

# BAYESIAN LEARNING - LECTURE 7

Mattias Villani and Per Sidén

**Division of Statistics and Machine Learning  
Department of Computer and Information Science  
Linköping University**

# LECTURE OVERVIEW

- ▶ Monte Carlo simulation and random number generation
- ▶ Gibbs sampling
- ▶ Data augmentation
  - ▶ Probit regression
  - ▶ Mixture models
- ▶ Regularized regression revisited

# Monte Carlo Sampling

- If  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)}$  is an **iid sequence** from a distribution  $p(\theta)$ , then

$$\frac{1}{N} \sum_{t=1}^N \theta^{(t)} \rightarrow E(\theta)$$
$$\frac{1}{N} \sum_{t=1}^N g(\theta^{(t)}) \rightarrow E[g(\theta)]$$

where  $g(\theta)$  is some well-behaved function.

- Easy to compute **tail probabilities**  $\Pr(\theta \leq c)$  by letting

$$g(\theta) = I(\theta \leq c)$$

and

$$\frac{1}{N} \sum_{t=1}^N g(\theta^{(t)}) = \frac{\# \theta\text{-draws smaller than } c}{N}.$$

# DIRECT SAMPLING BY THE INVERSE CDF METHOD

- ▶ How to **simulate** from a distribution?
- ▶ Let  $f(x)$  be the density function of a stochastic variable. CDF:  $F(x)$ .  
**Inverse CDF method:**
  1. Generate  $u$  from the uniform distribution on  $[0, 1]$ .
  2. Compute  $x = F^{-1}(u)$ .
- ▶ Example 1: **Exponential distribution:**

$$u = F(x) = 1 - \exp(-\lambda x)$$

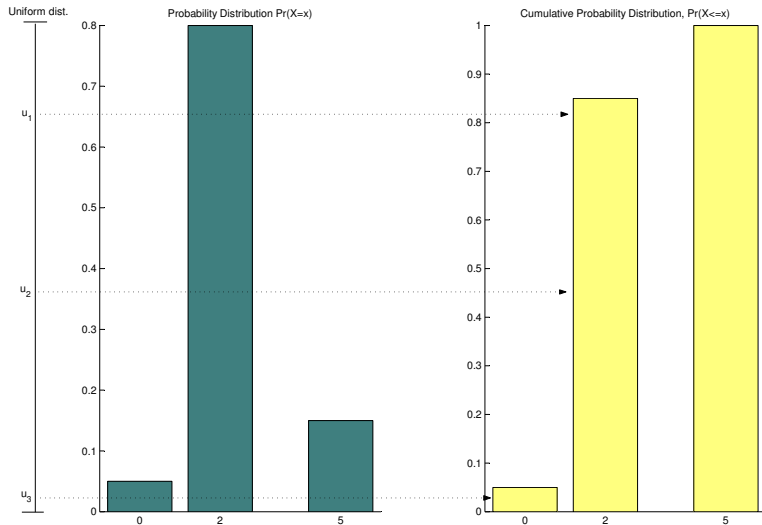
Inverting gives

$$x = -\ln(1 - u) / \lambda$$

But  $1 - u$  is also uniformly distributed on  $[0, 1]$ . So:

- ▶ If  $x = -(\ln u) / \lambda$  where  $u \sim \text{Unif}(0, 1)$ , then  $x \sim \text{Expon}(\lambda)$ .

# INVERSE CDF METHOD, DISCRETE CASE



# DIRECT SAMPLING BY THE INVERSE CDF METHOD

- ▶ Example 2: **Cauchy distribution**:

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$$
$$u = F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan(x)$$

Inverting ...

$$x = \tan[\pi(u - 1/2)].$$

- ▶ We can also use relations between distribution to sample from distributions.
- ▶ Cauchy-example, cont. If  $y$  and  $z$  are independent  $N(0, 1)$  variables, then  $z = \frac{y}{z} \sim \text{Cauchy}$ .
- ▶ Example: **Chi-square**. If  $x_1, \dots, x_v \stackrel{iid}{\sim} N(0, 1)$ , then  $y = \sum_{i=1}^v x_i^2 \sim \chi_v^2$ .

