AstroStat Presentation, or
A collection of stuff from Stat 225

Luke Bornn

Department of Statistics, Harvard University

October 22, 2013
Outline

Gaussian Processes

Hierarchical Bayes
  Intro to Bayes
  Bayesian Computation in 3 slides
  Sampling to the Rescue

Spatial GLMs
Gaussian process as prior over functions

Source: packages.python.org/infpy/gps.html
Gaussian process as prior over functions

Source: packages.python.org/infpy/gps.html
Suppose at each location \( s_i \), \( Z(s_i) \) is Gaussian with mean \( \mu_i \) and variance \( \sigma_i^2 \), and that the between-site covariance matrix is \( \Sigma \). \( \{Z(s_i), i = 1, \ldots, n\} \) is then multivariate normal with mean vector \( \mu \) and covariance matrix \( \Sigma \).
Covariance and Prediction: An Example

- Suppose at each location $s_i$, $Z(s_i)$ is Gaussian with mean $\mu_i$ and variance $\sigma_i^2$, and that the between-site covariance matrix is $\Sigma$.  

\[ \{Z(s_i), i = 1, \ldots, n\} \]  

is then multivariate normal with mean vector $\mu$ and covariance matrix $\Sigma$.

- Now suppose we observe sites $2, \ldots, n$, i.e. we have observations $z_{2:n} = z(s_2), \ldots, z(s_n)$.

- Can we work out the distribution of $Z(s_1)$, given the observations?
Covariance and Prediction: An Example

- Suppose at each location $s_i$, $Z(s_i)$ is Gaussian with mean $\mu_i$ and variance $\sigma^2_i$, and that the between-site covariance matrix is $\Sigma$. The set $\{Z(s_i), i = 1, \ldots, n\}$ is then multivariate normal with mean vector $\mu$ and covariance matrix $\Sigma$.

- Now suppose we observe sites 2, \ldots, $n$, i.e. we have observations $z_{2:n} = z(s_2), \ldots, z(s_n)$.

- Can we work out the distribution of $Z(s_1)$, given the observations?

- Yes! As the process is multivariate normal, we know that

$$Z(s_1)|z_{2:n} \sim \mathcal{N} \left( \mu_1 + \Sigma_{12} \Sigma^{-1}_{22} (z_{2:n} - \mu_{2:n}), \sigma^2_1 - \Sigma_{12} \Sigma^{-1}_{22} \Sigma_{21} \right)$$
Covariance and Prediction: An Example

Suppose at each location \( s_i \), \( Z(s_i) \) is Gaussian with mean \( \mu_i \) and variance \( \sigma_i^2 \), and that the between-site covariance matrix is \( \Sigma \). \{\( Z(s_i), i = 1, \ldots, n \)\} is then multivariate normal with mean vector \( \mu \) and covariance matrix \( \Sigma \).

Now suppose we observe sites 2, \ldots, n, i.e. we have observations \( z_{2:n} = z(s_2), \ldots, z(s_n) \).

Can we work out the distribution of \( Z(s_1) \), given the observations?

Yes! As the process is multivariate normal, we know that

\[
Z(s_1) | z_{2:n} \sim \mathcal{N} \left( \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (z_{2:n} - \mu_{2:n}), \sigma_1^2 - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right)
\]

What happens when we add a new location?
Covariance functions

- Prediction requires estimating the mean $\mu$ ($n$ parameters) and covariance $\Sigma$ ($n(n + 1)/2$ parameters). What to do for unobserved locations?
Covariance functions

- Prediction requires estimating the mean $\mu$ ($n$ parameters) and covariance $\Sigma$ ($n(n+1)/2$ parameters). What to do for unobserved locations?
- We need to simplify things via some assumptions, the most common of which is to assume second-order (or weak) stationarity:

$$E[Z(s)] = \mu$$

$$\text{Cov}[Z(s), Z(s')] = \text{Cov}[Z(s + \delta), Z(s' + \delta)] \quad \forall\delta.$$ 

