Bayesian Learning - Lecture 7

Mattias Villani

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

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

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

  where $g(\theta)$ is some well-behaved function.
- Easy to compute **tail probabilities** $\Pr(\theta \leq c)$ by letting

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

  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**:

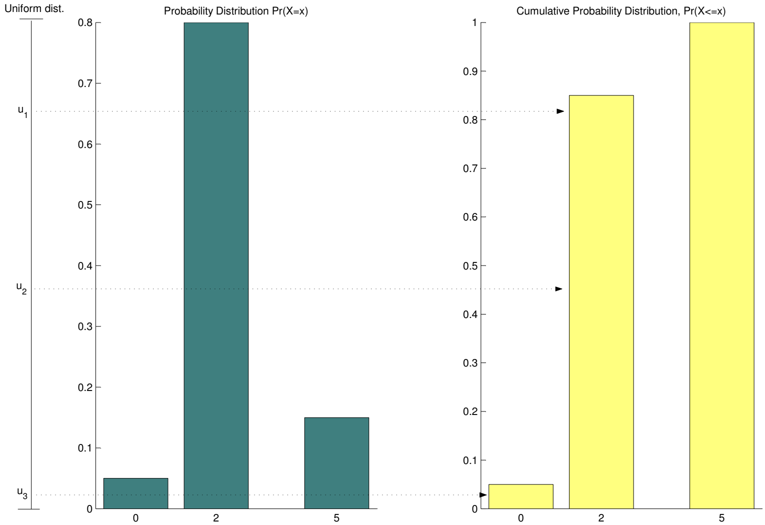
$$u = F(x) = 1 - \exp(-\lambda x)$$

  Inverting gives

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

  But $1 - u$ is also uniformly distributed on [0,1]. So:
- If $x = -(\ln u)/\lambda$ where $u \sim \mathit{Unif}(0, 1)$, then $x \sim \mathit{Expon}(\lambda)$.

Inverse CDF method, discrete case



Direct sampling by the inverse CDF method

- Example 2: **Cauchy distribution**:

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

  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 $z = \frac{y}{z} \sim Cauchy$.
- Example: **Chi-square**. If $x_1, ..., x_v \overset{iid}{\sim} N(0,1)$, then $y = \sum_{i=1}^{v} x_i^2 \sim \chi_v^2$.

GIBBS SAMPLING

- Easily implemented methods for **sampling from multivariate distributions**, $p(\theta_1, ..., \theta_k)$.
- Requirements: Easily sampled **full conditional posteriors**:
  - $p(\theta_1|\theta_2, \theta_3 ..., \theta_k)$
  - $p(\theta_2|\theta_1, \theta_3, ..., \theta_k)$
  - $\vdots$
  - $p(\theta_k|\theta_1, \theta_2, ..., \theta_{k-1})$
- 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 $\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_k^{(0)}$.

B: B_1 Draw $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_k^{(0)})$

B_2 Draw $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_k^{(0)})$

$\vdots$

B_n Draw $\theta_k^{(1)}$ from $p(\theta_n|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{k-1}^{(1)})$

C: Repeat Step B N times.

Gibbs sampling, cont.

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

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

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

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

Gibbs sampling multivariate normal

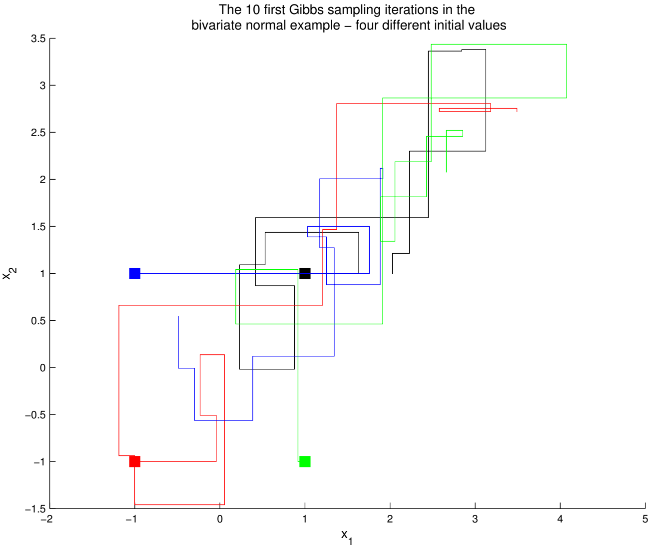
- Bivariate normal:
  - Joint distribution