# GIBBS SAMPLING

- ▶ Easily implemented methods for **sampling from multivariate distributions**,  $p(\theta_1, \dots, \theta_k)$ .
- ▶ Requirements: Easily sampled **full conditional posteriors**:
  - ▶  $p(\theta_1|\theta_2, \theta_3, \dots, \theta_k)$
  - ▶  $p(\theta_2|\theta_1, \theta_3, \dots, \theta_k)$
  - ▶  $\vdots$
  - ▶  $p(\theta_k|\theta_1, \theta_2, \dots, \theta_{k-1})$
- ▶ Started out in the early 80's in the image analysis literature.
- ▶ Gibbs sampling is a **special case of Metropolis-Hastings** (see Lecture 8)
- ▶ Metropolis-Hastings is a Markov Chain Monte Carlo (MCMC) algorithm.

# THE GIBBS SAMPLING ALGORITHM

- A:** Choose initial values  $\theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}$ .
- B:**  $B_1$  Draw  $\theta_1^{(1)}$  from  $p(\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_k^{(0)})$   
 $B_2$  Draw  $\theta_2^{(1)}$  from  $p(\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)})$   
 $\vdots$   
 $B_n$  Draw  $\theta_k^{(1)}$  from  $p(\theta_k | \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{k-1}^{(1)})$
- C:** Repeat Step B  $N$  times.



## GIBBS SAMPLING, CONT.

- ▶ The Gibbs draws  $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)}$  are **dependent** (autocorrelated), but **arithmetic means converge to expected values**

$$\frac{1}{N} \sum_{t=1}^N \theta_j^{(t)} \rightarrow E(\theta_j)$$

$$\frac{1}{N} \sum_{t=1}^N g(\theta^{(t)}) \rightarrow E[g(\theta)]$$

- ▶  $\theta^{(1)}, \dots, \theta^{(N)}$  **converges in distribution** to the target  $p(\theta)$ .
- ▶  $\theta_j^{(1)}, \dots, \theta_j^{(N)}$  converge to the marginal distribution of  $\theta_j$ ,  $p(\theta_j)$ .
- ▶ **Dependent** draws  $\rightarrow$  **less efficient** than iid sampling.
- ▶ Compare sampling from:
  - ▶  $x_t \stackrel{iid}{\sim} N(0, \sigma^2)$
  - ▶  $x_t = 0.9x_{t-1} + \varepsilon_t$  with  $\varepsilon_t \stackrel{iid}{\sim} N(0, \sigma^2)$ .

# GIBBS SAMPLING MULTIVARIATE NORMAL

- ▶ Bivariate normal:

- ▶ Joint distribution

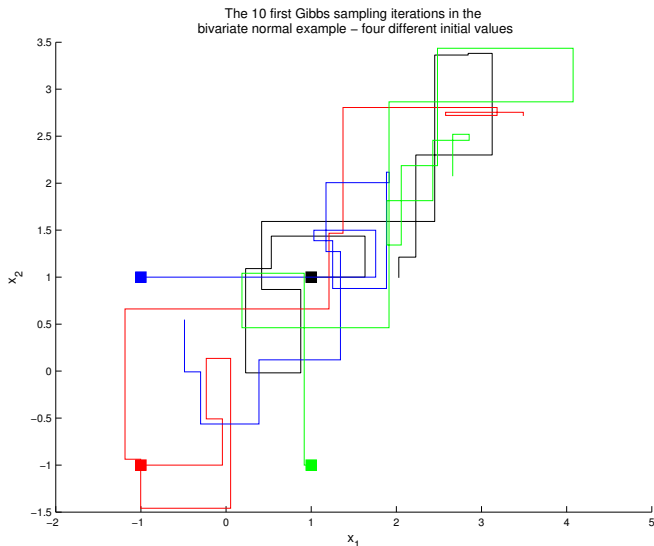
$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim \mathcal{N}_2 \left[ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

- ▶ Full conditional posteriors:

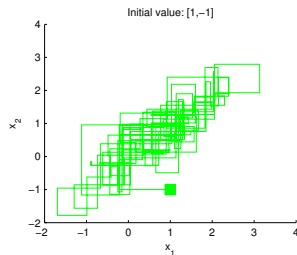
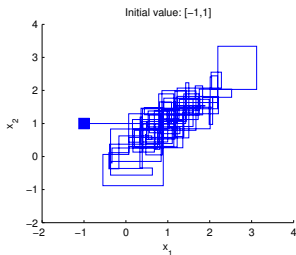
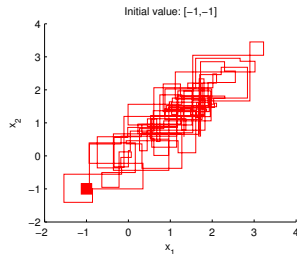
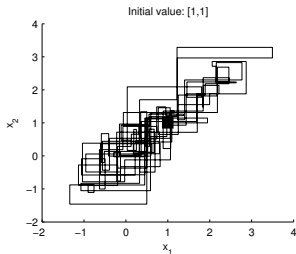
$$\theta_1 | \theta_2 \sim \mathcal{N}[\mu_1 + \rho(\theta_2 - \mu_2), 1 - \rho^2]$$

