

## Hierarchical Bayes modeling

- (Hierarchical) Bayesian modeling of stationary processes
- Bayesian modeling of Generalized spatial processes
- Bayesian modeling of nonstationary processes
- Hierarchical models for areal data (disease mapping)

## Bayesian modeling of stationary processes

The basic model:

$$Y(s) = \mu(s) + w(s) + \epsilon(s)$$

where the mean  $\mu(s) = X(s)\beta$ . The residual has two components;  $w(s)$  a correlated error term, and  $\epsilon(s)$  an uncorrelated error term.

$w(s)$  introduces the partial sill  $\sigma^2$  and range  $\phi$ .

$\epsilon(s)$  introduces the nugget effect  $\tau^2$ . The nugget effect represents measurement error and/or microscale variability.

### *Isotropic models*

The covariance matrix  $\Sigma$  of  $\mathbf{Y} = (Y(s_1), \dots, Y(s_n))$  is modeling as

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I$$

where  $H$  is a correlation matrix with  $H_{ij} = \rho(s_i - s_j; \phi)$  and  $\rho$  is a valid isotropic correlation function.

Let us called  $\theta = (\beta, \sigma^2, \tau^2, \phi)$ .

Parameter estimates are obtained from the posterior:

$$p(\theta|y) \propto f(y|\theta)p(\theta)$$

where  $p(\theta)$  is the prior, and  $f$  is the likelihood.

## PRIORS:

Independent priors are usually chosen:

$$p(\theta) = p(\beta)p(\sigma^2)p(\tau^2)p(\phi)$$

good candidates are multivariate normal for  $\beta$ , and inverse gamma for  $\sigma^2$  and  $\tau^2$ . The prior of  $\phi$  depends of what  $\phi$  represents, if  $\phi$  is the inverse of the range, a gamma prior is commonly used.

Generally we prefer relatively noninformative priors. For  $\beta$  we use a flat (improper uniform) that will be a proper posterior.

However, for the covariance parameters improper priors can lead to improper posteriors.

For inferential statements about parameters separately, we need the **marginal** posterior distributions. For example, a point estimate or credible interval for  $\beta$  can be obtained from

$$\begin{aligned} p(\beta|y) &= \int \int \int p(\beta, \sigma^2, \tau^2, \phi|y) d\sigma^2 d\tau^2 d\phi \\ &\propto p(\beta) \int \int \int f(y|\theta) p(\sigma^2) p(\tau^2) p(\phi) d\sigma^2 d\tau^2 d\phi. \end{aligned}$$

There will be no closed form for the integrations, and we will resort to MCMC.

## *Hierarchical modeling*

First stage:

$$Y|\theta, W \sim N(X\beta + W, \tau^2 I)$$

where  $W = (w(s_1), \dots, w(s_n))$ .

Second stage:

$$W|\sigma^2, \phi \sim N(0, \sigma^2 H(\phi))$$

where  $H$  is the correlation matrix.

Third stage:

prior specification for  $\beta$ ,  $\tau^2$  and  $\sigma^2$  and  $\phi$ .

Regardless, the resulting  $p(\theta|y)$  is the same. In this case

$$p(\theta|y) \propto f(y|\theta, W)p(W|\theta)p(\theta),$$

which gives the same answer as

$$p(\theta|y) \propto f(y|\theta)p(\theta).$$

Since the matrix  $\sigma^2 H(\phi) + \tau^2 I$  is better behaved than  $\sigma^2 H(\phi)$  (can be close to singular) we prefer with the former approach.

Draws of  $W^{(g)}$  (simulated values from  $P(W|y)$ ) can be easily obtained, since

$$p(W|y) = \int p(W|\sigma^2, \phi)p(\sigma^2, \phi|y)d\sigma^2 d\phi$$

If  $(\sigma^{2(g)}, \phi^{(g)})$  are MCMC simulated values of the parameters, then the corresponding draws  $W^{(g)}$  are obtained from

$$p(W|\sigma^{2(g)}, \phi^{(g)}).$$

*Posterior predictive distribution (ppd)*

Denote the unknown value at location  $x_0$  of  $Y$ , by  $Y_0$  the prediction is done, using the ppd:

$$\begin{aligned} p(y_0|y, X, x_0) &= \int p(y_0, \theta|y, X, x_0)d\theta \\ &= \int p(y_0|y, \theta, x_0)p(\theta|y, X)d\theta, \end{aligned}$$

where  $(y_0|y, \theta, x_0)$  has a conditional normal distribution.

MCMC methods are used to estimate  $p(y_0|y, X, x_0)$ . Supposed that we draw (after bur-in) our posterior sample:  $\theta^{(1)}, \dots, \theta^{(G)}$  from the posterior  $p(\theta|y, X)$ .