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N_2\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]$$

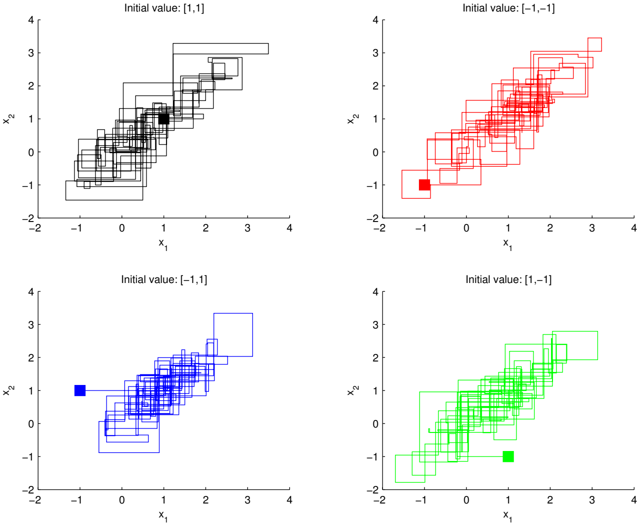
  - Full conditional posteriors:

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

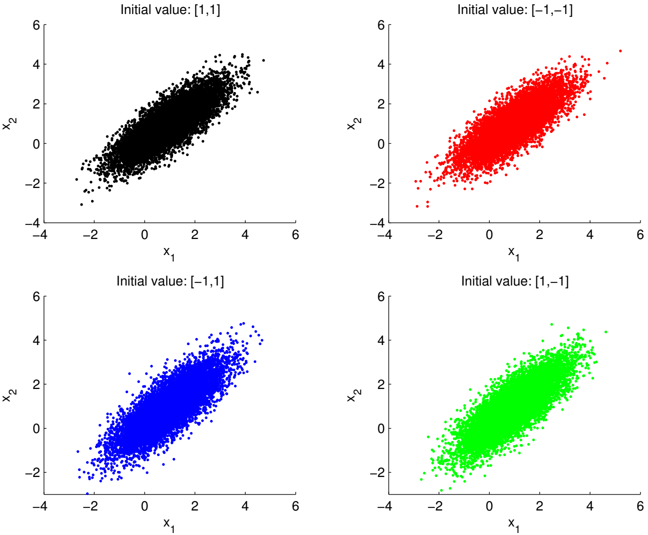
Gibbs sampling - Bivariate normal



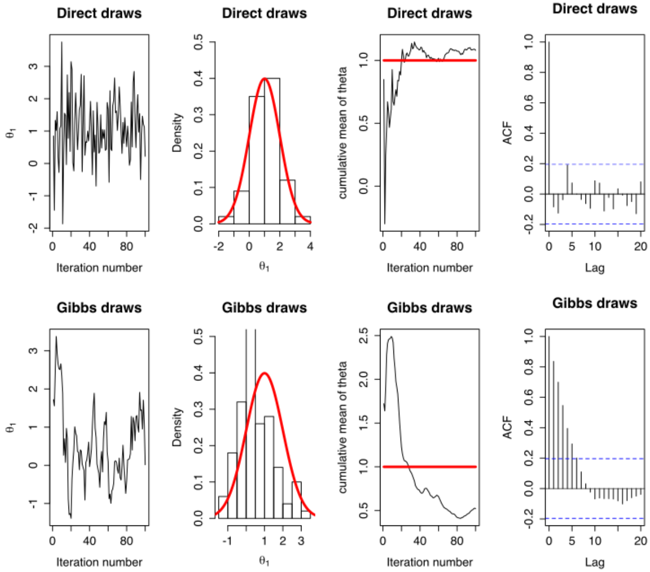
Gibbs sampling - Bivariate normal



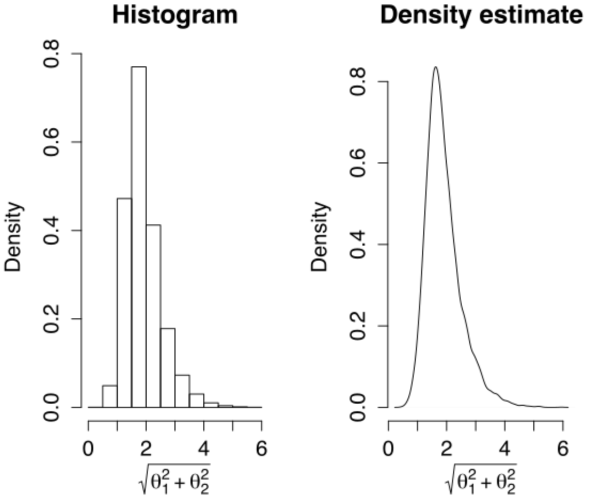
Gibbs sampling - Bivariate normal



Direct sampling vs Gibbs sampling



Estimating the density of $g(\theta_1, \theta_2) = \sqrt{\theta_1^2 + \theta_2^2}$

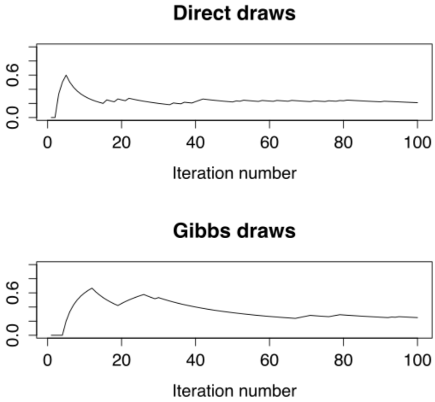


ESTIMATING $Pr(\theta_1 > 0, \theta_2 > 0)$

- We can estimate a joint probability by counting:

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

.



Gibbs sampling for normal model with non-conjugate prior

- Normal model with semi-conjugate prior

$$\mu \sim N(\mu_0, \tau_0^2)$$
$$\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2)$$

- Conditional posteriors

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

Gibbs sampling for AR processes

- **AR(p) process**

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

- Let $\phi = (\phi_1, ..., \phi_p)'$.
- **Prior**:
  - $\mu \sim$Normal
  - $\phi \sim$Multivariate Normal
  - $\sigma^2 \sim$Scaled Inverse χ^2.
- The **posterior** can be simulated by Gibbs sampling:
  - $\mu|\phi, \sigma^2, x \sim$ Normal
  - $\phi|\mu, \sigma^2, x \sim$ Multivariate Normal
  - $\sigma^2|\mu, \phi, x \sim$ Scaled Inverse χ^2

