)^T\Sigma(\theta)^{-1}(Y - X\beta)\right\}$$

The *maximum likelihood estimators* of $\beta$ and $\theta$ maximize this function, or, equivalently, the *log-likelihood function*

$$\ell(\beta, \theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}(Y - X\beta)^T\Sigma(\theta)^{-1}(Y - X\beta)$$

Actually, the linear structure of the mean is not crucial, and could be replaced by a nonlinear function $f(X(s), \beta)$. However...



The linear form of the mean allows us to simplify the optimization problem by *profiling*.

If we were to fix $\theta$ at a given value, we could find a corresponding $\beta$ to maximize the likelihood. It turns out we can find this in closed form, and it's just the GLS estimator

$$\hat{\beta}(\theta) = (X^T\Sigma(\theta)^{-1}X)^{-1}X^T\Sigma(\theta)^{-1}Y$$

The profile log-likelihood is a function only of $\theta$ :

$$\begin{aligned}\ell(\theta) &\equiv \ell(\hat{\beta}(\theta), \theta) \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}(Y - X\hat{\beta}(\theta))^T\Sigma(\theta)^{-1}(Y - X\hat{\beta}(\theta))\end{aligned}$$

Maximizing $\ell(\theta)$ to obtain $\hat{\theta}$ and then plugging in to get $\hat{\beta}(\hat{\theta})$ is equivalent mathematically to maximizing $\ell(\beta, \theta)$ over both parameters.

However, since the dimension of $\theta$, is often larger than that of $\beta$ this can substantially simplify the maximization problem.

Closed form solutions for $\hat{\theta}$ are rare, and we typically turn to numerical optimization techniques. We need to consider such things as

- Starting values and stopping rules
- Enforcing any parameter constraints
- Checking whether the solution is a global max



In the case that $C(s,t) = \sigma^2 R(s,t;\phi)$, we can also profile over $\sigma^2$. Writing $\Sigma(\sigma^2, \phi) = \sigma^2 K(\phi)$, we have

$$\ell(\beta, \sigma^2, \phi) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|K(\phi)| - \frac{1}{2\sigma^2}(Y - X\beta)^T K(\phi)^{-1}(Y - X\beta)$$

For fixed $\phi$, the maximizing values are

$$\hat{\beta}(\phi) = (X^T K(\phi)^{-1} X)^{-1} X^T K(\phi)^{-1} Y$$

$$\sigma^2(\phi) = (Y - X\hat{\beta}(\phi))^T K(\phi)^{-1}(Y - X\hat{\beta}(\phi))/n$$

So the profile likelihood is

$$\ell(\phi) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2(\phi)) - \frac{1}{2}\log|K(\phi)| - \frac{n}{2}$$



The "large n" problem

When the number of observations is large, evaluating the log-likelihood or profile log-likelihood can be very slow.

This is because two key operations involving $\Sigma$, finding its determinant and solving a linear system, both require $O(n^3)$ operations.

Some options:

- Change the algorithm (special cases)
- Change the model (e.g. low-rank or lattice model)
- Approximate the likelihood



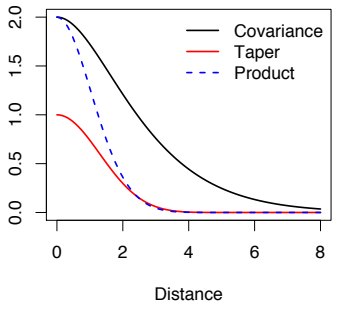
Vecchia (1988) proposed the following approximation, later extended by Stein et al. (2004):

$$\begin{aligned}\mathcal{L}(\beta,\theta) = p(Y;\beta,\theta) &= p(Y_1;\beta_1,\theta)\prod_{i=1}^{n} p(Y_i|Y_1,\ldots,Y_{i-1};\beta,\theta) \\ &\approx p(Y_1;\beta_1,\theta)\prod_{i=1}^{n} p(Y_i|S_{(i-1)};\beta,\theta)\end{aligned}$$

subset consisting of nearby points

Kaufman et al. (2008) proposed *covariance tapering*.

The idea is to multiply the original covariance function by a correlation function with compact support.

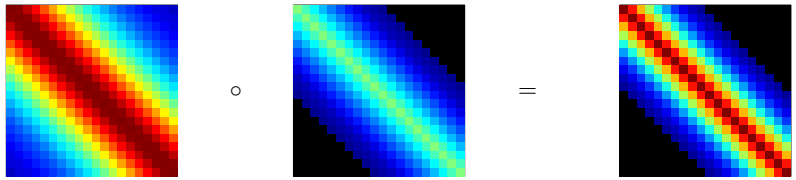



This produces another valid covariance function.

The resulting matrices can be manipulated using sparse matrix algorithms.



Equivalently, we can think of tapering a matrix. This uses the Schur (direct) matrix product $\Sigma \circ T$.



$$\begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{bmatrix} \circ \begin{bmatrix} T_{11} & \cdots & T_{1n} \\ \vdots & \ddots & \vdots \\ T_{n1} & \cdots & T_{nn} \end{bmatrix} = \begin{bmatrix} \Sigma_{11}T_{11} & \cdots & \Sigma_{1n}T_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1}T_{n1} & \cdots & \Sigma_{nn}T_{nn} \end{bmatrix}$$

$$\{K(||s_i - s_j||;\theta)\} \circ \{K_{tap}(||s_i - s_j||;\gamma)\} = \{KK_{tap}(||s_i - s_j||;\theta,\gamma)\}$$



## Likelihood Approximations

- Original log-likelihood:

$$\ell(\theta) = -\frac{1}{2}\log\det\Sigma(\theta) - \frac{1}{2}Z'\Sigma(\theta)^{-1}Z$$

- Approximation 1: Replace $\Sigma(\theta)$ by $\Sigma(\theta) \circ T$

$$\ell_{1taper}(\theta) = -\frac{1}{2}\log\det[\Sigma(\theta) \circ T] - \frac{1}{2}Z'[\Sigma(\theta) \circ T]^{-1}Z$$

- Approximation 2: Note that $Z'\Sigma^{-1}Z = \mathrm{tr}\{ZZ'\Sigma^{-1}\} = \mathrm{tr}\{\hat{\Sigma}\Sigma^{-1}\}$

$$\begin{aligned} \ell_{2tapers}(\theta) &= -\frac{1}{2}\log\det[\Sigma(\theta) \circ T] - \frac{1}{2}\mathrm{tr}\left\{[\hat{\Sigma} \circ T][\Sigma(\theta) \circ T]^{-1}\right\} \\ &= -\frac{1}{2}\log\det[\Sigma(\theta) \circ T] - \frac{1}{2}Z'\left([\Sigma(\theta) \circ T]^{-1} \circ T\right)Z \end{aligned}$$



Hierarchical models and Bayesian spatial statistics

A *hierarchical model* specifies a joint probability distribution by decomposing it into conditional distributions. These are often easier to specify.

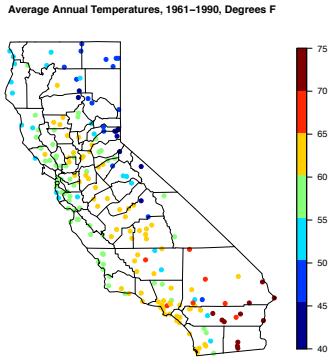
In the context of spatial statistics, what we typically want to make inference about is some underlying *process* that generated the data. The distribution of this process is often where the spatial dependence is specified.

We will discuss hierarchical models as used in Bayesian statistics.

Example 1: Temperatures in California, 1961-1990

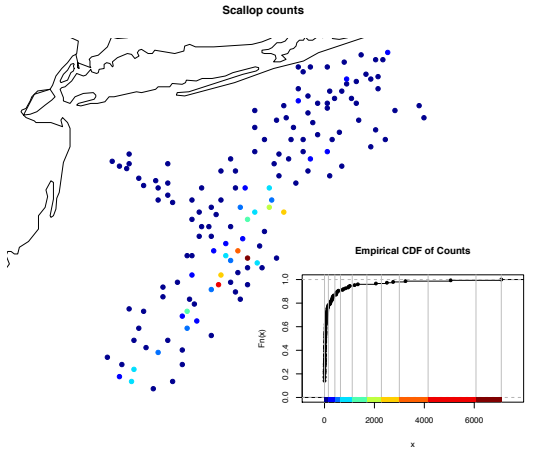
Data is from NCDC surface meteorology stations and represents average temperatures over the period 1961-1990.

The process of interest is the field of true average temperatures over this time period. It seems reasonable to model the observations themselves as being conditionally independent, given this process.






Example 2: Scallop abundance during a survey cruise
The process of interest is a smooth field representing the expected number of counts at a given location.






We'll return to the spatial context next time. For now, let's consider Bayesian statistics more generally.

Bayesian statistics is built on Bayes Theorem.

Likelihood

Posterior density

Prior density

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

Normalizing constant

Shorthand: Posterior $\propto$ Likelihood x Prior



There is a special case for which calculating $p(\theta|y)$ is mathematically tractable. This is when the prior $p(\theta)$ is *conjugate*. This means that $p(\theta)$ and $p(\theta|y)$ come from the same family; conditioning on the data only changes the parameters of that family.

Some common examples:

| Likelihood | Prior |
|---|---|
| $\sim N(\theta, \sigma^2)$, $\sigma^2$ know | $\theta \sim N(a, b)$ |
| $Y \sim N(\mu, \theta)$, $\mu$ known | $\theta \sim InvGamma(a, b)$ |
| $Y \sim Poisson(\theta)$ | $\theta \sim Gamma(a, b)$ |
| $Y \sim Binomial(n, \theta)$ | $\theta \sim Beta(a, b)$ |

The key to calculating the posterior when the prior is conjugate is to recognize the *kernel* of the prior and posterior densities. This is the part that depends on $\theta$.

Knowing the kernel of a density is the same as knowing the density itself, since the density must integrate to one.

Example: $\theta \sim InvGamma(a, b)$

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{-a-1} \exp\left\{-\frac{b}{\theta}\right\}$$

the kernel



To calculate the posterior, then, we find the terms in Bayes theorem that depend on $\theta$ and isolate the kernel of a known density.

Example, continued: $Y|\theta \sim N(\mu, \theta)$ with $\mu$ known

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^{-1/2} \exp\left\{-\frac{(y-\mu)^2}{2\theta}\right\} \theta^{-a-1} \exp\left\{-\frac{b}{\theta}\right\} \\ &= \theta^{-(a+1/2)-1} \exp\left\{-\frac{1}{\theta}\left(b + \frac{(y-\mu)^2}{2}\right)\right\} \end{aligned}$$

We've shown $\theta|Y \sim InvGamma(a + 1/2, b + (Y-\mu)^2/2)$



Another example: Over Feb 2-5, 2011, Gallup conducted a random digit dialing survey in which 1,015 adult participants were asked

*Overall, are you sympathetic or unsympathetic to the protestors in Egypt who have called for a change in the government?*

Let $\theta$ denote the true proportion of U.S. adults who would say they are sympathetic, and consider the model

$$\begin{aligned} Y|\theta &\sim Binomial(1015, \theta) \\ \theta &\sim Beta(a, b) \end{aligned}$$



- Choose values of $a$ and $b$ subjectively, to reflect your prior beliefs about $\theta$.
- Calculate the posterior distribution for $\theta$, given $Y =$ ____.
- Explore the effects of changing the prior parameters on the resulting posterior.
- What if the prior is instead $\theta \sim Unif(0, 1)$ ?
- What happens to the posterior if $Y/n$ is constant but $n$ increases or decreases?

It is fairly rare that we can calculate the posterior in closed form. Instead, we often make use of algorithms to *sample* from the posterior and then approximate distributional quantities by their sample equivalents.

One popular algorithm for doing this is the *Gibbs sampler*. It produces a Markov chain whose limiting distribution is the posterior.

Suppose our model has parameters $\theta_1, \ldots, \theta_k$. To implement the algorithm, we need to be able to sample from each of the *full conditional distributions*

$$p(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k, Data).$$



Gibbs sampling algorithm:

1. Choose starting values $\theta_1^{(1)}, \ldots, \theta_k^{(1)}$.

2. For $i = 2, \ldots, B$ sample

$$\theta_1^{(i)}|\theta_2^{(i-1)}, \ldots, \theta_k^{(i-1)}, Data$$
$$\theta_2^{(i)}|\theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_k^{(i-1)}, Data$$
$$\vdots$$
$$\theta_k^{(i)}|\theta_1^{(i)}, \ldots, \theta_{k-1}^{(i)}, Data$$

In practice, we then examine the sample for evidence of convergence.



Example: Consider data $Y_1 \ldots, Y_n|\mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$ and prior $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$ with $\mu \sim N(a, b)$ $\sigma^2 \sim InvGamma(c, d)$. Calculate the full conditionals and implement a Gibbs sampler.

Two notes:

• The priors in this case are what are called conditionally conjugate, meaning the full conditionals have the same form as the prior.

• The conjugate prior in this case is called a normal inverse gamma distribution and decomposes into

$$\mu|\sigma^2 \sim N(a, \sigma^2/b) \qquad \sigma^2 \sim InvGamma(c, d)$$



Diagnosing "convergence" of the Markov chain can be tricky. A variety of tests exist, but truly pathological cases will only be caught if the chain runs a long time.

At a minimum, you should examine the sample paths, estimate their autocorrelation function, and calculate the effective sample size, which is defined as

$$ESS = N/[1 + 2\sum_{k=1}^{\infty} \rho_k]$$

where $\rho_k$ is the autocorrelation at lag $k$.

We choose a value $b \in 1, \ldots, B$, called the *burn-in*, and discard samples prior to this.

A more general-purpose algorithm is *Metropolis-Hastings (MH) sampling*. We need a starting value $\theta^{(1)}$ and *proposal density* $q(\cdot;\theta)$.

For $i = 2, \ldots, B$

1. Draw $\theta^{cand} \sim q(\cdot;\theta^{(i-1)})$

2. Compute $r = \dfrac{p(y|\theta^{cand})p(\theta^{cand})q(\theta^{(i-1)};\theta^{cand})}{p(y|\theta^{(i-1)})p(\theta^{(i-1)})q(\theta^{cand};\theta^{(i-1)})}$

3. With probability $\min\{r, 1\}$, set $\theta^{(i)} = \theta^{cand}$. Otherwise, set $\theta^{(i)} = \theta^{(i-1)}$.



Returning to the case of spatial data, we now have some latent (unobserved) process of interest.

