AstroStat Presentation, or A collection of stuff from Stat 225

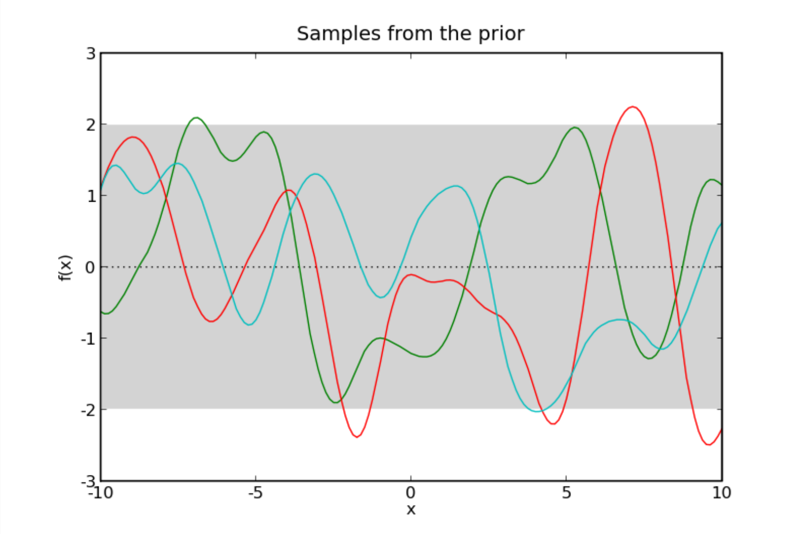
Luke Bornn

Department of Statistics, Harvard University

October 22, 2013

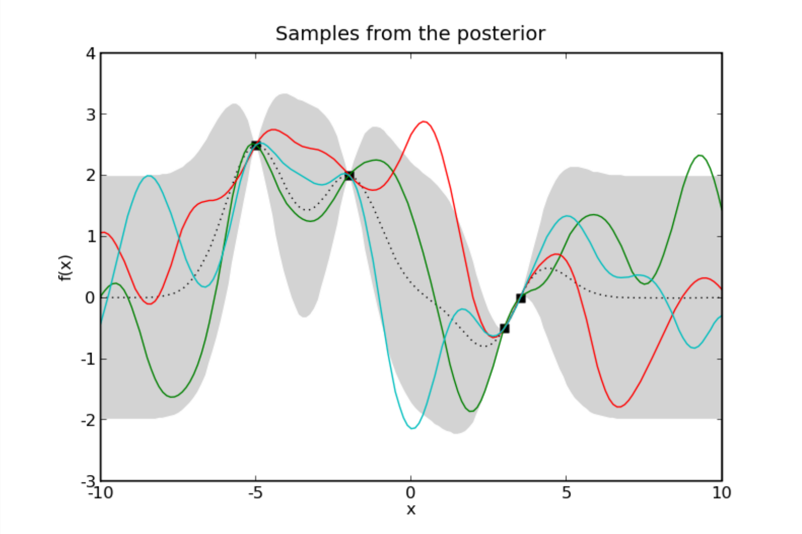
Outline

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

Covariance and Prediction: An Example

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

Covariance and Prediction: An Example

- Suppose at each location s_i, $Z(s_i)$ is Gaussian with mean μ_i and variance σ_i^2, and that the between-site covariance matrix is Σ. $\{Z(s_i), i = 1, \ldots, n\}$ is then multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ
- Now suppose we observe sites $2, \ldots, n$, i.e. we have observations $\mathbf{z}_{2:n} = z(s_2), \ldots, z(s_n)$.
- 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 μ_i and variance σ_i^2, and that the between-site covariance matrix is Σ. $\{Z(s_i), i = 1, \dots, n\}$ is then multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ
- Now suppose we observe sites $2, \dots, n$, i.e. we have observations $\mathbf{z}_{2:n} = z(s_2), \dots, z(s_n)$.
- Can we work out the distribution of $Z(s_1)$, given the observations?
- Yes! As the process is multivariate normal, we know that

$$Z(s_1)|\mathbf{z}_{2:n} \sim \mathbb{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{z}_{2:n} - \mu_{2:n}), \sigma_1^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

Covariance and Prediction: An Example

- Suppose at each location s_i, $Z(s_i)$ is Gaussian with mean μ_i and variance σ_i^2, and that the between-site covariance matrix is Σ. $\{Z(s_i), i = 1, \ldots, n\}$ is then multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ
- Now suppose we observe sites $2, \ldots, n$, i.e. we have observations $\mathbf{z}_{2:n} = z(s_2), \ldots, z(s_n)$.
- Can we work out the distribution of $Z(s_1)$, given the observations?
- Yes! As the process is multivariate normal, we know that

$$Z(s_1)|\mathbf{z}_{2:n} \sim \mathbb{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{z}_{2:n} - \mu_{2:n}), \sigma_1^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

- What happens when we add a new location?

Covariance functions

- Prediction requires estimating the mean μ (n parameters) and covariance Σ ($n(n+1)/2$ parameters). What to do for unobserved locations?