$$\theta_2 | \theta_1 \sim \mathcal{N}[\mu_2 + \rho(\theta_1 - \mu_1), 1 - \rho^2]$$

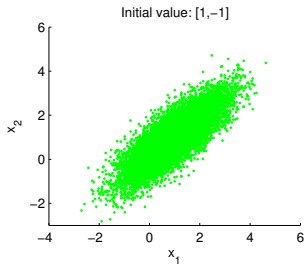
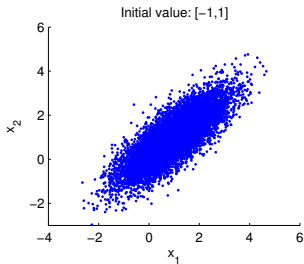
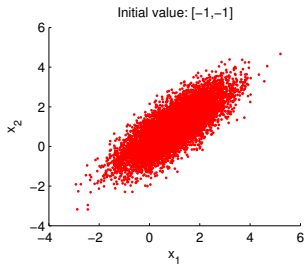
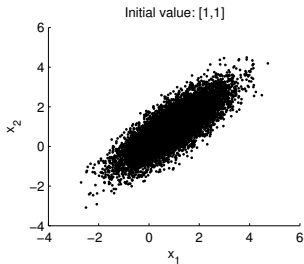
# GIBBS SAMPLING - BIVARIATE NORMAL