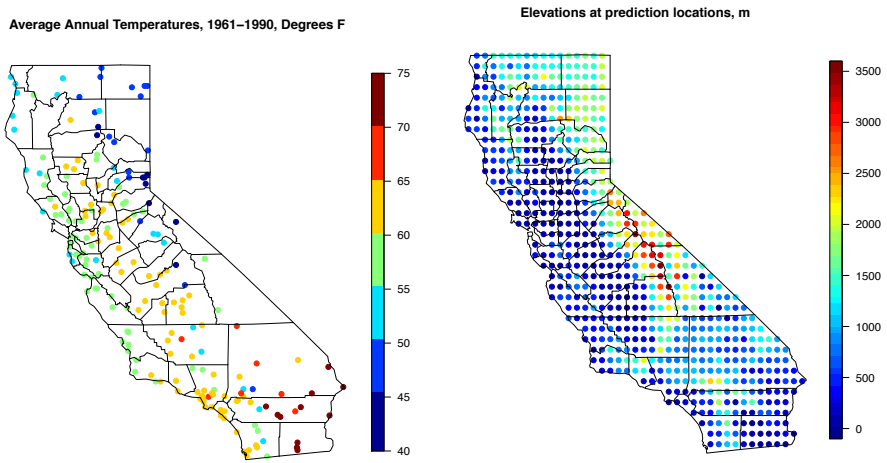
Letting $Y$ denote the data, $\eta$ the process, and $\theta$ the parameters, Bayes Theorem now becomes

$$\begin{aligned} p(\eta, \theta|Y) \quad \propto \quad & p(Y|\eta,\theta) \times && \text{Data model / likelihood} \\ & p(\eta|\theta) \times && \text{Process model / prior} \\ & p(\theta) && \text{Parameter model / hyperprior} \end{aligned}$$

Let's specify this for the CA temperature example.



The data: from NCDC and USGS






I find it easiest to start with the process model. Here, let $\eta$ represent the average temperature field, over all locations in California.

We can assume $\eta$ varies smoothly in space, and possible covariates we might relate it to are longitude, latitude, and elevation.

Some preliminary (classical geostatistical) analysis indicates that all three covariates are useful, and there is some evidence of anisotropy in the residuals. For now I will ignore the anisotropy.

Suppose we take

$$\eta|\beta,\sigma^2,\rho \sim GP(X(\cdot)^T\beta, \sigma^2 K(\cdot,\cdot;\rho,\nu))$$

where $X$ contains an intercept, longitude, latitude, and elevation, and

$$K(s_i,s_j;\rho,\nu) = \frac{(||s_i - s_j||/\rho)^\nu \mathcal{K}_\nu(||s_i - s_j||/\rho)}{2^{\nu-1}\Gamma(\nu)}$$

i.e., the Matern correlation function.

Also, define $\eta_{obs} = (\eta(s_1),\ldots,\eta(s_n))^T$
$\eta_{pred} = (\eta(s_1^*),\ldots,\eta(s_m^*))^T$



By specifying the GP distribution for $\eta$, we have the joint distribution for $\eta_{obs}$ and $\eta_{pred}$ (conditional on the relevant parameters), which we could write as

$$\begin{aligned} p(\eta_{obs},\eta_{pred}|\beta,\sigma^2,\rho,\nu) &= p(\eta_{obs}|\beta,\sigma^2,\rho,\nu) \times \\ & \quad p(\eta_{pred}|\eta_{obs},\beta,\sigma^2,\rho,\nu) \end{aligned}$$

For now, note that

$$\eta_{obs}|\beta,\sigma^2,\rho,\nu \sim MVN(X\beta,\sigma^2\Gamma(\rho,\nu))$$

where $\Gamma(\rho,\nu)_{ij} = K(s_i,s_j;\rho,\nu)$.



Moving now to the data model, our task is relatively easy. We take the observations to be conditionally iid, given $\eta$ and a parameter $\tau^2$ giving the variance of the measurement error, with

$$Y(s_i)|\eta,\tau^2 \sim N(\eta(s_i),\tau^2)$$

Note that an equivalent way to write this is

$$Y|\eta,\tau^2 \sim MVN(\eta_{obs},\tau^2 I_n)$$

where $I_n$ is the $n \times n$ identity matrix.

So far, this is very similar to what we've done before.



To see the connection to what we've done before, note what happens if we find the marginal distribution for $Y$ given the parameters (but not $\eta$).

$$Y|\beta,\sigma^2,\rho,\nu,\tau^2 \sim MVN(X\beta,\sigma^2\Gamma(\rho,\nu)+\tau^2 I_n)$$

This is the same form we considered when we calculated the MLEs, where in this case

$$\Sigma(\theta) = \sigma^2\Gamma(\rho,\nu)+\tau^2 I_n$$

The key distinction is that now, rather than estimating the parameters, fixing them, and plugging in to predict, we will base inference on the joint posterior of $\theta$ and $\eta_{pred}$.

To do this, we need to specify the final part of our model, which is $p(\theta)$. I will choose conditionally conjugate priors where possible:

$$\begin{aligned} \beta &\sim N(m_\beta, V_\beta) \\ \sigma^2 &\sim InvGamma(a_{\sigma^2}, b_{\sigma^2}) \\ \tau^2 &\sim InvGamma(a_{\tau^2}, b_{\tau^2}) \end{aligned}$$

There are no conditionally conjugate priors for $\rho$ and $\nu$. I will use Gamma distributions for them.

$$\begin{aligned} \rho &\sim Gamma(a_\rho, b_\rho) \\ \nu &\sim Gamma(a_\nu, b_\nu) \end{aligned}$$



The joint posterior distribution for all unobserved parts of the model, given the data, is

$$p(\eta_{obs}, \eta_{pred}, \beta, \sigma^2, \rho, \nu, \tau^2 | Y)$$

In practice, we'll probably be most interested in marginal posterior distributions, such as $p(\eta_{pred}|Y)$.

This is treated very easily within the context of MCMC: if we have a sample from the joint posterior, ignoring the other components of the sample gives us a sample from the marginal we want.



For everything but $\rho$ and $\nu$, we can find the full conditional distributions in closed form.

We can use a hybrid MCMC algorithm, in which we embed "Metropolis steps" within a Gibbs sampler. Whenever it is time to sample $\rho|Rest$ or $\nu|Rest$, we carry out a single iteration from the Metropolis algorithm, but using the ratio of, say,

$$p(\rho^{cand}|Rest)/p(\rho^{(i-1)}|Rest)$$

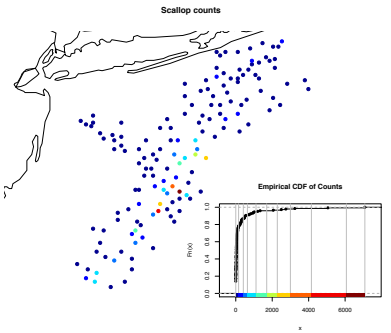
in the acceptance ratio. What makes this possible is that the normalizing constants we'd need to calculate either one of these separately cancel each other out.



Another class of models opens up to us if we keep the GP structure in the process model, but allow the data to be non-Gaussian.

This idea (from Diggle, Tawn, and Moyeed, 1998) is a spatial analogue to *generalized linear models* in regression.

For example, we could use it to model the scallop count data, conditional on an underlying smooth field modeled by a GP.



Let's review generalized linear models (GLMs) in the non-spatial context first.

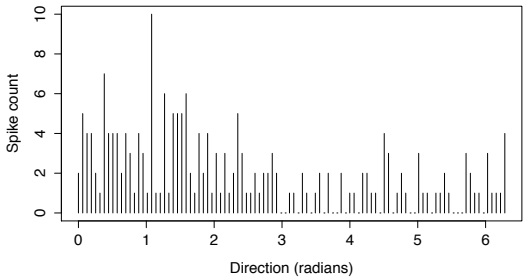
A GLM specifies a particular distribution for a set of observations, and then relates the mean of that distribution to a *linear predictor* $X\beta$ via a *link function* $g$.

$$E[Y_i] = \mu_i = g^{-1}(f_i^T \beta)$$

GLMs have become a standard tool in statistics and encompass a wide variety of models, including Poisson regression, logistic regression (for Binomial outcomes), and standard linear regression with Gaussian residuals.



Example: Neurons in the area of the motor cortex that controls arm movement show evidence of being tuned for a particular "preferred direction". For example, the number of times the neuron fires during a 2D motion task might look something like:

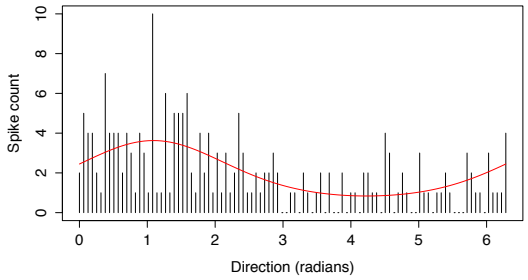





A Poisson regression model:

$$E[Y_i] = \mu_i = exp\{\beta_0 + \beta_1 cos(x_i) + \beta_2 sin(x_i)\}$$

Here the link is the natural log, and the covariates are the cosine and sine of the angle of movement x.






GLMs have been extended in a number of ways, the most relevant for us being the *generalized linear mixed model* (GLMM). We add a *random effect* $Z_i$

$$E[Y_i] = \mu = g^{-1}(x_i^T \beta + Z_i)$$

The distributional assumptions for the random effects depend on the problem. For example, if the observations are repeated measurements on a group of patients, each patient may have his or her own $Z_i$, drawn from a normal distribution where the variance represents patient-to-patient differences not captured by the regressors.

The models proposed by Diggle, Tawn, and Moyeed (1998) are a special class of GLMMs, for which the random effects are spatially correlated.

For example, a potential model for the scallop data is

$$\eta|\beta,\sigma^2,\rho \sim GP(X(\cdot)^T\beta, \sigma^2 K(\cdot,\cdot;\rho,\nu))$$

as before, but now take $Y(s_1),\ldots,Y(s_n)$ to be conditionally iid given $\eta$, with

$$Y(s_i)|\eta \sim Pois\{\exp(\eta(s_i))\}$$



Reduced rank models

We'll now consider a class of models that have shown up in the literature under different guises but which all share a common, computationally efficient representation.

First, recall the "large n" problem we discussed in the context of likelihood-based estimation. This is also a problem in kriging. We need to evaluate expressions like

$$(Y-\mu)^T\Sigma^{-1}(Y-\mu)$$

$$k^T\Sigma^{-1}(Y-\mu)$$



In practice, the most efficient way to compute such expressions is to first form the Cholesky decomposition of $\Sigma$, then backsolve into the vectors or matrices on either side, then take the product.

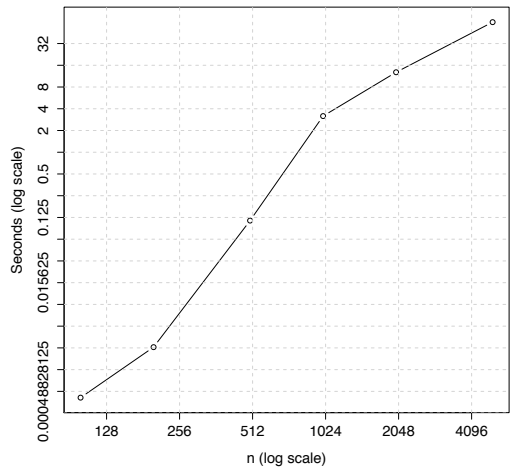
Example:

```
Q <- chol(Sigma)
v <- backsolve(Q, Y - mu, transpose = TRUE)
crossprod(v)
```

The most computationally expensive part of this calculation is the Cholesky decomposition, which requires $O(n^3)$ calculations.



Here's an illustration, with timings from my laptop:




| n | Time to calculate chol 10,000 times (e.g. for MCMC) |
|---|---|
| 100 | 5 seconds |
| 200 | 30 seconds |
| 500 | 20 minutes |
| 1000 | 7 hours |
| 2000 | 30 hours |
| 5000 | 6 days |

The models we'll discuss share a common idea:

When specifying the distribution for the process $\eta$, express it in terms of a random vector $\alpha$ of length $p$, where $p \ll n$.

This is not such an unusual idea. For example, in the random effects model

$$Y_{ij} = X_i^T \beta + Z_j + \epsilon_{ij}$$

with $Z_j \overset{iid}{\sim} N(0, \sigma_z^2)$ and $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ we are expressing the joint distribution of the residuals using a lower-dimensional vector $Z = (Z_1, \ldots, Z_J)$ and a simple covariance matrix $\sigma_\epsilon^2 I$ for the $\epsilon_{ij}$.



I will follow the text in making some simplifying assumptions to clarify the development. In particular, I'll assume the mean of the process $\eta$ is known to be zero. It's not difficult to extend this to be $X\beta$ for unknown $\beta$.

Another note: it's possible to view this approach as either *approximating the original Gaussian geostatistical model* or as *using a different but equally valid model.*

Both motivations take advantage of the computational simplifications, but this distinction has implications for how things are specified in the model.



Mean-zero low-rank Gaussian geostatistical model

$$Y = \eta + \epsilon \qquad \epsilon \sim MVN(0, \sigma^2 I)$$

dimension $n$

$$\eta = H\alpha + \xi \qquad \xi \sim MVN(0, \Sigma_\xi)$$
$$\eta_0 = H_0\alpha + \xi_0 \qquad \xi_0 \sim MVN(0, \Sigma_{\xi_0})$$

$$\alpha \sim MVN(0, \Sigma_\alpha)$$

dimension $p$

$\Sigma_\xi, \Sigma_{\xi_0}$ are diagonal and $\xi, \xi_0$ are independent



An equivalent representation for the process is

$$\begin{pmatrix} \eta \\ \eta_0 \end{pmatrix} | \alpha \sim MVN\left( \begin{pmatrix} H \\ H_0 \end{pmatrix} \alpha, \begin{pmatrix} \Sigma_\xi & 0 \\ 0 & \Sigma_{\xi_0} \end{pmatrix} \right)$$

Integrating out $\alpha$ and then $\eta$, we have

$$\begin{pmatrix} \eta \\ \eta_0 \end{pmatrix} \sim MVN\left(0, \begin{pmatrix} H\Sigma_\alpha H^T + \Sigma_\xi & H\Sigma_\alpha H_0^T \\ H_0\Sigma_\alpha H^T & H_0\Sigma_\alpha H_0^T + \Sigma_{\xi_0} \end{pmatrix}\right)$$

$$\begin{pmatrix} Y \\ \eta_0 \end{pmatrix} \sim MVN\left(0, \begin{pmatrix} H\Sigma_\alpha H^T + \Sigma_\xi + \sigma_\epsilon^2 I & H\Sigma_\alpha H_0^T \\ H_0\Sigma_\alpha H^T & H_0\Sigma_\alpha H_0^T + \Sigma_{\xi_0} \end{pmatrix}\right)$$

Using the properties of the MVN distribution, we therefore have $\eta_0|Y \sim MVN(\tilde{\mu}, \tilde{\Sigma})$, where

$$\tilde{\mu} = (H_0\Sigma_\alpha H^T)(H\Sigma_\alpha H^T + V)^{-1}Y \qquad (V = \Sigma_\xi + \sigma^2 I)$$

$$\tilde{\Sigma} = (H_0\Sigma_\alpha H_0^T + \Sigma_{\xi_0}) - (H_0\Sigma_\alpha H^T)(H\Sigma_\alpha H^T + V)^{-1}(H\Sigma_\alpha H_0^T)$$

For both of these, the key calculation is $(H\Sigma_\alpha H^T + V)^{-1}$ which at first glance still looks intractable (n x n).

However, the form $(H\Sigma_\alpha H^T + V)^{-1}$ is amenable to a matrix identity known as the *Sherman-Morrison-Woodbury formula*.



Sherman-Morrison-Woodbury formula:

$$(A + BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1} + CA^{-1}B)^{-1}CA^{-1}$$

In our case, we have

$$(H\Sigma_\alpha H^T + V)^{-1} = V^{-1} - V^{-1}H(H^T V^{-1} H + \Sigma_\alpha^{-1})^{-1}H^T V^{-1}$$

By our construction,

- $V = \Sigma_\xi + \sigma_\epsilon^2 I$ is diagonal
- $\Sigma_\alpha$ is $p \times p$, where $p \ll n$.
- $(H^T V^{-1} H + \Sigma_\alpha^{-1})$ is also $p \times p$.