Specifically, the covariance only depends on the spatial lag $h = s - s'$ between locations. We call $C(h) = \text{Cov}[Z(0), Z(h)]$ the covariance function.
Bayes Theorem

- Assume we have some parameters $\theta = (\theta_1, \ldots, \theta_p)$ which come from the prior distribution $\pi(\theta|\eta)$ and we observe some data $z = (z_1, \ldots, z_n)$ which has the distribution $\pi(z|\theta)$

$$
\pi(\theta|z, \eta) = \frac{\pi(z|\theta)\pi(\theta|\eta)}{\int \pi(z|\theta)\pi(\theta|\eta)d\theta}
$$
Bayes Theorem

- Assume we have some parameters $\theta = (\theta_1, \ldots, \theta_p)$ which come from the prior distribution $\pi(\theta|\eta)$ and we observe some data $z = (z_1, \ldots, z_n)$ which has the distribution $\pi(z|\theta)$

$$
\pi(\theta|z, \eta) = \frac{\pi(z|\theta)\pi(\theta|\eta)}{\int \pi(z|\theta)\pi(\theta|\eta)d\theta}
$$

- Sometimes we put a hyperprior on $\eta$, $\pi(\eta)$, and the posterior then becomes

$$
\pi(\theta|z) = \frac{\int \pi(z|\theta)\pi(\theta|\eta)\pi(\eta)d\eta}{\int \int \pi(z|\theta)\pi(\theta|\eta)\pi(\eta)d\eta d\theta}
$$
Bayes Theorem

- Assume we have some parameters $\theta = (\theta_1, \ldots, \theta_p)$ which come from the prior distribution $\pi(\theta | \eta)$ and we observe some data $z = (z_1, \ldots, z_n)$ which has the distribution $\pi(z | \theta)$

$$
\pi(\theta | z, \eta) = \frac{\pi(z | \theta)\pi(\theta | \eta)}{\int \pi(z | \theta)\pi(\theta | \eta) d\theta}
$$

- Sometimes we put a hyperprior on $\eta$, $\pi(\eta)$, and the posterior then becomes

$$
\pi(\theta | z) = \frac{\int \pi(z | \theta)\pi(\theta | \eta)\pi(\eta) d\eta}{\int \int \pi(z | \theta)\pi(\theta | \eta)\pi(\eta) d\eta d\theta}
$$

- Alternatively, we can take an empirical Bayes approach and find a value of $\eta$ to maximize $\pi(z | \eta)$. 

Point Estimation

- Assuming we have the posterior distribution $\pi(\theta|z)$, we can find point estimates of $\theta$
Point Estimation

- Assuming we have the posterior distribution $\pi(\theta|z)$, we can find point estimates of $\theta$
- The mean:
  \[ \hat{\theta} = \mathbb{E}(\theta|z) \]
- The median:
  \[ \hat{\theta} : \int_{-\infty}^{\hat{\theta}} \pi(\theta|z) d\theta = 0.5 \]
Point Estimation

Assuming we have the posterior distribution $\pi(\theta|z)$, we can find point estimates of $\theta$

- The mean:
  $$\hat{\theta} = \mathbb{E}(\theta|z)$$

- The median:
  $$\hat{\theta} : \int_{-\infty}^{\hat{\theta}} \pi(\theta|z) d\theta = 0.5$$

- The mode (aka the MAP):
  $$\hat{\theta} : \sup_{\theta} \pi(\theta|z)$$
Interval Estimation

- We can create a 95% credible interval by finding the values

\[
\int_{-\infty}^{l_l} \pi(\theta | z) = \frac{\alpha}{2} \quad \text{and} \quad \int_{l_u}^{\infty} \pi(\theta | z) = \frac{\alpha}{2}
\]
Interval Estimation

- We can create a 95% credible interval by finding the values

\[ \int_{-\infty}^{l_l} \pi(\theta|z) = \alpha/2 \quad \text{and} \quad \int_{l_u}^{\infty} \pi(\theta|z) = \alpha/2 \]