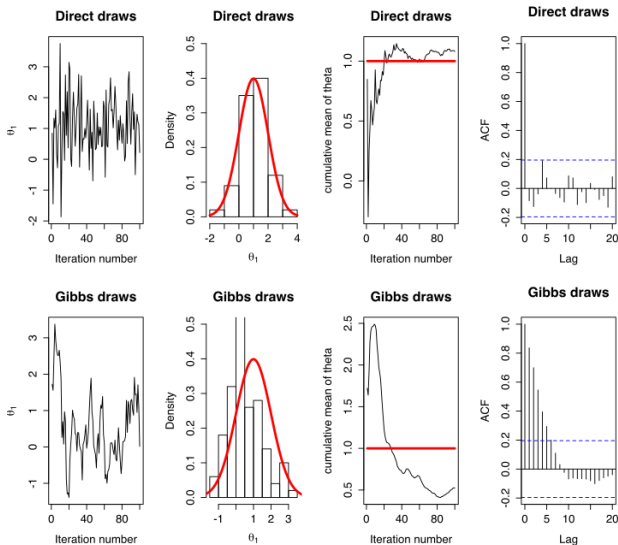
# GIBBS SAMPLING - BIVARIATE NORMAL



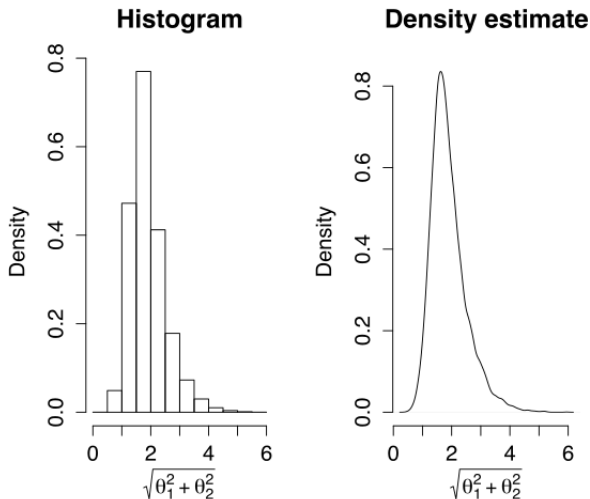
# GIBBS SAMPLING - BIVARIATE NORMAL



# DIRECT SAMPLING VS GIBBS SAMPLING



# ESTIMATING THE DENSITY OF $g(\theta_1, \theta_2) = \sqrt{\theta_1^2 + \theta_2^2}$

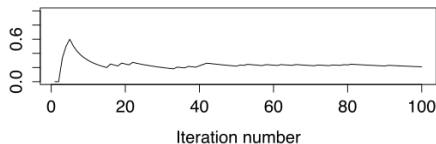


## ESTIMATING $Pr(\theta_1 > 0, \theta_2 > 0)$

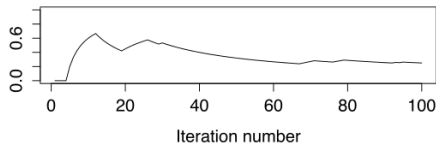
- We can estimate a joint probability by counting:

$$Pr(\theta_1 > 0, \theta_2 > 0) \approx N^{-1} \sum_{i=1}^N 1(\theta_1^{(i)} > 0, \theta_2^{(i)} > 0)$$

**Direct draws**



**Gibbs draws**





# GIBBS SAMPLING FOR NORMAL MODEL WITH NON-CONJUGATE PRIOR

- ▶ Normal model with semi-conjugate prior

$$\begin{aligned}\mu &\sim N(\mu_0, \tau_0^2) \\ \sigma^2 &\sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2)\end{aligned}$$

- ▶ Conditional posteriors

$$\begin{aligned}\mu | \sigma^2, x &\sim N(\mu_n, \tau_n^2) \\ \sigma^2 | \mu, x &\sim \text{Inv} - \chi^2 \left( \nu_n, \frac{\nu_0 \sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2}{n + \nu_0} \right)\end{aligned}$$

with  $\mu_n$  and  $\tau_n^2$  defined the same as when  $\sigma^2$  is known (Lecture 1).

# GIBBS SAMPLING FOR AR PROCESSES

## ► AR( $p$ ) process

$$x_t = \mu + \phi_1(x_{t-1} - \mu) + \dots + \phi_p(x_{t-p} - \mu) + \varepsilon_t, \quad \varepsilon_t \stackrel{iid}{\sim} N(0, \sigma^2).$$

► Let  $\phi = (\phi_1, \dots, \phi_p)'$ .

## ► Prior:

- $\mu \sim \text{Normal}$
- $\phi \sim \text{Multivariate Normal}$
- $\sigma^2 \sim \text{Scaled Inverse } \chi^2$ .

► The **posterior** can be simulated by Gibbs sampling:

- $\mu | \phi, \sigma^2, x \sim \text{Normal}$
- $\phi | \mu, \sigma^2, x \sim \text{Multivariate Normal}$
- $\sigma^2 | \mu, \phi, x \sim \text{Scaled Inverse } \chi^2$

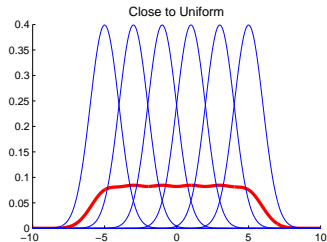
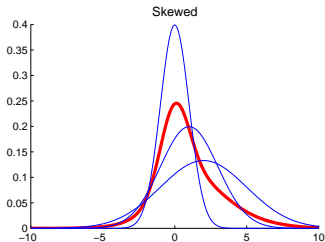
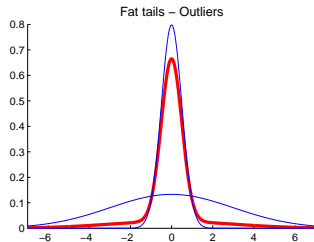
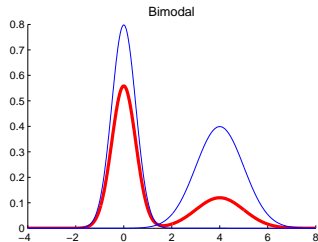
# DATA AUGMENTATION - MIXTURE DISTRIBUTIONS

- ▶ Let  $\phi(x|\mu, \sigma^2)$  denotes the **PDF** of a **normal** variable  $x \sim N(\mu, \sigma^2)$ .
- ▶ **Two-component mixture of normals** [MN(2)]

$$p(x) = \pi \cdot \phi(x|\mu_1, \sigma_1^2) + (1 - \pi) \cdot \phi(x|\mu_2, \sigma_2^2)$$

- ▶ **Simulate** from a MN(2):
  - ▶ Simulate an indicator  $I \in \{1, 2\}$ :  $I \sim \text{Bern}(\pi)$ .
  - ▶ If  $I = 1$ , simulate  $x$  from  $N(\mu_1, \sigma_1^2)$
  - ▶ If  $I = 2$ , simulate  $x$  from  $N(\mu_2, \sigma_2^2)$ .