We'll now consider three different constructions that arrive at this same basic model structure:

- Orthogonal basis-functions and the Karhunen-Loeve decomposition
- Predictive process models
- Kernel convolution models



Consider a stochastic process constructed as follows:

$$\eta(s) = \sum_{k=1}^{\infty} \phi_k(s) Z_k$$

where the $\phi_k$ form an orthonormal basis on $D \subseteq \mathcal{R}^d$ and the $Z_k$ are independent random variables with $Z_k \sim N(0, \lambda^k)$.

We can calculate the covariance of this process:

$$Cov(\eta(s), \eta(r)) = \sum_{k=1}^{\infty} \phi_k(s)\phi_k(r)\lambda_k$$

It turns out that any mean zero GP can be represented in this way. This is called the *Karhunen-Loeve decomposition*.

This might motivate our choices in

$$\eta = H\alpha + \xi \qquad \eta_0 = H_0\alpha + \xi_0$$

i.e. to model $\eta(s) = \sum_{k=1}^{p} \phi_k(s)\alpha_k + \xi(s)$

$$= H(s)^T\alpha + \xi(s)$$

where $\alpha \sim MVN(0, diag(\lambda_1, \ldots, \lambda_p))$

This falls under the motivation of approximating the true model. The difficulty with this approach is that except in special cases, one can't find $\phi_k$ and $\lambda_k$ in closed form.



To find $\phi_k$ and $\lambda_k$ requires solving

$$\int_D c_\eta(s,r)\phi_k(s)ds = \lambda_k\phi_k(r), \quad k = 1, \ldots, \infty$$

We can, however, solve the discrete version of these equations defined over the points for which we have observations:

$$\Sigma_\eta\Phi = \Phi\Lambda$$

where $\Sigma_{\eta,ij} = Cov(\eta(s_i), \eta(s_j))$

This is just an eigenvalue equation, with the solution being $\Phi$ equal to the matrix of eigen-vectors and $\Lambda$ a diagonal matrix of eigen-values for $\Sigma_\eta$.



The downsides to the approach on the last slide, known as Empirical Orthogonal Functions (EOFs) in meteorology and Principal Component Analysis (PCA) in statistics, are that

- We still need an estimate of $\Sigma_\eta$.
- We do not solve the equations for the locations where we want to predict, so we need to interpolate the $\phi_k$ in some way.

Due to these difficulties, people often *choose* a set of orthogonal basis functions, rather than matching them to a particular covariance function.



*Predictive process models* (Banerjee et al., 2008) are another way to approximate the original GP model.

The key idea here is that the process at the observation and prediction locations is modeled as the *kriged predictions of a latent set of observations* under the original GP model.

This latent set of observations is defined at a set of *knots* $S^* = \{s_1^* \ldots, s_p^*\}$. Specifically, take

$$\eta^* \sim MVN(0, C^*(\theta))$$

where $C^*(\theta)_{ij} = C(s_i^*, s_j^*; \theta)$ for some fixed covariance function $C$.

Now define a new process $\eta \sim GP(0, \tilde{C})$ which is obtained by kriging $\eta^*$ to any arbitrary new location. That is, define

$$\eta(s) = c(s;\theta)^T C^*(\theta)^{-1} \eta^*$$

where $c(s;\theta)^T$ contains the covariances for location $s$ and each of the knots in $S^*$. In matrix notation, we have

$$\alpha = \eta^* \sim MVN(0, C^*(\theta))$$

$$\eta = H\alpha + \xi$$

and the rows of $H$ are $c(s_i;\theta)^T C^*(\theta)^{-1}$ for $i = 1, \ldots, n$. In the paper, $\xi = 0$, so the approximation errors are absorbed into the measurement error term.

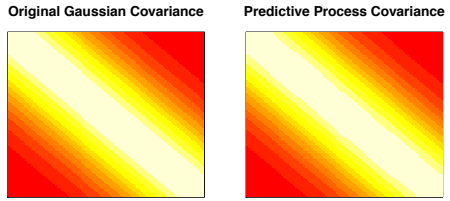


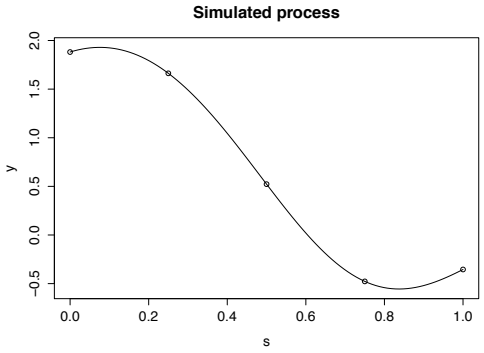
Example: Predictive process with Gaussian covariance function

$\sigma^2 = 1; \quad \theta = 0.2$

What do you expect will happen if we change $\theta$?

What if we change the number of knots?






*Kernel convolution models* (Higdon, Swall, and Kern, 1999) are motivated by the representation of a GP as a convolution of a kernel function $k$ and a Gaussian white noise process $x$.

$$\eta(s) = \int_D k(s - r;\theta) x(r) dr$$

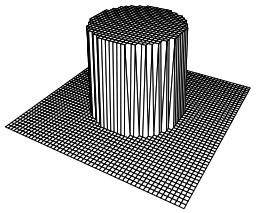
The kernel function satisfies

$$\int_D k(r) dr = 1 \qquad \int_D r k(r) dr = 0$$

and is often isotropic, although not always. The white noise process is independent at each $r$ with a common variance $\sigma_x^2$.

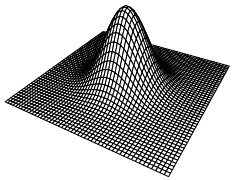


Example kernels:



## Top hat

$$k(x;\theta) = \frac{1}{\pi\theta^2} I\{||x|| < \theta\}$$



## Gaussian

$$k(x;\theta) = \frac{1}{2\pi\theta^2} \exp\left\{-\frac{||x||^2}{2\theta^2}\right\}$$

The covariance function can be calculated from a given kernel:

$$Cov(\eta(s), \eta(s')) = \sigma_x^2 \int_D k(s - r; \theta) k(s' - r; \theta) dr$$

In the isotropic case, there is a one-to-one relationship between the covariance function and the kernel, but in the general case, different kernels can give the same covariance function.

Note that $\sigma_x^2$ is not the marginal variance of this process; it is

$$\sigma_x^2 \int_D k(r; \theta)^2 dr$$



For computational purposes, rather than using a white noise process defined everywhere on $D$, consider a process defined only on the finite *support points*

$$R = \{r_1, \ldots, r_p\}$$

For notational consistency, let's now call this "process" (really just a random vector) $\alpha$. Now

$$\begin{aligned} \eta(s) &= \int_D k(s - r; \theta) \alpha(r) dr \\ &= \sum_{j=1}^{p} k(s - r_j; \theta) \alpha_j \end{aligned}$$



$$\eta(s) = \sum_{j=1}^{p} k(s - r_j; \theta) \alpha_j$$

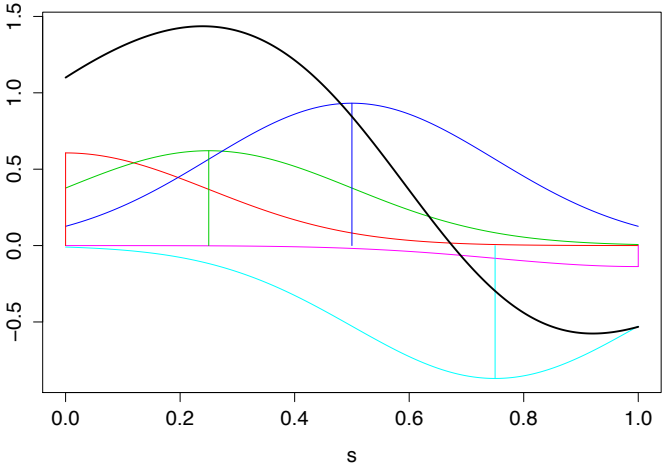
Another way to view this model is that we have chosen a non-orthogonal set of basis functions $k_j(s; \theta) = k(s - r_j; \theta)$.

If we are modeling the process $\eta$ at a finite number of locations $S = \{s_1, \ldots, s_n\}$, we can write this in matrix notation as

$$\eta = H\alpha \qquad H_{ij} = k(s_i - r_j; \theta) \qquad \alpha \sim MVN(0, \sigma_\alpha^2 I)$$

Again, there is no $\xi$ term.






What do you expect will happen as we change $\theta$? What about the number of support points?

*Nonstationary models* allow the distribution to vary spatially. Typically, people reserve this term for the covariance, since it isn't uncommon to have a spatially varying mean term (universal kriging).

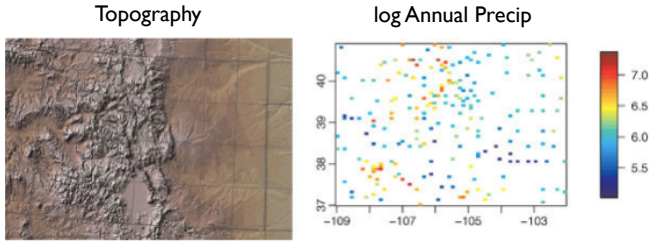
The intuitive motivation behind using a nonstationary model is that if the strength of the spatial correlation is changing, our predictions will suffer if we use a common correlation everywhere.

In practice, however, it can be very difficult to fit nonstationary models, and there is not yet a consensus that predictions always benefit by this extra flexibility in the model.



To think about why this might be the case, imagine a naive approach to nonstationary models: break the spatial area up into pieces, estimate a stationary model for each one, and krige using those parameters and that subset of the data.

An example to consider:




From Paciorek & Schervish, 2006



The idea of kernel convolution can be extended to nonstationary models.

Higdon (1998) and Higdon et al. (1999) proposed allowing the parameters of the kernel to vary in space. In the discrete version, we have

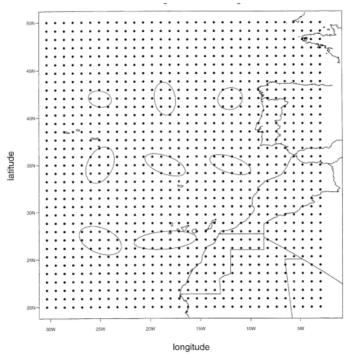
$$\eta(s) = \sum_{j=1}^{p} k(s - r_j; \theta(r_j))\alpha_j$$

Now the question arises of how we estimate the $\theta(r_j)$. Note that without constraining them in some way, we have dramatically increased the dimension of the parameter space.



Higdon (1998) used a 2D Gaussian kernel. This can be parameterized in terms of the lengths of the major and minor axes of the one standard deviation ellipse, and a rotation angle.

Higdon estimated the parameters by taking 8 locations, finding data within a given radius, and using the variogram just for that data. He then took the kernels (centered at the dots) to be weighted averages of those 8 kernels.



An alternative would be to specify the way the parameters are changing in space and then estimate this, say in a Bayesian hierarchical framework.

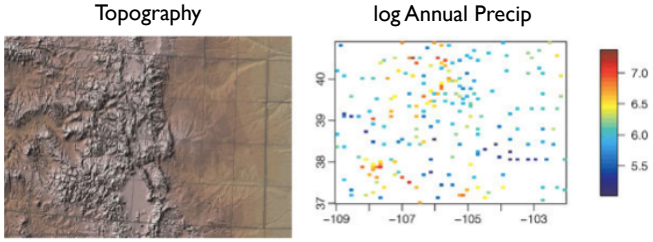
This is what Higdon et al. (1999) do. In fact they take the parameters of the kernel and treat them as coming from another (stationary) GP. The terms in the model are

| | |
|---|---|
| $Y\|\eta, \sigma^2_\epsilon$ | Data level: measurement error |
| $\eta\|\alpha, \theta$ | Process level: kernel representation |
| $\alpha\|\sigma^2_\alpha$ | Process level: generating random vector |
| $\theta\|\psi$ | Prior level: Smooth GPs |
| $\sigma^2_\epsilon, \sigma^2_\alpha, \psi$ | Prior level: hyper-priors |



A somewhat simpler but related approach would be to specify a parametric form for how $\theta$ varies. That is, write $\theta(s) = f(s; \psi)$ for a lower dimensional parameter $\psi$ and then estimate this parameter.

What might you consider for the rainfall dataset?




From Paciorek & Schervish, 2006



Returning to the matrix form of this model,

$$Y = X\beta + H\alpha + \epsilon$$
$$\alpha \sim MVN(0, \sigma^2_\alpha I_p) \qquad \epsilon \sim MVN(0, \sigma^2_\epsilon I_n)$$

this is the form of a linear mixed effects model. This connection between smoothing (including spatially) and mixed effects models was pointed out by Matt Wand.

For fixed $X$ and $H$, the model can be fit using standard software. In R, the function is called lme. The default estimation method is restricted maximum likelihood (REML).



Asymptotic properties of estimation and prediction

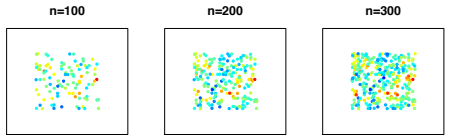
Usually in statistics, when we write $n \to \infty$, we mean that we are taking an increasing number of iid (independent and identically distributed) random variables.

In spatial statistics, the data are not iid, and we need to specify a sequence of sampling locations $S_1, S_2, \ldots$ where $S_n = \{s_1, \ldots, s_n\}$. How will we specify these?

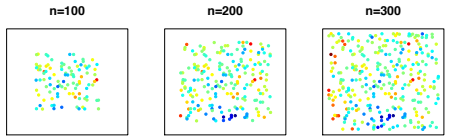
(Note: If we have independent replications in time at the same set of spatial locations, the usual framework applies.)

We can consider two basic schemes:

- Fixed domain or "infill": Increasingly dense set of locations in a bounded domain




- Increasing domain: Minimum distance is bounded away from zero






Neither framework is inherently the "correct" one to use. However, what we can prove differs between the two frameworks.

Increasing domain asymptotics is perhaps the natural one to use when proving properties of estimators, since increasing the domain allows us to get more and more (nearly) independent observations.

Mardia and Marshall (1984) gave conditions for the MLE to be consistent (converging in probability to the correct value) and asymptotically normal. The increasing domain asymptotics falls under these conditions.



However, fixed domain asymptotics seems more natural when considering prediction, since we are getting observations close to any point where we want to predict.

One might also argue that in many real world examples, fixed domain asymptotics is actually a better representation for how more samples could be taken. For example, over time the U.S. has increased the density of its weather monitoring stations.

Under fixed domain asymptotics, strong results are available for prediction, but less is possible for estimation. More on this shortly.



Practically, I think the usefulness of asymptotics is to give us insight into the behavior of estimators and predictors when the sample size is finite but "sufficiently large."

What "sufficiently large" means is something I consider relative to the degree of correlation in the process.

A highly correlated set of observations may be close to "asymptopia" under the fixed domain asymptotics, but not under the increasing domain asymptotics. For this reason I often find fixed domain results more interesting.