Covariance functions

- Prediction requires estimating the mean μ (n parameters) and covariance Σ ($n(n+1)/2$ 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:

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

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

Bayes Theorem

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

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

Bayes Theorem

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

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

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

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

Bayes Theorem

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

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

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

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

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

Point Estimation

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

Point Estimation

- Assuming we have the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{z})$, we can find point estimates of $\boldsymbol{\theta}$
- The mean:

$$\hat{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{z})$$

- The median:

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

Point Estimation

- Assuming we have the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{z})$, we can find point estimates of $\boldsymbol{\theta}$
- The mean:

$$\hat{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{z})$$

- The median:

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

- The mode (aka the MAP):

$$\hat{\boldsymbol{\theta}} : \sup_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\boldsymbol{z})$$

Interval Estimation

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

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

Interval Estimation

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

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

- A shorter 95% interval is the set

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

$$\int_{\pi(\boldsymbol{\theta}|\boldsymbol{z})>c} \pi(\boldsymbol{\theta}|\boldsymbol{z}) = 0.95$$

Gibbs Sampler

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

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

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

Gibbs Sampler

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

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

  One way to generate these samples is to iteratively sample from the full conditionals $\pi(\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{z})$
- The Gibbs sampler first finds starting values $\theta_1^{(0)}, \ldots, \theta_p^{(0)}$, then iterates, for t in $1, \ldots, T$
  1. Sample $\theta_1^{(t)}$ from $\pi(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_p^{(t-1)}, \boldsymbol{z})$
  2. Sample $\theta_2^{(t)}$ from $\pi(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_p^{(t-1)}, \boldsymbol{z})$

     $\vdots$
  3. Sample $\theta_p^{(t)}$ from $\pi(\theta_p|\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \boldsymbol{z})$

Metropolis-Hastings

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

Metropolis-Hastings

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

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

Metropolis-Hastings

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

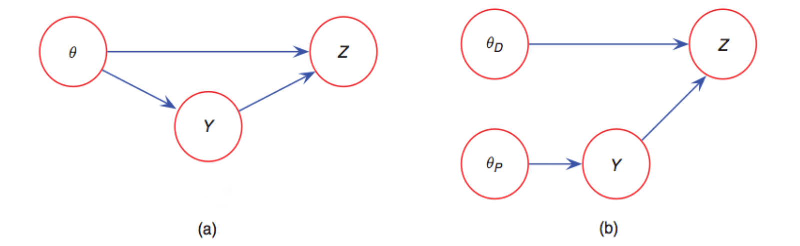
$$r = \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{z})}{\pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{z})}$$

- If $r \geq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$
- If $r \leq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ with probability r, and $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$ 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 θ 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 θ are some parameters.
- The process Y also often depends on some such parameters, via $\pi(Y|\theta)$
- 

Extending the hierarchical process model

- $Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + Y(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$ with some parameters $\boldsymbol{\theta}$ where

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

where $\Sigma_{ij} = C_{\sigma^2,\phi}(\boldsymbol{s}_i - \boldsymbol{s}_j) = \sigma^2 \rho_\phi(\boldsymbol{s}_i - \boldsymbol{s}_j)$. Here $\epsilon(\boldsymbol{s})$ is a white-noise process with parameter σ_ϵ^2.

Extending the hierarchical process model

- $Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + Y(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$ with some parameters $\boldsymbol{\theta}$ where

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

  where $\Sigma_{ij} = C_{\sigma^2,\phi}(\boldsymbol{s}_i - \boldsymbol{s}_j) = \sigma^2 \rho_\phi(\boldsymbol{s}_i - \boldsymbol{s}_j)$. Here $\epsilon(\boldsymbol{s})$ is a white-noise process with parameter σ^2_ϵ.
- We would also assign a prior distribution to $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$
- To approximate the posterior, iteratively sample $\boldsymbol{Y}$ as well as the parameters $\boldsymbol{\theta} = \beta, \phi, \sigma^2, \sigma^2_\epsilon$.

Marginalizing out the latent process

- We can alternatively write

$$Z(\boldsymbol{s}) \sim N(\boldsymbol{x}^T(\boldsymbol{s})\beta, \Sigma + \sigma_\epsilon^2 \boldsymbol{I})$$

and avoid the sampling of $\boldsymbol{Y}$. **DEMO**

Marginalizing out the latent process

- We can alternatively write

$$Z(\boldsymbol{s}) \sim N(\boldsymbol{x}^T(\boldsymbol{s})\beta, \Sigma + \sigma_\epsilon^2 \boldsymbol{I})$$

  and avoid the sampling of $\boldsymbol{Y}$. DEMO
- However, we're often interested in

$$\pi(\boldsymbol{Y}|z) = \int \pi(\boldsymbol{Y}, \boldsymbol{\theta}|z) d\boldsymbol{\theta}$$
$$= \int \pi(\boldsymbol{Y}|\boldsymbol{\theta}, z)\pi(\boldsymbol{\theta}|z) d\boldsymbol{\theta}$$