# ILLUSTRATION OF MIXTURE DISTRIBUTIONS



## MIXTURE DISTRIBUTIONS, CONT.

- ▶ Not easy to estimate directly - the likelihood is a product of sums.
- ▶ **Assume** that we knew which of the two densities each observation came from.

$$I_i = \begin{cases} 1 & \text{if } x_i \text{ came from Density 1} \\ 2 & \text{if } x_i \text{ came from Density 2} \end{cases}.$$

- ▶ Armed with knowledge of  $I_1, \dots, I_n$  it is now easy to estimate  $\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$  by separating the sample according to the  $I$ 's.
- ▶ But we do **not** know  $I_1, \dots, I_n$ !

# GIBBS SAMPLING FOR MIXTURE DISTRIBUTIONS

- ▶ Prior:  $\pi \sim \text{Beta}(\alpha_1, \alpha_2)$ . Conjugate prior for  $(\mu_j, \sigma_j^2)$ , see Lecture 5.
- ▶ Define:  $n_1 = \sum_{i=1}^n (I_i = 1)$  and  $n_2 = n - n_1$ .
- ▶ **Gibbs sampling:**
  - ▶  $\pi \mid \mathbf{l}, \mathbf{x} \sim \text{Beta}(\alpha_1 + n_1, \alpha_2 + n_2)$
  - ▶  $\sigma_1^2 \mid \mathbf{l}, \mathbf{x} \sim \text{Inv-}\chi^2(\nu_{n_1}, \sigma_{n_1}^2)$  and  $\mu_1 \mid \mathbf{l}, \sigma_1^2, \mathbf{x} \sim N\left(\mu_{n_1}, \frac{\sigma_1^2}{\kappa_{n_1}}\right)$
  - ▶  $\sigma_2^2 \mid \mathbf{l}, \mathbf{x} \sim \text{Inv-}\chi^2(\nu_{n_2}, \sigma_{n_2}^2)$  and  $\mu_2 \mid \mathbf{l}, \sigma_2^2, \mathbf{x} \sim N\left(\mu_{n_2}, \frac{\sigma_2^2}{\kappa_{n_2}}\right)$
  - ▶  $I_i \mid \pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \mathbf{x} \sim \text{Bern}(\theta_i)$ ,  $i = 1, \dots, n$ ,

$$\theta_i = \frac{(1 - \pi)\phi(x_i; \mu_2, \sigma_2^2)}{\pi\phi(x_i; \mu_1, \sigma_1^2) + (1 - \pi)\phi(x_i; \mu_2, \sigma_2^2)}.$$

# GIBBS SAMPLING FOR MIXTURE DISTRIBUTIONS

- ▶  **$K$ -component mixture of normals**

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \phi(\mathbf{x}; \mu_k, \sigma_k^2),$$

where  $\sum_{k=1}^K \pi_k = 1$ .

- ▶ **Multi-class indicators:**  $I_i = k$  if observation  $i$  comes from density  $k$ .

- ▶ **Gibbs sampling** with

- ▶  $(\pi_1, \dots, \pi_K) \mid \mathbf{I}, \mathbf{x} \sim \text{Dirichlet}(\alpha_1 + n_1, \alpha_2 + n_2, \dots, \alpha_K + n_K)$
- ▶  $\sigma_k^2 \mid \mathbf{I}, \mathbf{x} \sim \text{Inv-}\chi^2$  and  $\mu_k \mid \mathbf{I}, \sigma_k^2, \mathbf{x} \sim \text{Normal}$ , for  $k = 1, \dots, K$ ,
- ▶  $I_i \mid \pi, \mu, \sigma^2, \mathbf{x} \sim \text{Multinomial}(\theta_{i1}, \dots, \theta_{iK})$ , for  $i = 1, \dots, n$ ,

$$\theta_{ij} = \frac{\pi_j \phi(\mathbf{x}_i; \mu_j, \sigma_j^2)}{\sum_{r=1}^K \pi_r \phi(\mathbf{x}_i; \mu_r, \sigma_r^2)}.$$

- ▶ Gibbs sampling is very powerful for **missing data** problems.  
**Semi-supervised learning.**