Two main threads appear in the literature on fixed domain asymptotics:

- The properties of estimators of spatial covariance parameters, e.g. the MLE
- The properties of predictions under a model that is not the true model, but is fixed as $n \to \infty$.

So far, there is very little to address what we actually do in practice, which is to do both estimation and prediction using the same dataset.

Both sets of results can make use of the idea of *equivalence and orthogonality* of Gaussian distributions.



Consider two distributions, denoted $P_0$ and $P_1$, for a random quantity $X$. $X$ may be finite-dimensional (e.g., a random variable or random vector), or it may be infinite-dimensional (e.g., a stochastic process).

Now suppose we observe a realization of $X$. $P_0$ and $P_1$ are *equivalent*, written $P_0 \equiv P_1$, if we are unable to determine which one generated $X$.

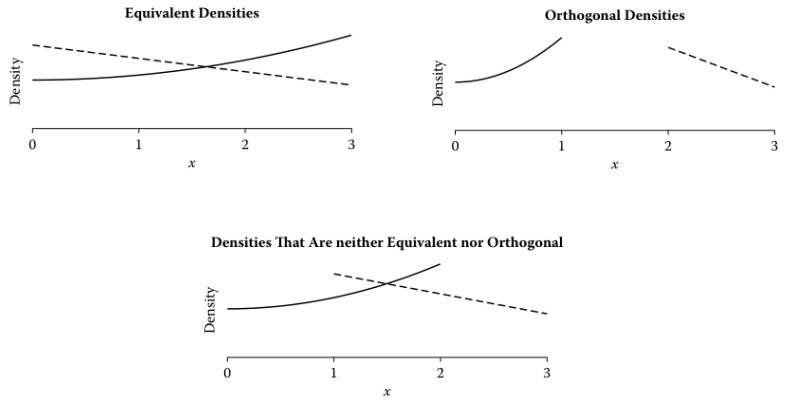
Formally, $P_0 \equiv P_1$ means that for any event $A$,

$$P_0(A) = 0 \Leftrightarrow P_0(A) = 0$$

The distributions are *orthogonal* if after observing $X$ we can determine with certainty which one is correct.



Example: observe scalar random variable $X = x$.






In great generality, Gaussian distributions are either equivalent or orthogonal (Kuo, 1975).

Example: Recall the Matern covariance function

$$C(d) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(d/\rho)^\nu \mathcal{K}_\nu(d/\rho)$$

For any $\nu > 0$, let $P_0 = G(0, C(\sigma_0^2, \rho_0, \nu))$ and $P_1 = G(0, C(\sigma_1^2, \rho_1, \nu))$.

Zhang (2004) showed that $P_0 \equiv P_1$ on any bounded domain if and only if

$$\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$$

Equivalence has important implications for estimation. For example, an immediate consequence of the last result is that $\sigma^2$ and $\rho$ are not consistently estimable under fixed domain asymptotics. That is, there are no estimators satisfying

$$\hat{\sigma}_n^2 \xrightarrow{P} \sigma_0^2 \quad \text{and} \quad \hat{\rho}_n \xrightarrow{P} \rho_0$$

However, Zhang (2004) showed that if you fix $\rho = \rho^*$ and find the corresponding MLE $\hat{\sigma}_n^2(\rho^*)$, then

$$\hat{\sigma}_n^2(\rho^*)/\rho^{*2\nu} \xrightarrow{P} \sigma_0^2/\rho_0^{2\nu}$$



The main results for prediction (see Stein, 1999) consider having a true GP distribution $G(0, K_0)$ but using a misspecified distribution $G(0, K_1)$ to make predictions. If $G(0, K_0) \equiv G(0, K_1)$, we have

- Asymptotic efficiency:

$$\frac{Var_0[\hat{Y}_1(s)]}{Var_0[\hat{Y}_0(s)]} \to 1$$

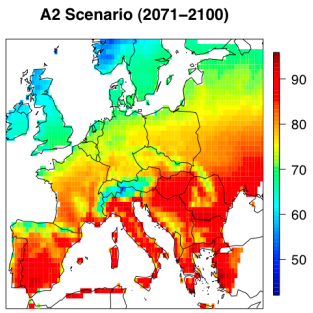
- Asymptotically correct variance estimation:

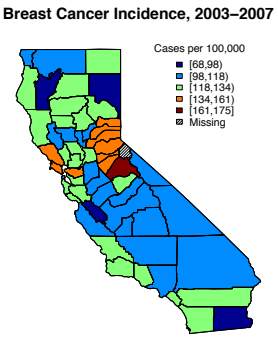
$$\frac{Var_1[\hat{Y}_1(s)]}{Var_0[\hat{Y}_1(s)]} \to 1$$



*Areal or lattice data* consist of observations that are associated with spatial regions.

The set of locations may be regular (e.g. a rectangular grid) or irregular (e.g. geographic units).





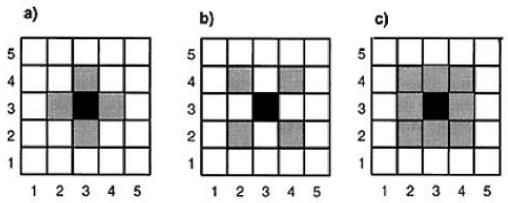

Areal data can often be thought of a "coarser-resolution" version of other data types, such as

- averaging a geostatistical field, rather than observing it at a point

- summing up observations from a point process.

In geostatistical data, the notion of distance was key. For areal data, we will sometimes still use distance, but more often we specify spatial dependence in terms of a more general notion of spatial neighborhoods or spatial proximity.

For data on a regular lattice, it's fairly straightforward to think about what a neighborhood means. You'll often see terminology referring to chess moves.




From Schabenberger & Gotway, 2005

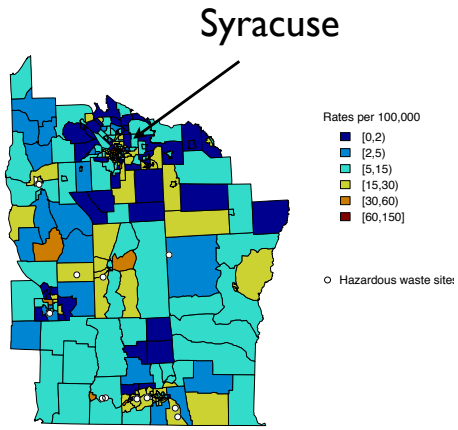
For irregular lattices, there are many ways to define the neighborhood structure.



Example dataset I: Leukemia cases from 1978-1982 in eight-county region of upstate NY, from Waller et al. (1992, 1994).

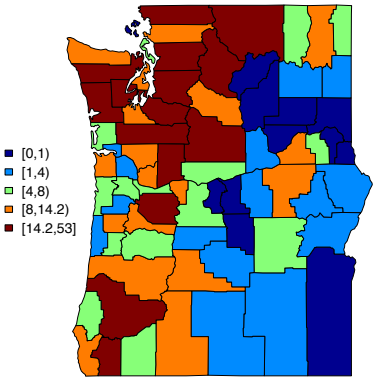
The map shows observed *rates* per 100,000 people per year.

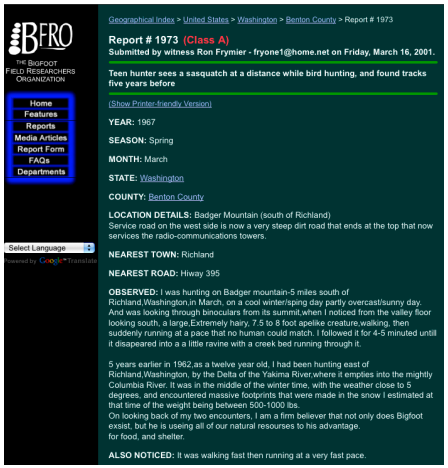
We also have demographic information for each census tract, as well as the locations of some (inactive) hazardous waste sites.






Example dataset II: Bigfoot sightings reports from bfro.net (no date range specified)






These datasets may be efficiently stored, manipulated, and visualized using the SpatialPolygonsDataFrame class in R, from the spdep package.

We need to be careful about what an areal unit is: many are simply polygons, but we may also have
- multiple polygons (e.g. islands)
- polygons with holes (e.g. lakes)

For this reason, there are different classes
- Polygon - single shape
- Polygons - collection of Polygon objects with extra information
- SpatialPolygons - collection of Polygons objects, similar in overall structure to SpatialPoints

We will also consider the *spatial proximity or weighting matrix*, W.

This is an nxn matrix whose entries are numbers from 0 to 1. It reflects

- Proximity: is j a neighbor of i?
- Strength: what is the strength of influence of j on i?

Note: these relationships are often defined in a way that is symmetric, but not always. For example, a large urban county may have more influence on a nearby rural one than vice versa.

By convention, $w_{ij} = 0$.



Testing for spatial association: Moran's I

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{ij} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

Under $H_0 : Y_i \, iid, \quad \frac{I + 1/(n-1)}{\sqrt{Var(I)}} \xrightarrow{D} N(0,1)$

Alternatively, the p-value may be found using Monte Carlo sampling, permuting the indices.



A few important caveats:

The results of the test can depend heavily on the form of W. Perhaps the chosen weights do not reflect the true scales of interaction between entities.

The observations may be non-iid in ways that are not due to spatial correlation. For example, perhaps there is an underlying spatial trend (mean) we have neglected. Or perhaps the observations have different variances.

There is a lack of consensus about how areas with no neighbors should be treated.



Spatial autoregressive models

We now consider a class of models, analogous to autoregressive models in time-series, for spatial dependence in areal data. The main context will be that we are fitting a regression model like

$$Y_i = x_i^T \beta + \epsilon_i,$$

where $i$ indexes spatial region. It's often useful to do some exploratory analysis using ordinary least squares (OLS), then look at maps of the residuals.

Working example: NY Leukemia data

For each census tract, we have a count of cases, population, percentage of population over age 65, percentage of population owning a home, and average inverse distance to a hazardous waste site.

For now we will only consider normal outcome variables. Later we will build hierarchical models and relax this assumption. The following transformation gives an approximately normal distribution.

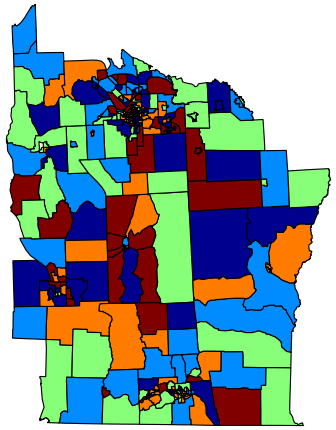
$$Z_i = log\left(\frac{1000(C_i+1)}{n_i}\right)$$



Visually, the residuals appear to have some spatial dependence.

A test using Moran's I rejects the null hypothesis of independence.

We will consider two classes of models for building dependence into such models: *simultaneous autoregressive (SAR)* and *conditional autoregressive (CAR)* models.





Simultaneous autoregressive (SAR) models date back to Whittle (1954). We build an autoregressive models for the residuals:

$$\epsilon_i = \sum_{j=1}^{N} b_{ij}\epsilon_j + \nu_i$$

indep Gaussian

$$\text{So } Y_i = x_i^T\beta + \sum_{j=1}^{N} b_{ij}\epsilon_j + \nu_i$$

$$= x_i^T\beta + \sum_{j=1}^{N} b_{ij}(Y_j - x_j^T\beta) + \nu_i$$

or in matrix notation, $(I-B)(Y-X\beta) = \nu$



If $(I-B)$ is invertible, then $Y - X\beta = (I-B)^{-1}\nu$ and so

$$\Sigma_Y = Var(Y) = (I-B)^{-1}\Sigma_\nu(I-B^T)^{-1}$$

How do we specify the matrix $B$? One common choice is to take $\alpha W$, where $W$ is a spatial proximity matrix. If $\lambda_{max}$ and $\lambda_{min}$ are the largest and smallest eigen-values of $W$ and $\lambda_{min} < 0$ and $\lambda_{max} > 0$, then constraining

$$1/\lambda_{min} < \alpha < 1/\lambda_{max}$$

guarantees $(I-B)$ is invertible. If we use $\tilde{W}$ instead, with $\tilde{W}_{ij} = W_{ij}/\sum_j W_{ij}$, $\lambda_{max} = 1$ and $\lambda_{min} \leq -1$.

The parameters in a SAR model can be estimated using maximum likelihood. The calculations are very similar to what we saw for point-referenced data; see e.g. page 50, where instead of $K(\phi)$ we have $K(\alpha)$, where

$$\Sigma_Y = \sigma^2(I - \alpha W)^{-1}V_\nu(I - \alpha W^T)^{-1} = \sigma^2 K(\alpha)$$

and $V_\nu$ is a diagonal matrix, either the identity if the variances of $\nu$ are equal, or something like $V_{\nu,ii} = 1/n_i$. We could also replace $W$ with $\tilde{W}$.

The function spautolm in the R package spdep fits this model using maximum likelihood.



Conditional autoregressive (CAR) models work with the *full conditional distributions*

$$p(Y_i|Y_j, j \neq i).$$

Although the full conditionals can always be derived from the joint distribution, a set of full conditionals does not necessarily define a valid joint distribution.

The Hammersley-Clifford theorem (first proved by Besag, 1974) gives the necessary conditions on the full conditionals. A set of full conditional distributions satisfying the conditions defines a *Markov random field*.



We will not discuss the conditions but instead consider a class of models satisfying them, taking everything to be Gaussian and with

$$E[Y_i|Y_j, j \neq i] = x_i^T\beta + \sum_{j=1}^{N} c_{ij}[Y_j - x_i^T\beta]$$
$$Var[Y_i|Y_j, j \neq i] = \tau_i^2$$

In spatial models, $c_{ij}$ is typically nonzero only if $s_j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the neighborhood of region $s_i$.

But how do we choose the $c_{ij}$ and $\tau_i^2$?



A result called Brook's Lemma can be used to show that the joint distribution is proportional to

$$\exp\{-\frac{1}{2}(Y - X\beta)^T[\Sigma_c^{-1}(I - C)](Y - X\beta\}$$

where $\Sigma_c = diag\{\tau_1^2, \ldots, \tau_n^2\}$.

This is the kernel of a multivariate normal distribution, but only if $\Sigma_c^{-1}(I - C)$ is symmetric, positive definite, and invertible.

Symmetry is imposed by $\frac{c_{ij}}{\tau_i^2} = \frac{c_{ji}}{\tau_j^2} \quad \forall i, j$

A specification that satisfies this condition if $W$ is symmetric is to take

$$c_{ij} = \frac{w_{ij}}{w_{i+}} \qquad \tau_i^2 = \frac{\tau^2}{w_{i+}}$$

so that $Y_i|y_j, j \neq i \sim N(x_i^T\beta + \sum_j \frac{w_{ij}}{w_{i+}}(y_j - x_j^T\beta), \frac{\tau^2}{w_{i+}})$

and Brooks Lemma gives the joint distribution proportional to

$$\exp\{-\frac{1}{2\tau^2}(Y - X\beta)(D_w - W)(Y - X\beta)\}$$

$$D_w = diag\{w_{1+}, ..., w_{n+}\}$$



However, we still have a problem: $(D_w - W)$ is not invertible. Actually, with some rearranging, the joint distribution can be expressed as

$$\exp\{-\frac{1}{2\tau^2}\sum_{i\neq j} w_{ij}((Y_i - x_i\beta) - (Y_j - x_j\beta))^2\}$$

and we can see that this doesn't change if we add a constant to all the $Y_i$.