Marginalizing out the latent process

- We can alternatively write

$$Z(\boldsymbol{s}) \sim N(\boldsymbol{x}^T(\boldsymbol{s})\beta, \Sigma + \sigma_\epsilon^2 \boldsymbol{I})$$

and avoid the sampling of $\boldsymbol{Y}$. **DEMO**

- However, we're often interested in

$$\begin{aligned}\pi(\boldsymbol{Y}|z) &= \int \pi(\boldsymbol{Y}, \boldsymbol{\theta}|z)d\boldsymbol{\theta} \\ &= \int \pi(\boldsymbol{Y}|\boldsymbol{\theta}, z)\pi(\boldsymbol{\theta}|z)d\boldsymbol{\theta}\end{aligned}$$

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

$$\boldsymbol{Y}^{(t)} \sim \pi(Y|\boldsymbol{\theta}^{(t)}, z)$$

will be distributed as $\pi(\boldsymbol{Y}|z)$, as desired.

What about predicting at an unknown location s_0?

- We need to find the predictive distribution

$$\pi(Z(\boldsymbol{s}_0)|\boldsymbol{z},\boldsymbol{\theta}) = \int \pi(Z(\boldsymbol{s}_0),\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{x}(\boldsymbol{s}_0))d\boldsymbol{\theta}$$
$$= \int \pi(Z(\boldsymbol{s}_0)|\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{x}(\boldsymbol{s}_0))\pi(\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{x})d\boldsymbol{\theta}$$

where $\pi(Z(\boldsymbol{s}_0)|\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{x}(\boldsymbol{s}_0))$ is a conditional normal, given the joint multivariate normal structure of $Z(\boldsymbol{s}_0)$ and the data $\boldsymbol{z}$.

What about predicting at an unknown location s_0?

- We need to find the predictive distribution

$$\pi(Z(\boldsymbol{s}_0)|\boldsymbol{z},\boldsymbol{\theta}) = \int \pi(Z(\boldsymbol{s}_0),\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{x}(\boldsymbol{s}_0))d\boldsymbol{\theta}$$
$$= \int \pi(Z(\boldsymbol{s}_0)|\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{x}(\boldsymbol{s}_0))\pi(\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{x})d\boldsymbol{\theta}$$

  where $\pi(Z(\boldsymbol{s}_0)|\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{x}(\boldsymbol{s}_0))$ is a conditional normal, given the joint multivariate normal structure of $Z(\boldsymbol{s}_0)$ and the data $\boldsymbol{z}$.
- If we have samples $\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^T$ from $\pi(\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{x})$, then the predictive integral is computed via a Monte Carlo mixture,

$$\hat{\pi}(Z(\boldsymbol{s}_0)|\boldsymbol{z},\boldsymbol{x},\boldsymbol{x}(\boldsymbol{s}_0)) = \frac{1}{T}\sum_{t=1}^{T}\pi(Z(\boldsymbol{s}_0)|\boldsymbol{z},\boldsymbol{\theta}^{(t)},\boldsymbol{x}(\boldsymbol{s}_0))$$

Sampling to the rescue, again

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

$$z_0^{(t)} \sim \pi(Z(\boldsymbol{s}_0)|\boldsymbol{z}, \boldsymbol{\theta}^{(t)}, \boldsymbol{x}(\boldsymbol{s}_0))$$

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

Sampling to the rescue, again

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

$$z_0^{(t)} \sim \pi(Z(\boldsymbol{s}_0)|\boldsymbol{z}, \boldsymbol{\theta}^{(t)}, \boldsymbol{x}(\boldsymbol{s}_0))$$

  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

$$Z(\boldsymbol{s}) \sim \mathbb{N}(\boldsymbol{X}\boldsymbol{\beta} + Y(\boldsymbol{s}), \tau^2 I)$$
$$Y(\boldsymbol{s}) \sim \mathbb{N}(0, \Sigma_{\sigma^2,\phi})$$

plus potentially further prior information on $\tau^2, \sigma^2, \boldsymbol{\beta}, \phi$, etc.

From continuous to binary data

- In the hierarchical framework, we modeled Gaussian data as

$$Z(\boldsymbol{s}) \sim \mathbb{N}(\boldsymbol{X}\boldsymbol{\beta} + Y(\boldsymbol{s}), \tau^2 I)$$
$$Y(\boldsymbol{s}) \sim \mathbb{N}(0, \Sigma_{\sigma^2,\phi})$$

  plus potentially further prior information on $\tau^2, \sigma^2, \boldsymbol{\beta}, \phi$, 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,

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

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

Connection to Generalized Linear Mixed Models

- Assume our data comes from an exponential family,

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

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

- As before, we assume

$$Y(\boldsymbol{s}) \sim \mathbb{N}(0, \Sigma_{\sigma^2,\phi})$$

  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 $\pi(Z(\boldsymbol{s})|\beta, \sigma^2, \phi, \kappa)$, namely

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

Notes on the GLMM framework

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

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

- Secondly, there's no need to include random (white) noise ϵ because stochastic variability is already included in the specification of $\pi(Z(\mathbf{s})|\beta, Y(\mathbf{s}), \kappa)$