# DATA AUGMENTATION - PROBIT REGRESSION

- **Probit** model:

$$\Pr(y_i = 1 \mid x_i) = \Phi(x_i^T \beta)$$

- **Random utility formulation** of the probit:

$$\begin{aligned} u_i &\sim N(x_i^T \beta, 1) \\ y_i &= \begin{cases} 1 & \text{if } u_i > 0 \\ 0 & \text{if } u_i \leq 0 \end{cases} . \end{aligned}$$

- Check:  $\Pr(y_i = 1 \mid x_i) = \Pr(u_i > 0) = 1 - \Pr(u_i \leq 0) = 1 - \Pr(u_i - x_i^T \beta < -x_i^T \beta) = 1 - \Phi(-x_i^T \beta) = \Phi(x_i^T \beta)$ .
- If  $u = (u_1, \dots, u_n)$  were observed, then  $\beta$  could be analyzed by traditional linear regression. But,  $u$  is **not observed**. Gibbs sampling to the rescue!



# GIBBS SAMPLING FOR THE PROBIT REGRESSION

- ▶ Simulate from joint posterior  $p(u, \beta|y)$  iterating between the **full conditional posteriors**:
  - ▶  $p(\beta|u, y)$ , which is multivariate normal (this is just a linear regression)
  - ▶  $p(u_i|\beta, y)$ ,  $i = 1, \dots, n$ .
- ▶ The full conditional posterior distribution of  $u_i$  is:

$$\begin{aligned} p(u_i|\beta, y) &\propto p(y_i|\beta, u_i)p(u_i|\beta) \\ &= \begin{cases} N(u_i|x_i'\beta, 1) & \text{truncated to } u_i \in (-\infty, 0] \text{ if } y_i = 0 \\ N(u_i|x_i'\beta, 1) & \text{truncated to } u_i \in (0, \infty) \text{ if } y_i = 1 \end{cases} \end{aligned}$$

- ▶ Collect the  $\beta$ -draws. A histogram of these draws approximates  $p(\beta|y) = \int p(u, \beta|y)du$ .

# REGULARIZED REGRESSION WITH GIBBS

- ▶ Recap: The joint posterior of  $\beta$ ,  $\sigma^2$  and  $\lambda$  is

$$\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X} \sim N(\mu_n, \Omega_n^{-1})$$

$$\sigma^2|\lambda, \mathbf{y}, \mathbf{X} \sim \text{Inv} - \chi^2(\nu_n, \sigma_n^2)$$

$$p(\lambda|\mathbf{y}, \mathbf{X}) \propto \sqrt{\frac{|\Omega_0|}{|\mathbf{X}'\mathbf{X} + \Omega_0|}} \left(\frac{\nu_n \sigma_n^2}{2}\right)^{-\nu_n/2} \cdot p(\lambda)$$

where  $p(\lambda)$  is the  $\text{Inv} - \chi^2$  prior for  $\lambda$ .

- ▶ This is the **conditional-marginal decomposition**

$$p(\beta, \sigma^2, \lambda|\mathbf{y}, \mathbf{X}) = p(\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X})p(\sigma^2|\lambda, \mathbf{y}, \mathbf{X})p(\lambda|\mathbf{y}, \mathbf{X})$$

- ▶ **Gibbs sampling** can instead be used:

- ▶ Sample  $\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X}$  from Normal
- ▶ Sample  $\sigma^2|\beta, \lambda, \mathbf{y}, \mathbf{X}$  from  $\text{Inv} - \chi^2$
- ▶ Sample  $\lambda|\beta, \sigma^2, \mathbf{y}, \mathbf{X}$  from  $\text{Inv} - \chi^2$

- ▶ Note that  $\lambda$  is now **easy** to simulate **once we condition** on  $\beta$  and  $\sigma^2$ .

# IMPROVING THE EFFICIENCY OF THE GIBBS SAMPLER

- ▶ *Efficient blocking*. Correlated parameters should ideally be included in the same updating block.
- ▶ *Reparametrization*. Convergence can improve dramatically in alternative parametrizations.
- ▶ *Data augmentation*. Bring in latent (unobserved) variables that make the full conditional posteriors more easily sampled (Probit, Mixture models etc). Downside: Typically increases the autocorrelation between draws.
- ▶ *Parameter expansion*. Introducing (non-sense) parameters in the model may break the dependence between the original parameters (Example probit).