This is called an *intrinsic autoregressive model.* The fact that it isn't proper isn't necessarily a problem if we use it as a prior distribution for some spatial latent effects, but we need to modify something if we want to use it as a model for the data.



If instead we take $\Sigma_Y^{-1} = \frac{1}{\tau^2}[D_w - \rho W]$, we can impose conditions on $\rho$ so that the matrix is invertible.

Specifically, let $\lambda_{max}$ and $\lambda_{min}$ be the largest and smallest eigen-values of

$$D_w^{-1/2} W D_w^{-1/2}.$$

Then $1/\lambda_{min} < \rho < 1/\lambda_{max}$ produces a proper joint distribution. The full conditionals are

$$Y_i|y_j, j \neq i \sim N(x_i^T\beta + \rho\sum_j \frac{w_{ij}}{w_{i+}}(y_j - x_j^T\beta), \frac{\tau^2}{w_{i+}})$$



When the joint distribution is proper, we can use properties of the multivariate normal distribution to interpret the entries of $\Sigma_Y^{-1} = \frac{1}{\tau^2}[D_w - \rho W]$.

- The diagonal entries:

$$1/(\Sigma_Y^{-1})_{ii} = Var(Y_i|Y_j, j \neq i) = \tau^2/w_{i+}$$

- The off-diagonal entries

$(\Sigma_Y^{-1})_{ij} = 0 \Rightarrow Y_i, Y_j$ cond indep given $Y_k, k \neq i, j$

Since this happens when $w_{ij} = 0$, by choosing a particular neighborhood structure, we are actually imposing a set of conditional independence relationships.

CAR models can also be fit using maximum likelihood.

Note the different formulations we have for the joint covariance matrix.

SAR: $\Sigma_Y = \sigma^2[I - \alpha W]^{-1}[I - \alpha W^T]^{-1}$

CAR: $\Sigma_Y = \tau^2[D_w - \rho W]^{-1}$

Any SAR model can also be expressed as a CAR model, but a CAR model cannot necessarily be expressed as a SAR model without changing the neighborhood structure (Cressie, 2005).



Waller and Gotway (2004) suggest incorporating weights into a CAR model using the model

$$\Sigma_{CAR} = \sigma^2(I - \rho W)^{-1}V_c$$

where $V_c = diag\{1/n_i\}$. However, this matrix is only symmetric if we take $n_i = w_{i+}$. Indeed, debugging the spautolm function using a weights argument, the covariance matrix being calculated is not symmetric. So I don't recommend this option.

More often, researchers use a CAR model within the context of a hierarchical model, in which unequal variances are more naturally accommodated.



For our disease mapping example, some alternative hierarchical models would be

$$Z_i|\eta \overset{indep}{\sim} N(x_i^T\beta + \eta_i, \tau^2/n_i)$$
$$\eta \sim MVN(0, \Sigma_{CAR}(\tau^2, \rho))$$

or, going back to our original count data,

$$Y_i|\lambda \overset{indep}{\sim} Pois(n_i\lambda_i)$$
$$\lambda_i = \exp\{x_i^T\beta + \eta_i\}$$
$$\eta \sim MVN(0, \Sigma_{CAR}(\tau^2, \rho))$$



Using a CAR model in a hierarchical context, we have the choice of whether to keep the "propriety parameter" $\rho$ in the model, or use the (improper) intrinsic autoregressive model.

If we keep $\rho$, we need to be careful in specifying a prior distribution for it. Only values of $\rho$ close to the upper bound for propriety produce a reasonable amount of spatial correlation.

However, if we remove $\rho$, the posterior is proper, but not all parameters in the model are identifiable. Although this is not theoretically a problem, in practice MCMC algorithms can have numerical issues.

SAR and CAR models on lattices

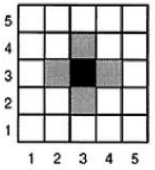
To further examine the properties of these models in a simplified setting, we'll consider the special case that the spatial areas are defined by a regular lattice, i.e. a two-dimensional array in which the spacings are equal.

When these models were introduced (Whittle, 1954; Besag, 1974), they were proposed for *infinite* lattices, meaning every location has the same neighborhood structure.

In practice, of course, we never have data on such objects.

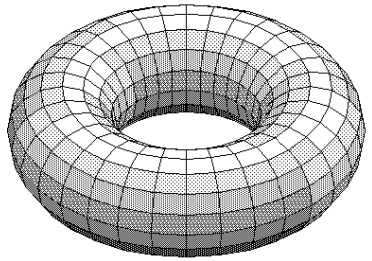


If we have a finite 2D lattice and adopt, say, the "rook" definition of neighbors, then points on the boundary of the array will have fewer neighbors.




To construct stationary processes for the purpose of study, we can "wrap" the 2D array to create a *torus*.

Now every point has the same number of neighbors.





One advantage of using CAR models, from a computational perspective, is that when each point has a relatively small number of neighbors, the precision matrix $\Sigma^{-1}$ is *sparse*, meaning it contains many zeroes. Sparse matrix algorithms can be used to manipulate it efficiently.

This advantage is even more striking for regular lattices with circular boundary conditions. In 1D, the corresponding precision matrix is *circulant*, and in 2D it is *block circulant*. In these cases the discrete Fourier transform can be used to calculate the covariance matrix.



The CAR models we've been discussing are special cases of *Gaussian Markov Random Fields (GMRFs)*.

A random vector $x$ is called a GMRF with respect to a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\mu$ and symmetric positive definite precision matrix $Q$ iff its density has the form

$$p(x) = (2\pi)^{-n/2}|Q|^{1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T Q(x-\mu)\right\}$$

and $$Q_{ij} \neq 0 \iff \{i,j\} \in \mathcal{E}\, \forall i \neq j$$

Recall the interpretation of $Q$ when $x$ is Gaussian:

$$x_i \perp x_j | x_{-\{i,j\}} \iff Q_{i,j} = 0$$

Aside: Reminder about conditional independence

$x \perp y | z$ means that $p(x, y|z) = p(x|z)p(y|z)$

This is equivalent to a definition based on the factorization criterion

$p(x, y, z) = f(x, z)g(y, z)$ for some functions $f$ and $g$.



We can read off conditional independence relationships from the graph. These are equivalent:

Pairwise Markov Property:

$x_i \perp x_j | x_{-\{i,j\}}$ if $\{i, j\} \notin \mathcal{E}$ and $i \neq j$

Local Markov Property:

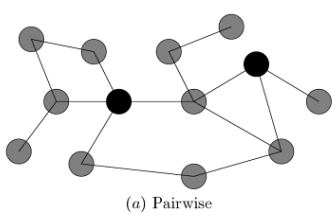
$x_i \perp x_{-\{i,\delta(i)\}} | x_{\delta(i)}$ for every $i \in \mathcal{V}$

Global Markov Property:

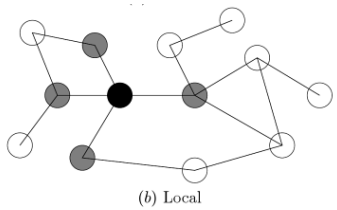
$x_A \perp x_B | x_C$ for disjoint sets $A, B, C$ if $C$ separates $A$ from $B$



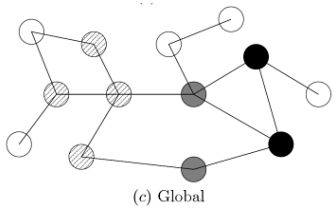
Example: What can we say in each case?



(a) Pairwise



(b) Local



(c) Global

Taken from Rue & Held, 2005

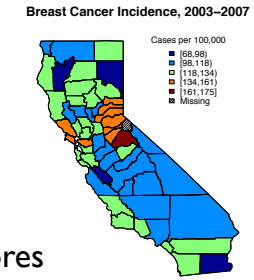


Using GMRFs in hierarchical models

Consider the following *disease mapping* problem. We observe disease counts $y_i$ in each of $I$ spatial regions. For each region, we also have a population at risk $n_i$.

If each person is at equal risk, an estimate of the expected number of counts for region $i$ is

$$E_i = n_i \frac{\sum_{i=1}^{I} y_i}{\sum_{i=1}^{I} n_i}$$




This estimate is kind of extreme: it ignores any differences between regions.

At the other extreme, we might use only the data for each region to estimate its risk, i.e. $y_i/n_i$.

However, these estimates are likely to be quite variable, due to small $n_i$. A model-based approach lets us "borrow strength" across regions to come up with less variable estimates. This results in a smoothing of extreme rates toward the mean.

The models can be cast in a Generalized Linear Mixed Model (GLMM) framework, similar to what we saw for geostatistical data (e.g. the scallop data).



For counts, we could use a Binomial (logit link) or Poisson (log link) model. The Poisson is appropriate as an approximation to the Binomial when the probability of an event (disease case) is small.

$$Y_i|\eta \overset{indep}{\sim} Poisson(n_i \exp\{\eta(s_i)\})$$

The value $n_i$ is an *offset* that reflects a different population at risk in each region. It's also possible to use $E_i$ as an offset, but this puts Y on the RHS.



Some flavors of this...

$\eta(s_i) = x_i^T\beta$ Fixed effects

$\eta(s_i) = x_i^T\beta + z_i$
$z \sim MVN(0, \sigma_z^2 I)$ + Exchangeable random effects

$\eta(s_i) = x_i^T\beta + u_i$
$u \sim MVN(0, Q^{-1})$ + Spatial (GMRF) random effects

$\eta(s_i) = x_i^T\beta + z_i + u_i$ + Both ("convolution prior")



$$Y_i|\eta \overset{indep}{\sim} Poisson(n_i \exp\{\eta(s_i)\})$$

You may wonder why we fall back on a latent GMRF to induce spatial correlation. Basically, it is because specifying MRFs for non-Gaussian variables is difficult to do in a flexible way.

For example, auto-Poisson models are possible, but the conditions of the Hammersley-Clifford theorem force them to have negative correlation between neighbors (Besag, 1974).

Constructing MCMC algorithms

Recall that for a GMRF,

$$x_i \perp x_j | x_{-\{i,j\}} \iff Q_{i,j} = 0$$

The sparsity of $Q$ can be used to construct efficient MCMC sampling algorithms.

To simplify the discussion, let's first consider the more abstract task of sampling from a GMRF, i.e.

sample $x \sim MVN(\mu, Q^{-1})$ where $Q$ is the precision matrix.



If we have specified the distribution for $x$ through its full conditionals, one option is to use the Gibbs sampler. However, this algorithm is not exact; it only converges to the correct distribution.

A better method is to sample the elements of $x$ all at once. This is called "blocking" $x$.

Algorithm:
1. Compute $L$, where $Q = LL^T$.
2. Sample $Z \sim N(0, I)$.
3. Solve $L^T v = Z$ for $v$.
4. Compute $x = \mu + v$.

Why does this work?



A simple GMRF example: AR(1) process

$X_1 \sim N(0, \sigma^2/(1-\phi^2))$

$X_t | X_1, \ldots, X_{t-1} \sim N(\phi x_{t-1}, \sigma^2)$

What are the full conditionals for the X's?
Sample $X_1, \ldots, X_T$ using a Gibbs sampler.

What is the precision matrix?
Sample the X's as a block.



Another task that often comes up in the context of Bayesian models is sampling

$$x \sim MVN(Q^{-1}a, Q^{-1})$$

For example, if

$$Y|\eta \sim MVN(\eta, \sigma^2 I) \qquad \eta \sim MVN(0, \tau^2 Q^{-1})$$

then $\eta | Y \sim MVN(Q^{*-1}a, Q^{*-1})$ where

$$Q^* = \frac{1}{\sigma^2} I + \frac{1}{\tau^2} Q \qquad a = \frac{1}{\sigma^2}(Y - \eta)$$

Note this is just another GMRF!

Here's an algorithm for this case:

Sampling $x \sim MVN(Q^{-1}a, Q^{-1})$

1. Compute $L$, where $Q = LL^T$.
2. Solve $Lv = a$ for $v$.
3. Sample $Z \sim N(0, I)$.
4. Solve $L^T x = v + z$ for $x$.



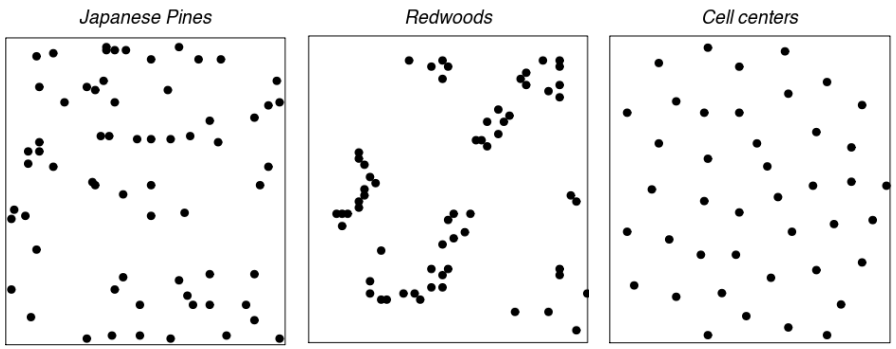
*Spatial point patterns* consist of observations that are the *locations* of events. We often use *spatial point process* models to describe them. These are stochastic processes that generate a countable set of events.

The locations may be associated with additional variables, known as a *marked point process*.

A null model for spatial point processes is *complete spatial randomness (CSR),* an idea we will make more precise shortly. Processes that don't follow CSR may have features of *clustering* or *regularity*.



Examples




CSR? Clustering Regularity



The locations of events are not always completely mapped. Two "sparse" sampling methods are

• *Quadrat counts*: count the number of events in a number of smaller, pre-defined areas

• *Distance based sampling*: choose an event, then record the distances to the first few nearest events

The methods for sparse sampling are different than for complete mapping, and the objectives are often more modest. When the areas in quadrat sampling completely partition in the region of interest, methods for areal data are appropriate.

A fundamental point process model is the *(spatial) homogeneous Poisson process*. Let $N(A)$ denote the number of events in a region $A$ and $|A|$ its area. There are several equivalent definitions; here is one:

1) For some $\lambda > 0$ and any finite region $A$,

$$N(A) \sim Poisson(\lambda|A|)$$

2) Given $N(A) = n$, the $n$ events form an iid sample from the uniform distribution on $A$.

This is what is meant by complete spatial randomness. The parameter $\lambda$ is called the *rate* of the process.



1) and 2) imply another condition, which is

3) For any two disjoint regions $A$ and $B$, the random variables $N(A)$ and $N(B)$ are independent.

Suppose we observe $n$ events in region $A$. Under CSR, the estimator $\hat{\lambda} = n/|A|$ is unbiased.

Simulating a homogeneous Poisson process can be done either unconditionally (draw $n \sim Pois(\lambda|A|)$ and then sample $n$ points uniformly on $A$) or conditionally on $n$ (skip the first step).



Testing for CSR is often the first step in analyzing a spatial point pattern. It's often not a very interesting hypothesis, but it's useful because

• If CSR is not rejected, there is not much reason to build a more complicated model.

• Many tests are associated with plots that are useful for diagnosing the nature of a deviation from CSR.