The the above integral can be obtained as a Monte Carlo mixture of

the form

$$\hat{p}(y_0|y, X, x_0) = \frac{1}{G} \sum_{g=1}^G p(y_0|y, \theta^{(g)}, x_0)$$

In practice, we use composition sampling to draw, one for one for each  $\theta^{(g)}$ , a

$$y_0^{(g)} \sim p(y_0|y, \theta^{(g)}, x_0).$$

Then, the collection  $\{y_0^{(1)}, \dots, y_0^{(G)}\}$  is a sample from the posterior predictive density.

The models seeing here can be implemented and used in Winbugs (example in class).

## Modeling geometric anisotropy

We have that the covariance function  $\rho$  is written in terms of an isotropic covariance  $\rho_0$  (after rotation and stretching of coordinates),

$$\rho(h, \phi) = \rho_0(\|Lh\|, \phi)$$

where  $h$  is a vector distance, and  $L$  is a matrix describing the linear transformation. If  $L$  is the identity this reduces to the isotropic case.

The covariance matrix of  $Y$  would be

$$\Sigma(\alpha) = \tau^2 I + \sigma^2 H((h' B h)^{1/2})$$

where  $B = L^T L$ .

A customary prior for a positive definite matrix such as  $B$  is a *Wishart*( $R, p$ ) where

$$\pi(B) \propto |B|^{(p-n-1)/2} \exp(-1/2 \text{tr}(pBR^{-1}))$$

so that  $E(B) = R$  and  $p \geq n$  is a precision parameter, i.e.  $var(B)$  increases as  $p$  decreases (generally,  $p = 2$ .)

A priori, it is easier to assume that the process is isotropic and set  $R = \delta I$ , and model  $\delta$  as random, using an inverse gamma prior for  $\delta$ , with mean the range parameter and infinite variance.

## Generalized linear spatial models

In some point-reference dataset we obtain measurements of a variable  $Y$ , that might not be continuous. For instance,  $Y$  might be binary. We formulate a hierarchical model with the Gaussian model for  $Y$  replaced by another member of the class of exponential family models.

Assume  $Y(s_i)$  are conditionally independent given  $\beta$  and  $w(s_i)$  with distribution:

$$f(y(s_i)|\beta, w(s_i), \gamma) = h(y(s_i), \gamma) \exp(\gamma[y(s_i)\eta(s_i) - \psi(\eta(s_i))])$$

where  $\eta(s_i) = x(s_i)\beta + w(s_i)$  for a LINK function  $\eta$ ,  $\gamma$  is a dispersion parameter.

$w(s_i)$  are spatial random effects coming from a Gaussian process, i.e.

$$W \sim N(0, \sigma^2 H(\phi))$$

We have not created yet a process for  $Y$ .

1. We need to create JOINT distribution for

$$f(y(s_1), \dots, y(s)_n) | \beta, \sigma^2, \phi, \gamma$$

$$\int \left( \prod_{i=1}^n f(y(s_i) | \beta, w(s_i), \gamma) \right) p(W | \sigma^2, \phi) dW$$

2. We could add to  $w(s_i)$  a pure error term  $\epsilon(s_i)$  in the definition of  $\eta(s_i)$ . Though, this might not make sense, since  $w$  is not a residual term, it is part of the (transformed) mean structure (first stage).

## Areal data models

### Disease mapping

A very common area of epidemiology and biostatistics is DISEASE MAPPING.

$Y_i$  = observed number of cases of disease in county  $i$ ,  $i = 1, \dots, I$ .

$E_i$  = expected number of cases of disease in county  $i$ ,  $i = 1, \dots, I$ .

The  $Y_i$  are random variables, while  $E_i$  are known functions of  $n_i$ , the number of persons at risk for the disease in county  $i$ .

We might assume,

$$E_i = n_i \bar{r} = n_i \left( \frac{\sum_i y_i}{\sum_i n_i} \right)$$

$\bar{r}$  is the overall disease rate in the entire study region. This is called *internal standardization* (IS).

IS is empirical bayes, since  $\bar{r}$  is estimated from our current data. A better approach is to make reference to an existing standard table of age-adjusted rates for the disease. After stratifying the population by age group,

$$E_i = \sum_j n_{ij} r_j$$

where  $n_{ij}$  is the person-years at risk in area  $i$  for age group  $j$ , and  $r_j$  is the disease rate in age group  $j$  (taken from the table). This is *external standardization*.

A disease map is a display of the disease rates overlaid on the areal units.

