MAST90125: Bayesian Statistical learning

Lecture 18: Data augmentation

Feng Liu and Guoqi Qian

Data augmentation

- Imagine you have specified a likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$ such that, regardless of your choice of prior $p(\boldsymbol{\theta})$, analytic determination of (conditional) posteriors is difficult/impossible.
- Now assume that the joint distribution $p(\mathbf{y}, \boldsymbol{\theta})$ is a marginalisation of the joint distribution $p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta})$

$$p(\mathbf{y}, \boldsymbol{\theta}) = \int_z p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) \mathrm{d}\mathbf{z}.$$

- Sometimes, for an appropriately chosen augmenting variable $\mathbf{z}$, we may find that

$$p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) p(\mathbf{z}|\mathbf{y}) p(\mathbf{y})$$

  can be decomposed such that the conditional posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$ and the posterior of the augmented variable, $p(\mathbf{z}|\mathbf{y})$, can be derived analytically or are easy to find.

Have you previously encountered Data augmentation?

- Where have we used data augmentation before?

Have you previously encountered Data augmentation?

- Where have we used data augmentation before?
    - Probit regression: In order to obtain full conditional posteriors, rather than work with the observed likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$, we used the augmented likelihood, $p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \boldsymbol{\beta}) = p(\mathbf{y}|\mathbf{z}, \mathbf{X}, \boldsymbol{\beta})p(\mathbf{z}|\mathbf{X}, \boldsymbol{\beta})$.

Have you previously encountered Data augmentation?

- Where have we used data augmentation before?
  - Probit regression: In order to obtain full conditional posteriors, rather than work with the observed likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$, we used the augmented likelihood, $p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \boldsymbol{\beta}) = p(\mathbf{y}|\mathbf{z}, \mathbf{X}, \boldsymbol{\beta})p(\mathbf{z}|\mathbf{X}, \boldsymbol{\beta})$.
  - LASSO: In order to obtain full conditional posteriors, rather than work directly with the Laplace prior $p(\beta_j) = \frac{\gamma}{2}e^{-\gamma|\beta_j|}$, we used the augmented prior, $p(\beta_j, \sigma_j^2) = p(\beta_j|\sigma_j^2)p(\sigma_j^2)$.

Have you previously encountered Data augmentation?

- Where have we used data augmentation before?
  - Probit regression: In order to obtain full conditional posteriors, rather than work with the observed likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$, we used the augmented likelihood, $p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \boldsymbol{\beta}) = p(\mathbf{y}|\mathbf{z}, \mathbf{X}, \boldsymbol{\beta})p(\mathbf{z}|\mathbf{X}, \boldsymbol{\beta})$.
  - LASSO: In order to obtain full conditional posteriors, rather than work directly with the Laplace prior $p(\boldsymbol{\beta}_j) = \frac{\gamma}{2}e^{-\gamma|\boldsymbol{\beta}_j|}$, we used the augmented prior, $p(\boldsymbol{\beta}_j, \sigma_j^2) = p(\boldsymbol{\beta}_j|\sigma_j^2)p(\sigma_j^2)$.
- Are we restricted to augmenting just the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, or just the prior $p(\boldsymbol{\theta})$?

Have you previously encountered Data augmentation?

- Where have we used data augmentation before?
  - Probit regression: In order to obtain full conditional posteriors, rather than work with the observed likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$, we used the augmented likelihood, $p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \boldsymbol{\beta}) = p(\mathbf{y}|\mathbf{z}, \mathbf{X}, \boldsymbol{\beta})p(\mathbf{z}|\mathbf{X}, \boldsymbol{\beta})$.
  - LASSO: In order to obtain full conditional posteriors, rather than work directly with the Laplace prior $p(\boldsymbol{\beta}_j) = \frac{\gamma}{2}e^{-\gamma|\boldsymbol{\beta}_j|}$, we used the augmented prior, $p(\boldsymbol{\beta}_j, \sigma_j^2) = p(\boldsymbol{\beta}_j|\sigma_j^2)p(\sigma_j^2)$.
- Are we restricted to augmenting just the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, or just the prior $p(\boldsymbol{\theta})$?
  - Having discussed two previously encountered examples of data augmentation, it is clear that augmentation can be considered for either the likelihood or the prior.

Data augmentation: an example

- To further illustrate data augmentation, consider a Poisson regression.
- If we assume the link is $\eta(\lambda_j) = \log(\lambda_j) = \mathbf{x}_j'\boldsymbol{\beta}$, we know that the likelihood,

$$\Pr(\mathbf{y}|\boldsymbol{\beta}) = \prod_{j=1}^{n} \frac{1}{y_j!} e^{y_j(\mathbf{x}_j'\boldsymbol{\beta})} e^{-e^{\mathbf{x}_j'\boldsymbol{\beta}}},$$

  is not in a form amenable to Gibbs sampling.
- After fitting a Poisson regression, imagine you found evidence for over-dispersion. While your instinct may be to change to a negative binomial likelihood, but you could instead change the representation of the link function.