• CSR as a hypothesis acts as a dividing line between those that are "clustered" vs. those that are "regular".



Many of the tests we'll examine are based on Monte Carlo sampling. Let $u_1$ be the observed value of a statistic $U$. The sampling distribution of $U$ under the null hypothesis may be difficult to derive.

Let $u_2, \ldots, u_s$ be a sample of test statistics generated by independent random sampling under a *simple* null hypothesis. Then a test that rejects when $u_1$ ranks $k^{th}$ or lower or $(s-k+1)^{th}$ or higher has size $\alpha = 2k/s$.

This is because under the null hypothesis,

$$P(u_1 \text{ has rank } j) = 1/s$$

for $j = 1, \ldots, s$.

Aside on the empirical distribution function

These methods will also make use of the *empirical distribution function (ECDF)*, an estimate of the CDF.

Suppose $X_1, \ldots, X_n$ are $iid$ with CDF $F$. The ECDF

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} I\{X_i \leq x\}}{n} = \frac{\#\{X_i \leq x\}}{n}$$

is an unbiased estimator of

$$F(x) = P(X_i \leq x) = \int_{-\infty}^{x} f(y)dy = \int_{-\infty}^{\infty} I\{y < x\} f(y)dy$$



Classical tests of CSR are based on derived distances.

- pairwise distances

$$s_{ij} = ||x_i - x_j||, \quad i \neq j$$

- nearest neighbor distances

$$t_i = \min_{j \neq i} s_{ij}, \quad i = 1, \ldots, n$$

- empty space distances

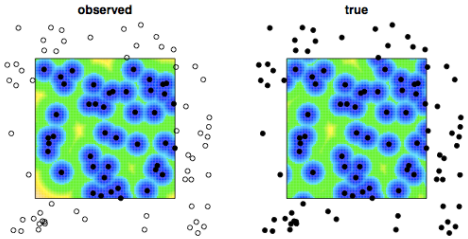
$$d(u) = \min_i ||u - x_i||$$



First consider the empty-space distances. There are two ways of thinking about their distribution.

One is to fix the observation window $A$. The distribution of $d(u)$ depends on $A$ and is hard to derive in closed form. Alternatively, we could imagine there is a process defined on $\Re^2$ and we only observe it on $A$, so the observed distances are biased relative to the "true" distances.

In other words, there is an "edge effect."






Putting this issue aside for a moment, consider the CDF

$$\begin{aligned} F_u(r) &= P(d(u) \leq r) \\ &= P(\text{at least one point within radius } r \text{ of } u) \\ &= 1 - P(\text{no points within radius } r \text{ of } u) \end{aligned}$$

For a homogeneous Poission process on $\Re^2$, this doesn't depend on $u$, and

$$F(r) = 1 - \exp\{-\lambda \pi r^2\}$$

Conditioning on $n$, we can replace $\lambda$ by $\hat{\lambda}$ and get a useful baseline for comparison.

Suppose we define a set of locations $u_1, \ldots, u_m$ for estimating $F(r)$ under an assumption of stationarity. The estimator

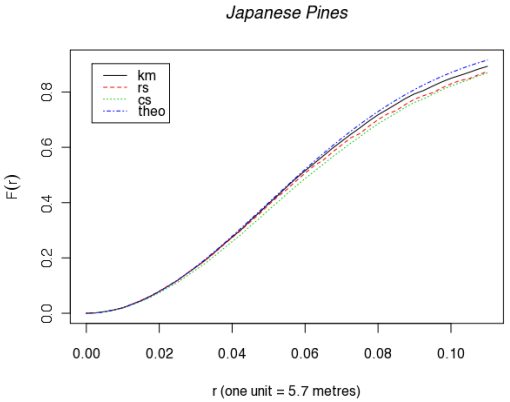
$$\hat{F}(r) = \frac{1}{m} \sum_{j=1}^{m} I\{d(u_j) \leq r\}$$

is biased due to the edge effects. There are a variety of corrections. For example, one simple one is when estimating for a given $r$ to not consider $u's$ within distance $r$ of the boundary. Another is to think of this as a censored data problem and use a spatial variant of the Kaplan-Meier correction.



Interpreting the estimate

By comparing $\hat{F}$ to our estimate under CSR, we can see whether points are more clustered or more regular.





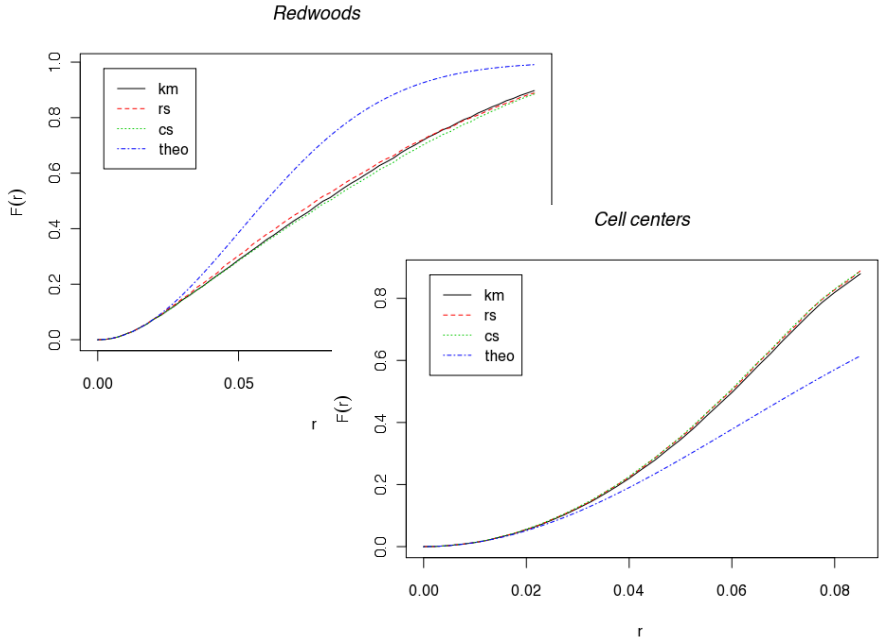
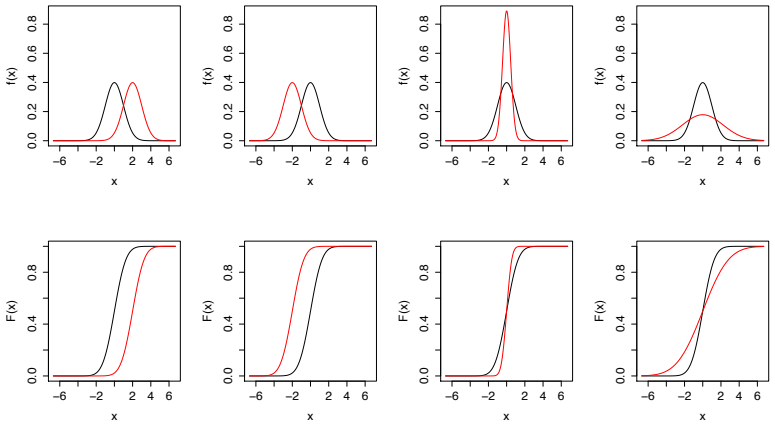






Illustration for interpreting changes in the CDF plots



We can use Monte Carlo sampling to construct "simulation envelopes" under CSR. Simply sample $n$ locations uniformly on $A$ and construct $\hat{F}$.

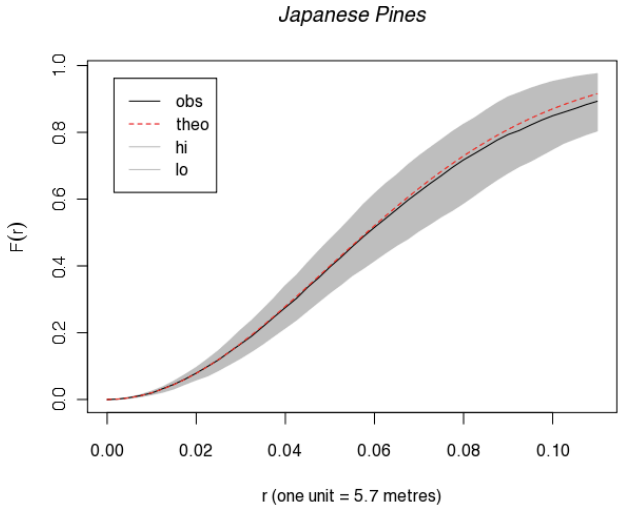
Do this $s-1$ times, then let $L_k(r)$ be the rank $k$ sample at $r$ and $U_k(r)$ the rank $k-s+1$ sample (including the original data in the $s$ samples).

$(L_k(r), U_k(r))$ can then be used to construct Monte Carlo tests for the CSR null hypothesis (coming up).

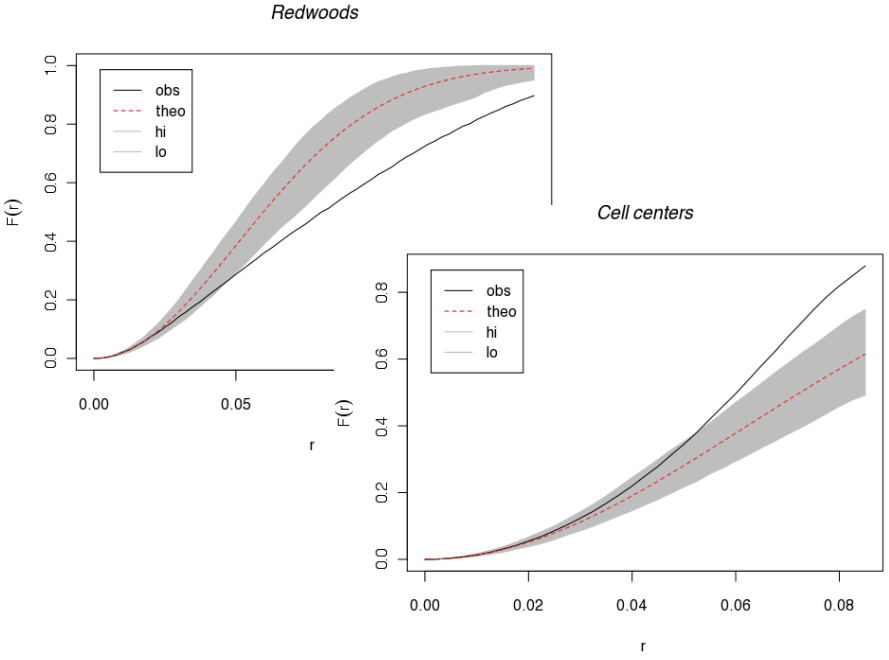
Note: they are not "confidence intervals" for the true $F$.



Text








Working instead with the nearest-neighbor distances from the events themselves, we can define

$$G(r) = P(t_i \leq r)$$

where $t_i = \min_{j \neq i} ||X_i - X_j||$ for an arbitrary point $X_i$.

We can estimate it by the ECDF of our observed $t_i's$.

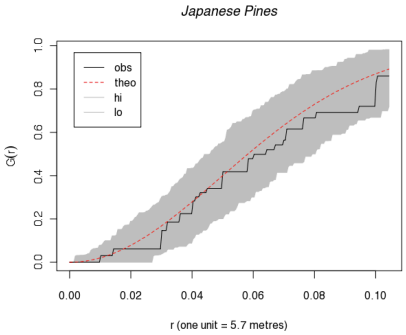
$$\hat{G}(r) = \frac{1}{n} \sum_{i=1}^{n} I\{t_i \leq r\}$$

As before, we may or may not make an edge correction.

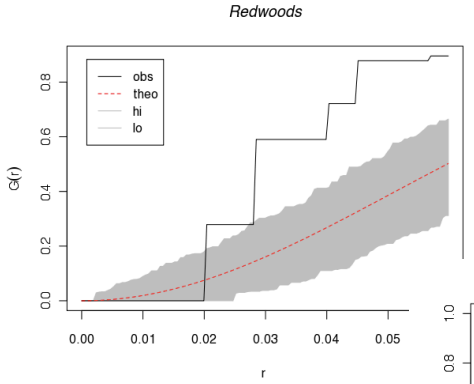
For a homogeneous Poisson process on $\Re^2$, the true value is again

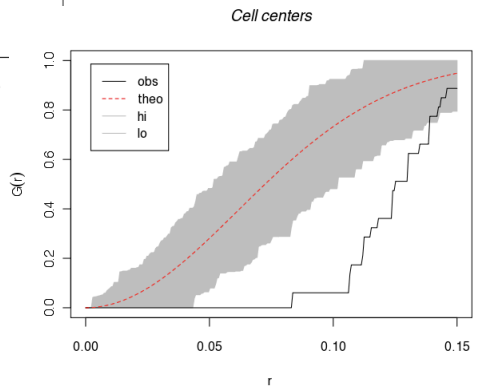
$$G(r) = 1 - \exp\{-\lambda\pi r^2\}$$

and we can again form Monte Carlo based simulation envelopes.








Returning to testing, we have many options for test statistics constructed from $\hat{F}$ and $\hat{G}$ (as well as many others).

For example, for any $r$, $\hat{F}(r)_1$ (from the observations) and $\hat{F}(r)_2, \ldots, \hat{F}(r)_s$ (from simulations under CSR), the test that rejects when the rank of $\hat{F}(r)_1$ is either $k$ or lower, or $s - k + 1$ or higher, has size $\alpha = 2k/s$.

This is equivalent to a test that rejects when $\hat{F}(r)_1$ falls outside of $(L_k(r), U_k(r))$.



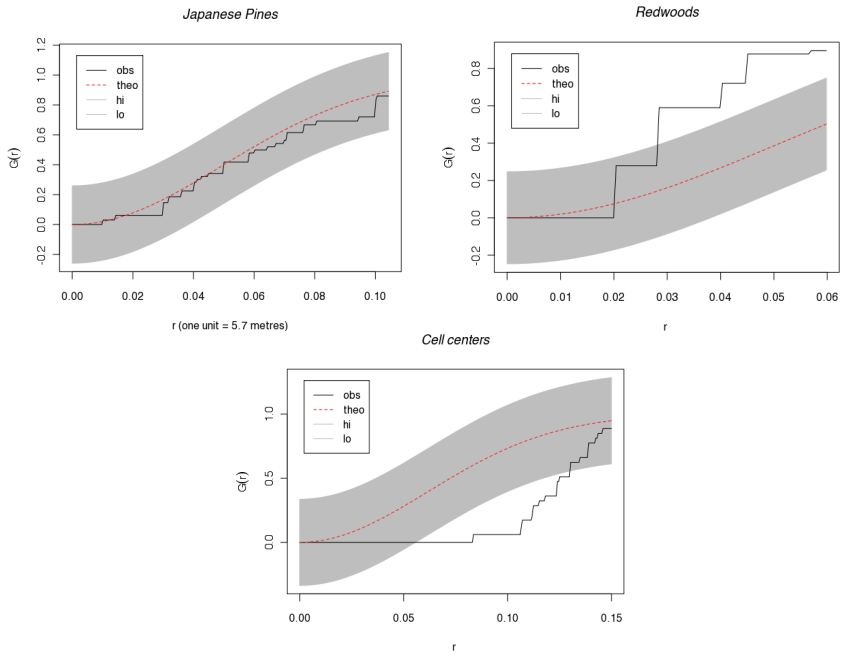
However, we are potentially doing a very large number of tests this way. A single "global" test can be constructed as follows: for each $i = 1, \ldots, s$, let

$$D_i = \max_r |\hat{H}_i(r) - H(r)|$$

where $H(r)$ is the theoretical value under CSR. Then define $D_{crit}$ to be the $k^{th}$ largest among the $D_i$. Then, define "global" envelopes

$$L(r) = H(r) - D_{crit}$$
$$U(r) = H(r) + D_{crit}$$

Then the test that rejects if $\hat{H}_1(r)$ ever wanders outside $(L(r), U(r))$ has size $\alpha = 1 - k/s$.