## Frequentist methods

If  $E_i$  is not too large (disease is rare), we model

$$Y_i | \eta_i \sim Po(E_i \eta_i)$$

where  $\eta_i$  is the true relative risk of disease in region  $i$ . The MLE of  $\eta_i$

$$\hat{\eta}_i \equiv SMR_i = Y_i / E_i$$

The standard morbidity (or mortality) ration (SMR), the ration of observed to expected.

$$\text{var}(SMR_i) = \text{var}(Y_i) / E_i^2 = \eta_i / E_i.$$

So, we might take

$$\hat{\text{var}}(SMR_i) = \hat{\eta}_i / E_i = Y_i / E_i^2.$$

## Hierarchical methods

If we want to estimate and map the underlying relative risk surface ( $\eta_i$ ), we would want to assume that  $\eta_i$  are random effects.

If we would like take into account spatial dependencies. The natural way of handling this complex model is through hierarchical Bayesian modeling.

## Poisson-gamma model

A simple model would be

$$Y_i | \eta_i \sim (\text{ind}) \text{Po}(E_i \eta_i)$$

for  $i = 1, \dots, I$ . And,

$$\eta_i \sim (\text{ind}) G(a, b)$$

where  $G(a, b)$  denotes the gamma distribution with mean  $\mu = a/b$ , and variance  $\sigma^2 = a/b^2$ . We get then, that  $a = \mu^2/\sigma^2$ , and  $b = \mu/\sigma^2$ .

This is a conjugate model. The posterior

$$p(\eta_i | y_i) \sim G(y_i + a, E_i + b).$$

An estimate of  $\eta_i$  is

$$E(\eta_i | y) = w_i SMR_i + (1 - w_i)\mu.$$

where  $w_i = E_i/[E_i + (\mu/\sigma^2)]$ . This is a weighted average of the SMR and the prior mean  $\mu$ .

This is equal to  $SMR_i$  when  $w_i$  is close to 1 ( $E_i$  is big, the data are informative, or  $\sigma^2$  is big, the prior is weakly informative). This will be approximately  $\mu$  when  $w_i$  is zero.

## Poisson-lognormal models

The gamma prior is convenient for computation, but it does not allow spatial correlation among the  $\eta_i$ . Instead we use,

$$Y_i|\psi \sim (ind)Po(E_i e^{\psi_i})$$

where

$$\psi_i = X_i\beta + \theta_i + \phi_i.$$

The  $x_i$  are explanatory variables.

$\theta_i$  captures region-wide heterogeneity,

$$\theta_i \sim N(\mu, 1/\tau_h)$$

$\tau_h$  is the precision. It captures global extra-Poisson variability in the log-relative risk.

$\phi_i$  make this a truly spatial model. It captures local extra-Poisson

variability in the log-relative risk. We might assume

$$\phi|\mu, \lambda \sim N_I(0, H(\lambda))$$

$N_I$  is a  $I$ -dimensional normal,  $\mu$  is the mean, and  $H$  the covariance.

Due to the heavy computational needs, we replace the Gaussian model by a CAR model.

$$\phi \sim CAR(\tau_c)$$

this is the improper car (IAR), with 0-1 adjacency weights,  $\tau_c$  is a precision parameter.

## Difficulties of CAR models

The previous CAR framework is very convenient, it can be efficiently implemented using Gibbs sampling.

The full conditional of  $\phi_i$  is

$$p(\phi_i | \phi_j, \theta, \beta, y) \propto Po(y_i | E_i \exp x_i \beta + \theta_i + \phi_i) \times N(\phi_i | \bar{\phi}_i, 1/(\tau_c m_i)).$$

This conditional model is computationally convenient. Though, CAR models have numerous theoretical difficulties.

## 1. Impropriety.

The IAR model is improper. It can be made proper by adding a "propriety parameter"  $\alpha$ . (where  $|\alpha| < 1$ ). This new proper prior does not deliver enough spatial similarity unless  $\alpha$  is close to 1, getting us to the same problem.

A common approach is to use the IAR model since the posterior is proper. But, this requires some care. This improper prior is identified only up to an additive constant. We must add the constraint  $\sum \phi_i = 0$ .

## 2. Selection of $\tau_c$ and $\tau_h$ .

These parameters can not be chose arbitrarily large, since  $\phi_i$  and  $\theta_i$  would be unidentifiable.

We can place a third-stage priors (hyperpriors), with the gamma (conjugate) prior

$$\tau_h \sim G(a_h, b_h)$$

$$\tau_c \sim G(a_c, b_c).$$

A fair specification might be,

$$sd(\theta_i) = \frac{1}{\sqrt{\tau_h}} \sim \frac{1}{0.7\sqrt{(\bar{m}\tau_c)}} \sim sd(\phi_i).$$

where  $\bar{m}$  is the average number of neighbors.