- A shorter 95% interval is the set

\[ \theta : \pi(\theta|z) > c \quad \text{where we maximize } c \text{ such that} \]

\[ \int_{\pi(\theta|z) > c} \pi(\theta|z) = 0.95 \]
Gibbs Sampler

Recall that we can use samples $\theta^{(1)}, \ldots, \theta^{(T)}$ from $\pi(\theta|z)$ to estimate expectations $\mathbb{E}(g(\theta)|z)$ via

$$\hat{\mathbb{E}}(g(\theta)|z) = \frac{1}{T} \sum_{t=1}^{T} g(\theta^{(t)})$$

One way to generate these samples is to iteratively sample from the full conditionals $\pi(\theta_i|\theta_{-i}, z)$.
Gibbs Sampler

- Recall that we can use samples $\theta^{(1)}, \ldots, \theta^{(T)}$ from $\pi(\theta|z)$ to estimate expectations $\mathbb{E}(g(\theta)|z)$ via

$$
\hat{\mathbb{E}}(g(\theta)|z) = \frac{1}{T} \sum_{1}^{T} g(\theta^{(t)})
$$

One way to generate these samples is to iteratively sample from the full conditionals $\pi(\theta_i|\theta_{-i}, z)$

- The Gibbs sampler first finds starting values $\theta^{(0)}_1, \ldots, \theta^{(0)}_p$, then iterates, for $t$ in $1, \ldots, T$

1. Sample $\theta^{(t)}_1$ from $\pi(\theta_1|\theta^{(t-1)}_2, \ldots, \theta^{(t-1)}_p, z)$
2. Sample $\theta^{(t)}_2$ from $\pi(\theta_2|\theta^{(t)}_1, \theta^{(t-1)}_3, \ldots, \theta^{(t-1)}_p, z)$
   
   $\vdots$
3. Sample $\theta^{(t)}_p$ from $\pi(\theta_p|\theta^{(t)}_1, \ldots, \theta^{(t)}_{p-1}, z)$
Metropolis-Hastings

- What if you can’t sample from $\pi(\theta | \theta_{-i}, z)$?
- You can instead propose $\theta^*$ from some (symmetric) proposal distribution $q(\theta)$
Metropolis-Hastings

- What if you can’t sample from $\pi(\theta|\theta_{-i}, z)$?
- You can instead propose $\theta^*$ from some (symmetric) proposal distribution $q(\theta)$
- Calculate

$$r = \frac{\pi(\theta^*|z)}{\pi(\theta^{(t-1)}|z)}$$
Metropolis-Hastings

- What if you can’t sample from $\pi(\theta|\theta_{-i}, z)$?
- You can instead propose $\theta^*$ from some (symmetric) proposal distribution $q(\theta)$
- Calculate

\[
r = \frac{\pi(\theta^*|z)}{\pi(\theta^{(t-1)}|z)}
\]

- If $r \geq 1$, set $\theta^{(t)} = \theta^*$
- If $r \leq 1$, set $\theta^{(t)} = \theta^*$ with probability $r$, and $\theta^{(t)} = \theta^{(t-1)}$ with probability $1 - r$. 
The hierarchical Bayes framework

- In the hierarchical process framework, we want to learn about an underlying process $Y$ through some (noisy, polluted, transformed) data $Z$.
- $Z$ comes from $Y$ through $\pi(Z|Y, \theta)$ where $\theta$ are some parameters.
- The process $Y$ also often depends on some such parameters, via $\pi(Y|\theta)$. 
The hierarchical Bayes framework