Suppose instead of a complete mapping, we only have *quadrat counts* $Y_1, \ldots, Y_n$ for subregions, each of area $D$.

We want to test for deviations from CSR and also estimate the overall *intensity* $\lambda_A = E[N(A)]/|A|$ for the study region $A$.

The key assumption is that $|A| >> nD$. If this were not true, it would have been feasible to do a complete mapping. So, it should also be the case that the subregions are spread out, far enough apart to be treated as approximately independent. (Of course, if CSR holds, the counts are independent as long as the subregions are disjoint.)



Define $\hat{\lambda}_A = (\sum_{i=1}^{n} Y_i)/(nD)$.

If the process is stationary, $\hat{\lambda}_A$ is unbiased, and its standard error is $\sqrt{\lambda/(nD)}$.

Consider that under CSR,

$$Y_1, \ldots, Y_n \overset{iid}{\sim} Poisson(\lambda D)$$

The test statistic $I = s^2/\bar{Y}$ is called the index of dispersion. Under CSR, $(n-1)I \sim \chi^2_{n-1}$ approximately. Large values of $I$ are indicative of clustering and small values of regularity.



There are a large number of tests available for the case that we sample a subset of *distances between events*, rather than mapping all the locations.

To get a flavor of the geometric arguments involved, consider T-square sampling (Besag and Gleaves, 1973).

Let O be an arbitrary point. Let P be the nearest event and X the distance between them. Then let Q be the nearest event to P with the restriction that the angle OPQ must be at least $\pi/2$. Let Z be the distance between P and Q.

Then $2\pi\lambda X^2$ and $\pi\lambda Z^2$ are iid $\chi^2_2$.

Let $(x_i, z_i)$, $i = 1, \ldots, m$ be $m$ such values. Typically the origins would be chosen on a grid. Then the test statistic

$$t_N = \frac{1}{m} \sum_{i=1}^{m} x_i^2/(x_i^2 + z_i^2/2)$$

is approximately $N(1, 1/(12m))$, with large values suggesting clustering and small values suggesting regularity.

The MLE for $\gamma = 1/\lambda$ under CSR is

$$\hat{\gamma} = \pi(\sum_i x_i^2 + \sum_i z_i^2/2)/(2m)$$



Point process models

A spatial point process is a stochastic process generating realizations of countable sets on the plane.

We'll start by defining first and second-order properties.

The *intensity function* $\lambda(x) = \lim_{|dx| \to 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\}$

For a stationary process, $\lambda$ is constant.



The *second order intensity function*

$$\lambda_2(x, y) = \lim_{|dx|, |dy| \to 0} \left\{ \frac{E[N(dx)N(dy)]}{|dx||dy|} \right\}$$

For a stationary process $\lambda_2(x, y) \equiv \lambda_2(x - y)$; for a stationary and isotropic process, $\lambda_2(x - y) \equiv \lambda_2(t)$, where $t = ||x - y||$.

The conditional intensity $\lambda_c(s|y) = \lambda_2(x, y)/\lambda(y)$



Another second-order summary is the *K function*

$$K(r) = \frac{1}{\lambda} E[N_o(r)]$$

where $N_o(r)$ denotes the number of events within distance $r$ of an arbitrary event.

We will come back to estimating $K$; it can be used in a similar manner to $F$ and $G$ in constructing tests of CSR.

We have already seen a homogeneous Poisson process. An *inhomogeneous Poisson process* is defined in terms of a spatially varying intensity function $\lambda(x)$, following

1) $N(A)$ has a Poisson distribution with mean

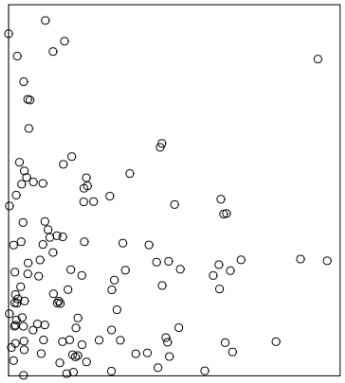
$$\mu(A) = \int_A \lambda(x)dx.$$

2) Given $N(A) = n$, the $n$ events in $A$ form an independent random sample from the distribution on $A$ with pdf $\lambda(x)/\mu(A)$.

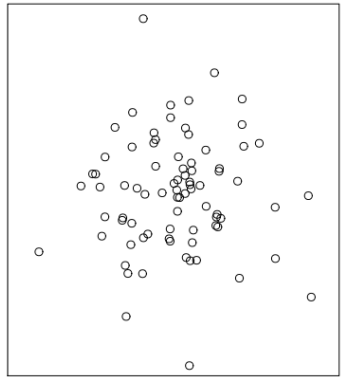


Two examples on $[0, 2]^2$

$\lambda(x) = 300 \exp\{-2x_1 - x_2\}$



$\lambda(x) = 300 \exp\{-||x - x_0||/0.2\}$
$x_0 = (1, 1)$





One way that an inhomogeneous Poisson process can arise is if we have a homogeneous Poisson process of rate $\lambda$ over a region $A$, and then the process is *thinned* in a non-constant way.

Specifically, suppose each event is deleted or not independently, with probability $0 \leq p(x_i) \leq 1$ of being kept. Then the result is a realization of an inhomogeneous Poisson process with

$$\lambda(x) = \lambda p(x).$$

This suggests a very simple algorithm for simulating an inhomogeneous Poisson process with intensity $\lambda(x)$....



1) Calculate $\lambda_{max} = \max_{x \in A} \lambda(x)$ (perhaps numerically)

2) Simulate from a homogeneous Poisson process with rate $\lambda_{max}$:

Simulate $n \sim Poisson(\lambda_{max}|A|)$

and $X_1, \ldots, X_n \overset{iid}{\sim} Unif(A)$

3) For $i = 1, \ldots, n$, retain $X_i$ with probability

$$p(x_i) = \lambda(x_i)/\lambda_{max}.$$

If we take the intensity function to be random as well, we have a "doubly stochastic" Poisson process, also known as a *Cox process*. That is,

let $\{\Lambda(x), x \in \Re^2\}$ be a non-negative stochastic process, and

conditional on $\{\Lambda(x) = \lambda(x), x \in \Re^2\}$ events are an inhomogeneous Poisson process with intensity $\lambda(x)$.

The distribution for $\Lambda$ completely determines the first and second order properties of the process. We can refer to it as a Cox process "driven by" $\Lambda$.



A few examples of Cox processes:

1) If $\Lambda(x)$ doesn't depend on $x$, we have a *mixed Poisson process*.

2) If $\Lambda(x) = \exp\{Z(x)\}$ and $Z$ is a Gaussian process, we call it a *log Gaussian Cox process*.

3) $\Lambda$ may be a *shot noise process* driven by unobserved homogeneous Poisson process $N_0$, with

$$\Lambda(x) = \gamma + \alpha \int_A K_h(x-u) dN_0(u)$$

for a kernel $K_h$. The case for which $K_h$ is a top hat function is called the Matern process.



As part of an exploratory data analysis, we can estimate the intensity function $\lambda(x)$. Because we often have only a single realization of the point process, what we tend to estimate is in fact a smoothed or spatially averaged version of $\lambda(x)$.

In *quadrat counting*, we divide the study area into subregions or quadrats, and we compute the number of points $n_j$ in each quadrat $B_j$.

$$E[n_j] = \int_{B_j} \lambda(x) dx$$

so the quadrat counts are unbiased estimates of the quantities on the RHS.



Another way to borrow strength locally is to use a kernel density estimator:

$$\hat{\lambda}(x) = e(x) \sum_{i=1}^{n} \kappa(x - x_i)$$

where $\kappa$ is a kernel and $e(x) = 1/\int_S \kappa(x-v)dv$

Then $\hat{\lambda}(x)$ is an unbiased estimator of

$$\lambda^*(x) = e(x) \int_S \kappa(x-v)\lambda(v)dv$$

The choice of $\kappa$ (particularly of its bandwidth) can be made to minimize an estimate of the mean squared error.

A *Poisson cluster process* (Neyman & Scott, 1958) is defined by

1) Parent events form a Poisson process with intensity $\lambda_c$.

2) Each parent produces a random number $M$ of offspring, iid for each parent according to discrete distribution $p_m$, $m = 0, 1, \ldots$

3) The positions of offspring relative to parents are iid according to bivariate pdf $f(x)$.

By convention, the point pattern consists of the offspring only, not the parents.



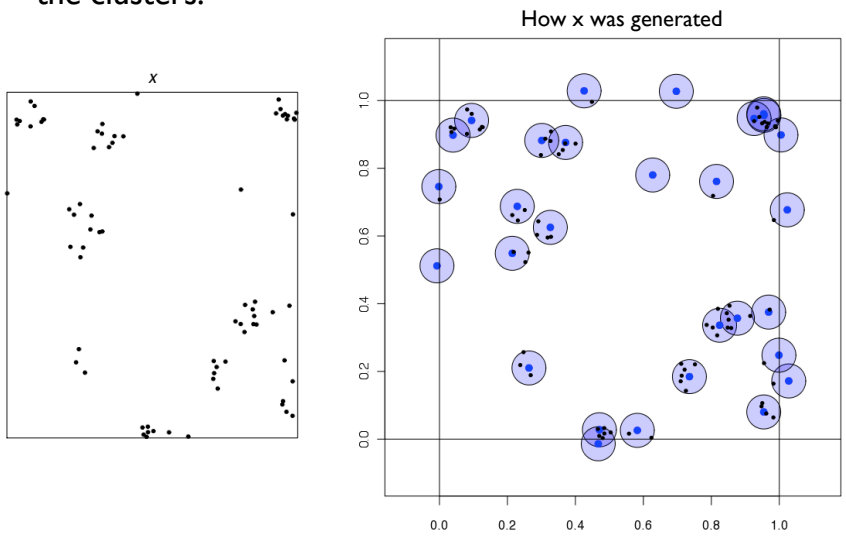
Two special cases (also special cases of shot noise)

1) Matern process
- Homogeneous Poisson parent process
- Poisson distributed number of children
- Children uniformly distributed on disc of radius $r$ centered at parent location

2) Thomas process
- Homogeneous Poisson parent process
- Poisson distributed number of children
- Locations of children according to isotropic bivariate normal distribution with variance $\sigma^2$

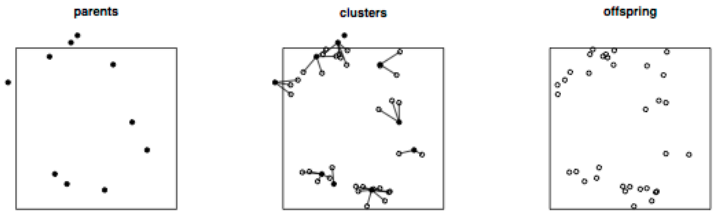


It can be difficult to look at a realization and identify the clusters.






When simulating a Poisson cluster process, we need to be careful of the boundary issue.




If the mass of $f$ is contained within radius $r$ of the origin, we can "dilate" the window by $r$ when we generate the parents. If this is not the case, we could still choose an $r$ to contain a large percentage of the mass.

Regular patterns can arise through the imposition of a minimum permissible distance $\delta$.

Processes that have no other departures from CSR are called *simple inhibition processes*. There are several, non-equivalent formulations.

They can be compared using using the *packing intensity* $\tau = \lambda\pi\delta^2/4$ , where $\lambda$ is the intensity. $\tau$ is the proportion of the plane covered by non-overlapping discs of diameter $\delta$. At most, it can be

$$\tau_{\max} = (\pi/\sqrt{3})/6 \approx 0.907$$

which is the packing intensity for an equilateral triangular lattice with spacing $\delta$.



Model 1 (Matern, 1960) A Poisson process of intensity $\rho$ is thinned by deletion of all pairs of events less than $\delta$ apart.

The probability an arbitrary event survives is $\exp(-\pi\rho\delta^2)$ so the intensity is $\lambda = \rho\exp(-\pi\rho\delta^2)$.

The corresponding packing intensity is at most $(4e)^{-1} \approx 0.092,$ which is only about 10% of $\tau_{\max}$.

$$\lambda_2(t) = \begin{cases} 0 & t < \delta \\ \rho^2\exp(-\rho U_\delta(t)) & t \geq \delta \end{cases}$$

where $U_\delta(t)$ is the area of the union of two discs, each of radius $\delta$ and with centers distance $t$ apart.



Model 2: Simple Sequential Inhibition

Define a sequence of $n$ events in $A$ with

1) $X_1 \sim Unif(A)$

2) Given $X_1, \ldots, X_{i-1},$ is uniformly distributed on

$$\{y \in A : ||y - x_j|| \geq \delta,\, j = 1, \ldots, i-1\}$$

The packing intensity is $\tau = n\pi\delta^2/(4A)$. The maximum attainable packing intensity is a random variable with an intractable distribution, but its mean has been estimated at about 0.547.



*Gibbs processes* are defined via their probability densities. The definition of the density for a point process $f(\{x_1, \ldots, x_n\})$ is complicated somewhat due to the fact that $n$ is random.

For a process $X$ defined on bounded region $S$,

1) The probability that $X$ contains exactly $n$ points is

$$p_n = \frac{e^-|S|}{n!}\int_S \cdots \int_S f(\{x_1, \ldots, x_n\})dx_1 \cdots dx_n$$

2) Conditionally on exactly $n$ points, the joint density of the locations is $f(\{x_1, \ldots, x_n\})/p_n$.

A *pairwise interaction process* has

$$f(\{x_1, \ldots, x_n\}) = \alpha \left[\prod_i b(x_i)\right] \left[\prod_{i<j} c(x_i, x_j)\right]$$

The term $\alpha$ is a normalizing constant. The function $c$ controls the pairwise interactions and must be symmetric. A number of restrictions can be made on $c$ to enforce a local dependence structure, leading to *Markov point processes*. (Markov processes can also be more general than only pairwise interactions.)

Rather than covering the general case, we'll consider two common examples.



The *hard core Gibbs process* takes $b(x) = \beta$ and

$$c(x_i, x_j) = \begin{cases} 1 & ||x_i - x_j|| > r \\ 0 & ||x_i - x_j|| \leq r \end{cases}$$

Therefore the density is

$$f(\{x_1, \ldots, x_n\}) = \begin{cases} \alpha\beta^n & ||x_i - x_j|| > r \text{ for all } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

This is the density of a Poisson process of intensity $\beta$ conditioned on the event that no two points lie closer with $r$ units apart.



Consider generalizing this to

$$c(x_i, x_j) = \begin{cases} 1 & ||x_i - x_j|| > r \\ \gamma & ||x_i - x_j|| \leq r \end{cases}$$

where $\gamma \in [0, 1]$ is a parameter controlling the strength of the interaction between points. $\gamma = 0$ corresponds to the hard core process, $\gamma = 1$ is the homogeneous Poisson process, and for values in between, there is some degree of inhibition.