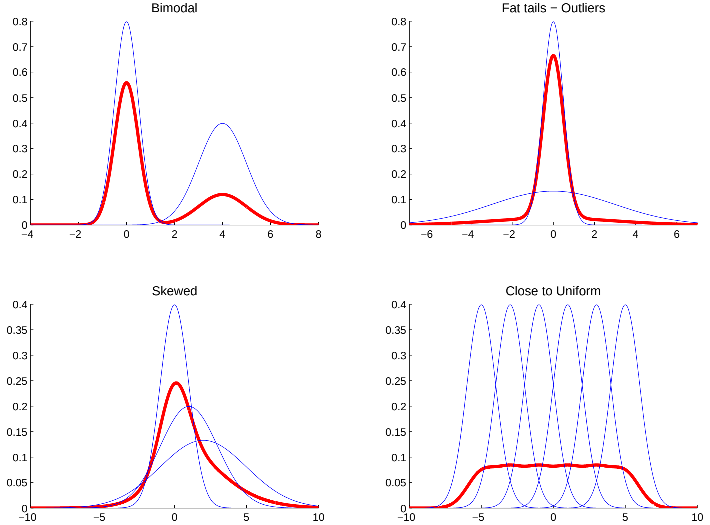
Data augmentation - Mixture distributions

- 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 $I \in \{1, 2\}$: $I \sim Bern(\pi)$.
  - 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 $I_1, ..., I_n$ it is now easy to estimate π, $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ by separating the sample according to the I's.
- But we do **not** know $I_1, ..., I_n$!

GIBBS SAMPLING FOR MIXTURE DISTRIBUTIONS

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

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

Gibbs sampling for mixture distributions

- *K*-component mixture of normals

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

  where $\sum_{k=1}^{K} \pi_k = 1$.
- **Multi-class indicators**: $I_i = k$ if observation i comes from density k.
- **Gibbs sampling** with
  - $(\pi_1, ..., \pi_K) \mid \mathbf{I}, \mathbf{x} \sim \mathit{Dirichlet}(\alpha_1 + n_1, \alpha_2 + n_2, ..., \alpha_K + n_K)$
  - $\sigma_k^2 \mid \mathbf{I}, \mathbf{x} \sim \mathit{Inv}\text{-}\chi^2$ and $\mu_k | \mathbf{I}, \sigma_k^2, \mathbf{x} \sim \mathit{Normal}, \quad \mathit{for}\ k = 1, ..., K,$
  - $I_i \mid \pi, \mu, \sigma^2, \mathbf{x} \sim \mathit{Multinomial}(\theta_{i1}, ..., \theta_{iK}),\ \mathit{for}\ i = 1, ..., n,$

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

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

Data augmentation - Probit regression

- **Probit** model:
$$\Pr(y_i = 1 \mid x_i) = \Phi(x_i'\beta)$$
- **Random utility formulation** of the probit:
$$\begin{aligned} u_i &\sim N(x_i'\beta, 1) \\ y_i &= \begin{cases} 1 & \text{om } u_i > 0 \\ 0 & \text{om } u_i \leq 0 \end{cases}. \end{aligned}$$
- Check: $\Pr(y_i = 1 \mid x_i) = \Pr(u_i > 0) = 1 - \Pr(u_i \leq 0) = 1 - \Pr(u_i - x_i'\beta < -x_i'\beta) = 1 - \Phi(-x_i'\beta) = \Phi(x_i'\beta)$.
- If $u = (u_1, ..., u_n)$ were observed, then β 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)
  - $p(u_i|\beta, y), \; i = 1, ..., n.$
- 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 β-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 β, σ^2 and λ is

$$\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X} \sim N\left(\mu_n, \Omega_n^{-1}\right)$$
$$\sigma^2|\lambda, \mathbf{y}, \mathbf{X} \sim Inv-\chi^2\left(\nu_n, \sigma_n^2\right)$$
$$p(\lambda|\mathbf{y}, \mathbf{X}) \propto \sqrt{\frac{|\Omega_0|}{|X'X + \Omega_0|}} \left(\frac{\nu_n \sigma_n^2}{2}\right)^{-\nu_n/2} \cdot p(\lambda)$$

  where $p(\lambda)$ is the $Inv-\chi^2$ prior for λ.
- 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 Inv-χ^2
  - Sample $\lambda|\beta, \sigma^2, \mathbf{y}, \mathbf{X}$ from Inv-χ^2
- Note that λ is now **easy** to simulate **once we condition** on β and σ^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).