- In the hierarchical process framework, we want to learn about an underlying process $Y$ through some (noisy, polluted, transformed) data $Z$.
- $Z$ comes from $Y$ through $\pi(Z|Y,\theta)$ where $\theta$ are some parameters.
- The process $Y$ also often depends on some such parameters, via $\pi(Y|\theta)$.
Extending the hierarchical process model

\[ Z(s) = \mu(s) + Y(s) + \epsilon(s) \] with some parameters \( \theta \) where

\[ \mu(s) = x^T(s)\beta \quad \text{and} \]
\[ Y(s) | \theta \sim N(0, \Sigma) \]

where \( \Sigma_{ij} = C_{\sigma^2,\phi}(s_i - s_j) = \sigma^2 \rho_\phi(s_i - s_j) \). Here \( \epsilon(s) \) is a white-noise process with parameter \( \sigma^2_\epsilon \).
Extending the hierarchical process model

- \( Z(s) = \mu(s) + Y(s) + \epsilon(s) \) with some parameters \( \theta \) where
  \[
  \mu(s) = x^T(s)\beta \quad \text{and} \quad Y(s) | \theta \sim N(0, \Sigma)
  \]
  where \( \Sigma_{ij} = C_{\sigma^2, \phi}(s_i - s_j) = \sigma^2 \rho(\phi)(s_i - s_j) \).
  Here \( \epsilon(s) \) is a white-noise process with parameter \( \sigma^2_\epsilon \).
- We would also assign a prior distribution to \( \theta \), \( \pi(\theta) \).
- To approximate the posterior, iteratively sample \( Y \) as well as the parameters \( \theta = \beta, \phi, \sigma^2, \sigma^2_\epsilon \).
Marginalizing out the latent process

- We can alternatively write

\[ Z(s) \sim N(x^T(s)\beta, \Sigma + \sigma^2_\epsilon I) \]

and avoid the sampling of \( Y \). DEMO
Marginalizing out the latent process

- We can alternatively write

\[ Z(s) \sim N(x^T(s)\beta, \Sigma + \sigma^2 \epsilon I) \]

and avoid the sampling of \( Y \). DEMO

- However, we’re often interested in

\[
\pi(Y|z) = \int \pi(Y, \theta|z) d\theta
\]

\[ = \int \pi(Y|\theta, z)\pi(\theta|z)d\theta \]
Marginalizing out the latent process

- We can alternatively write

\[ Z(s) \sim N(x^T(s)\beta, \Sigma + \sigma^2_\epsilon I) \]

and avoid the sampling of \( Y \). DEMO

- However, we’re often interested in

\[
\pi(Y|z) = \int \pi(Y, \theta|z)d\theta
\]

\[= \int \pi(Y|\theta, z)\pi(\theta|z)d\theta \]

- If we have samples \( \theta^1, \ldots, \theta^T \) from \( \pi(\theta|z) \), then samples \( Y^{(t)} \) from

\[ Y^{(t)} \sim \pi(Y|\theta^{(t)}, z) \]

will be distributed as \( \pi(Y|z) \), as desired.
What about predicting at an unknown location \( s_0 \)?

- We need to find the predictive distribution

\[
\pi(Z(s_0)|z, \theta) = \int \pi(Z(s_0), \theta|z, x, x(s_0))d\theta
\]

\[
= \int \pi(Z(s_0)|z, \theta, x(s_0))\pi(\theta|z, x)d\theta
\]

where \( \pi(Z(s_0)|z, \theta, x(s_0)) \) is a conditional normal, given the joint multivariate normal structure of \( Z(s_0) \) and the data \( z \).
What about predicting at an unknown location \( s_0 \)?

- We need to find the predictive distribution

\[
\pi(Z(s_0)|z, \theta) = \int \pi(Z(s_0), \theta|z, x, x(s_0))d\theta
\]

\[
= \int \pi(Z(s_0)|z, \theta, x(s_0))\pi(\theta|z, x)d\theta
\]

where \( \pi(Z(s_0)|z, \theta, x(s_0)) \) is a conditional normal, given the joint multivariate normal structure of \( Z(s_0) \) and the data \( z \).