For values of $\gamma > 1$ the density is not integrable, so this is a model for inhibition only. It's called the *Strauss process*.



As we will see, formal likelihood methods can be difficult computationally for point processes. Although these difficulties may be overcome using Monte Carlo methods, it is useful to have simpler estimators as well.

Consider an estimator based on matching second-order properties. Recall the K function, defined for a stationary and isotropic function as

$$K(r) = \frac{1}{\lambda} E[N_o(r)]$$

where $N_o(r)$ is the number of events within radius $r$ of an arbitrary event.

It turns out that the K function can be found in closed form for a few special cases, beyond just CSR. This is true for many Poisson cluster processes.

For example, consider a Thomas process for which the rate for the parent process is $\lambda_c$, each parent has a Poisson number of children with mean $\mu$, and the locations of the children relative to the parents are iid $MVN(0, \sigma^2 I_2)$. The K function is

$$K(r) = \pi r^2 + \frac{1}{\lambda_c}[1 - \exp\{-r^2/(4\sigma^2)\}]$$



The *method of minimum contrast* (coming up) estimates parameters in the model process by comparing an estimate of the K function to the theoretical K function with those parameters.

So we also need an estimate of K. A natural estimate of $\lambda$ is $N(S)/|S|$, where $S$ is the observation window.

Consider estimating $E[N_0(r)]$ by

$$\tilde{E}(r) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} I(||x_i - x_j|| \leq r)$$



$\tilde{E}(r)$ is biased due to edge effects. Ripley (1976) suggested the correction

$$\hat{E}(r) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} w_{ij}^{-1} I(||x_i - x_j|| \leq r)$$

Let $w(x, u)$ denote the proportion of the circumference of the circle with center $x$ and radius $u$ lying within the observation window $S$. Define $w_{ij} = w(x_i, u_{ij})$ and note that it is not necessarily true that $w_{ij} = w_{ji}$.

This estimator is unbiased. Other corrections have also been proposed.



Now define a discrepancy function

$$D(\theta) = \int_a^b |\hat{K}(r)^q - K_\theta(r)^q|^p dr$$

for parameters $\theta$ of the family of processes being fit. A minimum contrast estimator of $\theta$ minimizes this discrepancy function. The need to specify the tuning parameters $0 \leq a < b < \infty$ and $p, q > 0$ is one drawback to this method. For clustered data, Diggle (2003) recommends $a = 0, p = 2, q = 1/4$ and taking $b$ no more than $1/4$ for observations on the unit square.

This method can be extended to the case that the theoretical form of the K function is not known, by using

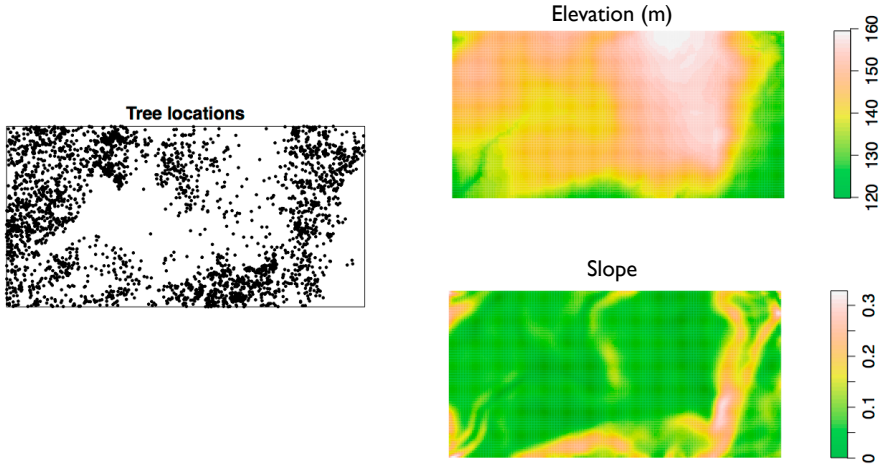
$$\hat{D}(\theta) = \int_a^b |\hat{K}(r)^q - \hat{K}_\theta(r)^q|^p dr$$

for an estimate $\hat{K}_\theta(r)$ taken to be the mean of the estimated K functions for a large number of simulated datasets under the model with parameter $\theta$.

Note that this method is quite computationally expensive; within each iteration of a numerical maximization algorithm, we need to take a new set of Monte Carlo samples.



A motivating example for fitting models with covariates:






The likelihood for Poisson processes can be written in closed form:

When the process is homogeneous with parameter $\lambda$, the likelihood for observations over region W is

$$\mathcal{L}(\lambda) = e^{-\lambda|W|}\lambda^n$$

where $n$ is the number of observations.

Therefore $\hat{\lambda} = n/|W|$ is the MLE. It has variance $\lambda/|W|^2$.



Now consider an inhomogeneous Poisson process with intensity $\lambda_\theta(x)$, where $\theta$ is a vector of unknown parameters. The likelihood is

$$\mathcal{L}(\theta) = \exp\left\{-\int_W \lambda_\theta(u)du\right\}\prod_{i=1}^n \lambda_\theta(x_i)$$

For example, suppose $\log\lambda(x) = \sum_{j=1}^p \beta_j z_j(x)$ for spatial covariates $z_j$.

To evaluate the likelihood, we need to have observed the covariates everywhere in the observation window, not just at the locations of events.

If $\log \lambda_\theta(x)$ is linear in $\theta$, then the log likelihood is concave and so there is a unique MLE. However, it is not analytically tractable.

We can maximize the log likelihood numerically, at each iteration calculating the integral numerically. This problem was treated by Berman and Turner (1992), who showed that using a quadrature rule for the integral produced problem equivalent to solving for $\theta$ in a weighted Poisson regression.

The implication is that these models can be fit using off-the-shelf software for GLMs. This is implemented in spatstat in the ppm function.



Models containing interaction (clustering and/or regularity) have likelihoods that are difficult to express in closed form. For example, Gibbs models have a density with an unknown normalizing constant, and Cox process models are defined hierarchically, so the (marginal) likelihood involves solving an integral.

MCMC can be used to approximate the likelihood in these cases, but we'll first consider some alternative methods that are less computationally intensive.

We saw one of these last time, the method of minimum contrast, good for Cox and cluster processes. We'll now consider a method suitable for models with regularity.



The *pseudo-likelihood* (Besag, 1978) is defined in terms of the *Papangelou conditional intensity*. If the probability density is $f_\theta(x),$ the conditional intensity is defined as

$$\lambda_\theta(s, x) = \begin{cases} f_\theta(x \cup \{s\})/f_\theta(x) & s \notin x \\ f_\theta(x)/f_\theta(x - \{s\}) & s \in x \end{cases}$$

For a Poisson process, $\lambda_\theta(s, x) = \lambda_\theta(s)$.

For the Strauss process, the density is $f(x) = \alpha\beta^{n(x)}\gamma^{t(x)}$ where $n(x)$ is the number of points in $x$ and $t(x)$ is the number of pairs of points within distance $r$. What is the conditional intensity?



The pseudo-likelihood function is

$$\mathcal{L}_p(\theta; x) = \exp\left\{-\int_W \lambda_\theta(u, x)dx\right\} \prod_{i=1}^{n} \lambda_\theta(x_i, x)$$

Since the pseudo-likelihood and the likelihood coincide for an inhomogeneous Poisson process, this means if we can write $\lambda_\theta$ as a log-linear function of $\theta$, we can again "trick" off-the-shelf software into fitting this model, even for non-Poisson models.

What happens for the Strauss process?

The parameters $\phi = (\log \beta, \log \gamma)$ are called *regular parameters*, whereas the radius $r$ is irregular.

Since $\mathcal{L}_p(\phi, r; x)$ is easily maximized for a fixed value of $r$, this suggests finding the maximum pseudo-likelihood estimate of $r$ by maximizing the *profile pseudo-likelihood*

$$p\mathcal{L}(r; x) = \max_{\phi} \mathcal{L}(\phi, r; x)$$

This is implemented in spatstat by the profilepl function.



We'll now consider using Monte Carlo methods to approximate likelihood functions. First, let's review some basic concepts from Monte Carlo integration.

Basic MC integration uses the following approximation:

density of r.v. X

$$\begin{aligned} E[g(X)] &= \int_{\mathcal{X}} g(x) f(x) dx \\ &\approx \frac{1}{B} \sum_{i=1}^{B} g(x^{(i)}), \quad x^{(1)}, \ldots, x^{(B)} \overset{iid}{\sim} f \end{aligned}$$

Actually, we can relax the assumption of independence, and the approximation is still unbiased.



Importance sampling modifies this to

another density with same support

$$\begin{aligned} E[g(X)] &= \int_{\mathcal{X}} g(x) \frac{f(x)}{h(x)} h(x) dx \\ &\approx \frac{1}{B} \sum_{i=1}^{B} g(x^{(i)}) \frac{f(x^{(i)})}{h(x^{(i)})}, \quad x^{(1)}, \ldots, x^{(B)} \overset{iid}{\sim} h \end{aligned}$$

The usual motivations for using importance sampling are either 1) if you can't sample from $f$ directly, or 2) to reduce the variance of the approximation, over the basic MC approximation. However, our motivation will be slightly different.



One class of densities we are interested in, for Gibbs processes, are a special case of *normalizing constant families*. That is, suppose we can write the density as

$$f_\theta(x) = \alpha(\theta) k_\theta(x)$$

where

$$\alpha(\theta)^{-1} = \int_{\mathcal{X}} k_\theta(x) dx$$

We want to maximize $f_\theta(x)$, or, equivalently, $\ell(\theta) = \log f_\theta(x)$, over $\theta$ to get the MLE, but the expression for the integral is not available in closed form.

For any fixed $\theta_0$, we can re-express the integral as follows:

$$\begin{aligned}
\alpha(\theta)^{-1} &= \int_{\mathcal{X}} k_\theta(x)dx \\
&= \int_{\mathcal{X}} k_\theta(x)\frac{\alpha(\theta_0)}{\alpha(\theta_0)}\frac{k_{\theta_0}(x)}{k_{\theta_0}(x)}dx \\
&= \alpha(\theta_0)^{-1}\int_{\mathcal{X}} \frac{k_\theta(x)}{k_{\theta_0}(x)}\alpha(\theta_0)k_{\theta_0}(x)dx \\
&= \alpha(\theta_0)^{-1}E_{\theta_0}[r_{\theta,\theta_0}(x)]
\end{aligned}$$

where $r_{\theta,\theta_0}(x) = \dfrac{k_\theta(x)}{k_{\theta_0}(x)}$



Plugging this into the log-likelihood, we have

$$\begin{aligned}
\ell(\theta) &= \log k_\theta(x) - \log[\alpha(\theta)^{-1}] \\
&= \log k_\theta(x) - \log E_{\theta_0}[r_{\theta,\theta_0}(x)] + \log\alpha(\theta_0)
\end{aligned}$$

can evaluate

can approximate by

$$\frac{1}{B}\sum_{i=1}^{B}\frac{k_\theta(x^{(i)})}{k_{\theta_0}(x^{(i)})}$$

does not depend on $\theta$; can ignore for purposes of maximization!

$$x^{(1)},\ldots,x^{(B)} \overset{iid}{\sim} f_{\theta_0}$$



Example: fitting the Strauss process to the cells data

Earlier, we found maximum pseudo-likelihood estimates for the cells data under the model

$$f_\theta(x) = \alpha(\beta,\gamma,r)\beta^{n(x)}\gamma^{s_r(x)}$$

where $s_r(x)$ is the number of pairs in $x$ whose distance is less than or equal to $r$.

A reasonable strategy is to use this initial estimate as our $\theta_0$. Note that we only need to do this initial sampling step once. But how to we sample from a Strauss process?



Sampling from Gibbs processes is what MCMC was originally invented to do. We won't obtain independent samples, but after a sufficiently long run of the chain, they can be treated as samples from the same (target or stationary) distribution.

The Metropolis-Hastings algorithm is a bit more complicated than what we have seen previously, because the dimension of x is also random. We need to allow points to "be born", "die", or move their location, while ensuring that the chain will still converge.

Luckily this is already implemeted in spatstat by the rmh function.

Model checking

There are a number of formal and informal methods for evaluating the goodness of fit of a particular fitted model.

Monte Carlo tests based on estimate F, G, or K functions can be computed as before for CSR, but now using the fitted models. These should not be interpreted too literally, since the null model we are checking against depends on the data, but the corresponding plots can be useful diagnostic tools.

The envelope function in spatstat can take a fitted model; the raw data is encapsulated with it.



A (signed) *residual measure* can be defined as

$$R(B) = n(X \cap B) - \int_B \hat{\lambda}(u, x) du$$

where $\hat{\lambda}(u, x)$ is the conditional intensity (just the regular intensity for Poisson models) under the fitted model.

This measure is neither continuous nor discrete; it puts atoms of mass at each of the data points but is smooth elsewhere.



We can apply the residual measure to quadrats $B_i$ to get residuals

$$r_i = n_i - \int_{B_i} \hat{\lambda}(u) du$$

where $n_i$ is the number of points in $B_i$.

If the process is Poisson, these residuals are independent, and we can use the standardized *Pearson residuals*

$$r_i^P = \frac{n_i - e_i}{\sqrt{e_i}}$$

where $e_i = \int_{B_i} \hat{\lambda}(u) du$ to construct a goodness-of-fit test.



Rather than using quadrats, we can smooth the residual measure. Define

$$\begin{aligned} s(u) &= e(u) \int k(u - v) dR(u) \\ &= e(u) \sum_{i=1}^{n} k(u - x_i) - e(u) \int k(u - v) \hat{\lambda}(v) dv \end{aligned}$$

usual kernel estimator of intensity

kernel smoothed fitted model intensity

This is implemented in spatstat by diagnose.ppm with argument which = “smooth”.

We can also create a *lurking variable plot* for a variable Z by plotting $C(z) = R(B(z))$ against $z$, where

$$B(z) = \{u \in W : Z(u) \leq z\}$$

For example, Z could be either of the coordinate axes, or it could be an explanatory variable.

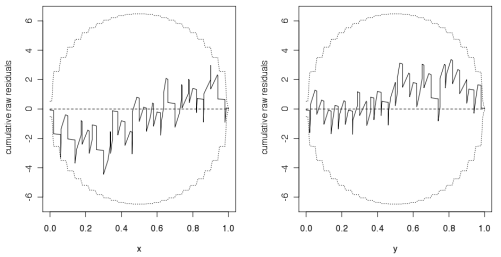
This is implemented in spatstat using the function lurking.

The derivative of $C(z)$ can be useful for diagnosing where the lack of fit occurs; this is obtained with argument cumulative = FALSE in lurking.



Note: The preceding tests and plots based on the residual measure are useful for detecting problems in the trend of the model, but they are not necessarily useful for diagnosing interaction (clustering or inhibition).

For example, here are lurking variable plots for the cells data, fitted using a homogeneous Poisson process.






To check the interaction terms in a point-process model, we need to examine the *distribution* of the residuals. For this, we can use a Q-Q plot. Pointwise simulation envelopes can be created by simulating under the fitted model. Here is the result for the cells data.