- If we have samples \( \theta^1, \ldots, \theta^T \) from \( \pi(\theta|z, x) \), then the predictive integral is computed via a Monte Carlo mixture,

\[
\hat{\pi}(Z(s_0)|z, x, x(s_0)) = \frac{1}{T} \sum_{t=1}^{T} \pi(Z(s_0)|z, \theta^{(t)}, x(s_0))
\]
Sampling to the rescue, again

- Again, we have samples $\theta^1, \ldots, \theta^T$ from $\pi(\theta | z, x)$. Next simulate

$$z_0^{(t)} \sim \pi(Z(s_0) | z, \theta^{(t)}, x(s_0))$$

which creates a set of samples from the posterior predictive density.
Sampling to the rescue, again

- Again, we have samples $\theta_1, \ldots, \theta_T$ from $\pi(\theta|z, x)$. Next simulate

  $$z_0^{(t)} \sim \pi(Z(s_0)|z, \theta^{(t)}, x(s_0))$$

  which creates a set of samples from the posterior predictive density.

- We can use these samples to find a point estimate (mean, median, etc.) as well as prediction variance.
From continuous to binary data

▶ In the hierarchical framework, we modeled Gaussian data as

\[ Z(s) \sim \mathcal{N}(X\beta + Y(s), \tau^2 I) \]
\[ Y(s) \sim \mathcal{N}(0, \Sigma_{\sigma^2,\phi}) \]

plus potentially further prior information on \( \tau^2, \sigma^2, \beta, \phi, \) etc.
From continuous to binary data

- In the hierarchical framework, we modeled Gaussian data as

\[ Z(s) \sim \mathcal{N}(X\beta + Y(s), \tau^2 I) \]
\[ Y(s) \sim \mathcal{N}(0, \Sigma_{\sigma^2,\phi}) \]

plus potentially further prior information on \( \tau^2, \sigma^2, \beta, \phi, \) etc.

- We could instead write, for example that \( Z|Y \) is Poisson or Binomial
Connection to Generalized Linear Mixed Models

Assume our data comes from an exponential family,

\[
\pi(Z(s)|\beta, Y(s), \kappa) = h(Z(s), \kappa) \exp\{\kappa(Z(s)\eta(s) - \Phi(\eta(s)))\}
\]

where \( g(\eta(s)) = x(s)\beta + Y(s) \) for some link function \( g \), where \( \kappa \) is a dispersion parameter. This family of distributions includes the Gaussian, Poisson, Binomial, and many others.
Assume our data comes from an exponential family,

\[ \pi(Z(s)|\beta, Y(s), \kappa) = h(Z(s), \kappa) \exp\{\kappa(Z(s)\eta(s) - \Phi(\eta(s)))\} \]

where \( g(\eta(s)) = x(s)\beta + Y(s) \) for some link function \( g \), where \( \kappa \) is a dispersion parameter. This family of distributions includes the Gaussian, Poisson, Binomial, and many others.

As before, we assume

\[ Y(s) \sim \mathcal{N}(0, \Sigma_{\sigma^2,\phi}) \]

If, on the contrary, \( Y \) was iid, then this would be the usual generalized linear mixed model (GLMM). Hence, what we have is still a GLMM, but with spatial correlation in the random effects.
Notes on the GLMM framework

▶ Firstly, we have not created a “spatial process” for $Z$. Rather, we have defined a joint distribution $\pi(Z(s)|\beta, \sigma^2, \phi, \kappa)$, namely

$$\int \left( \prod_{i=1}^{n} \pi(Z(s_i)|\beta, \sigma^2, \phi, \kappa) \right) \pi(Y(s)|\sigma^2, \phi) dY$$
Notes on the GLMM framework

Firstly, we have not created a “spatial process” for $Z$. Rather, we have defined a joint distribution $\pi(Z(s)|\beta, \sigma^2, \phi, \kappa)$, namely

$$\int \left( \prod_{i=1}^{n} \pi(Z(s_i)|\beta, \sigma^2, \phi, \kappa) \right) \pi(Y(s)|\sigma^2, \phi) dy$$

Secondly, there’s no need to include random (white) noise $\epsilon$ because stochastic variability is already included in the specification of $\pi(Z(s)|\beta, Y(s), \kappa)$.