Data augmentation: an example

- Your new representation is $\eta(\lambda_j) = \log(\lambda_j) = \mathbf{x}_j'\boldsymbol{\beta} + \epsilon_j$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I})$. As a result the distribution $p(y_1 \ldots y_n, \log(\lambda_1), \ldots, \log(\lambda_n), \boldsymbol{\beta}|\sigma^2)$ is

$$\prod_{j=1}^{n} \Pr(y_j|\log(\lambda_j), \boldsymbol{\beta}) \times \prod_{j=1}^{n} \Pr(\log(\lambda_j)|\boldsymbol{\beta}) \times p(\boldsymbol{\beta})$$
$$= \left(\prod_{j=1}^{n} \frac{1}{y_j!} e^{y_j \log(\lambda_j)} e^{-e^{\log(\lambda_j)}} \times (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(\log(\lambda_j) - \mathbf{x}_j\boldsymbol{\beta})^2}{2\sigma^2}} \right) \times p(\boldsymbol{\beta}).$$

- As in Probit regression, if we know λ_i, then Gibbs sampling can be used to determine the posterior distribution of $\boldsymbol{\beta}$.
- Also like Probit regression, the conditional posterior of $\log(\lambda_i)|\boldsymbol{\beta}, y$ can be found element-wise. However unlike Probit regression, this conditional posterior is not well-known in its closed form.

Outlining the algorithm for fitting Poisson-lognormal regression

- We will assume $p(\boldsymbol{\beta}) \propto 1$ and $p(\tau) = \text{Ga}(\alpha, \gamma)$, where $\tau = (\sigma^2)^{-1}$. This means the joint distribution is,

$$p(y_1, \log(\lambda_1), \ldots y_n, \log(\lambda_n), \boldsymbol{\beta}, \tau) = \frac{\gamma^{\alpha}\tau^{\alpha-1}e^{-\gamma\tau}}{\Gamma(\alpha)} \prod_{j=1}^{n} \frac{e^{y_j \log(\lambda_j)} e^{-e^{\log(\lambda_j)}}}{y_j!} \left(\frac{\tau}{2\pi}\right)^{-\frac{1}{2}} e^{-\frac{\tau(\log(\lambda_j)-\mathbf{x}_j\boldsymbol{\beta})^2}{2}}.$$

- The component of the joint distribution that is a function of $\boldsymbol{\beta}$ is,

$$\prod_{j=1}^{n} e^{-\frac{\tau(\log(\lambda_j)-\mathbf{x}_j\boldsymbol{\beta})^2}{2}} = e^{-\frac{\tau(\log(\boldsymbol{\lambda})-\mathbf{X}\boldsymbol{\beta})'(\log(\boldsymbol{\lambda})-\mathbf{X}\boldsymbol{\beta})}{2}} \propto e^{-\frac{\tau\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}}{2}} e^{\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\log(\boldsymbol{\lambda})}.$$

- This implies the conditional posterior of $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\tau, \boldsymbol{\lambda}) = \mathcal{N}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\log(\boldsymbol{\lambda}), (\mathbf{X}'\mathbf{X})^{-1}/\tau).$$

Outlining the algorithm for fitting Poisson-lognormal regression

- The joint distribution is

$$p(y_1, \log(\lambda_1), \ldots y_n, \log(\lambda_n), \boldsymbol{\beta}, \tau) = \frac{\gamma^{\alpha}\tau^{\alpha-1}e^{-\gamma\tau}}{\Gamma(\alpha)} \prod_{j=1}^{n} \frac{e^{y_j \log(\lambda_j)} e^{-e^{\log(\lambda_j)}}}{y_j!} \left(\frac{\tau}{2\pi}\right)^{-\frac{1}{2}} e^{-\frac{\tau(\log(\lambda_j)-\mathbf{x}_j\boldsymbol{\beta})^2}{2}}.$$

- The component of the joint distribution that is a function of τ is,

$$\frac{\gamma^{\alpha}\tau^{\alpha-1}e^{-\gamma\tau}}{\Gamma(\alpha)} \prod_{j=1}^{n} \tau^{-\frac{1}{2}} e^{-\frac{\tau(\log(\lambda)_j-\mathbf{x}_j\boldsymbol{\beta})^2}{2}} = \tau^{\alpha+n/2-1} e^{-\frac{\tau(2\gamma+(\log(\boldsymbol{\lambda})-\mathbf{X}\boldsymbol{\beta})'(\log(\boldsymbol{\lambda})-\mathbf{X}\boldsymbol{\beta}))}{2}}.$$

- This implies the conditional posterior of τ is

$$p(\tau|\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathrm{Ga}(\alpha + n/2, \gamma + (\log(\boldsymbol{\lambda}) - \mathbf{X}\boldsymbol{\beta})'(\log(\boldsymbol{\lambda}) - \mathbf{X}\boldsymbol{\beta})/2).$$

Outlining the algorithm for fitting Poisson-lognormal regression

- The joint distribution is

$$p(y_1, \log(\lambda_1), \dots y_n, \log(\lambda_n), \boldsymbol{\beta}, \tau) = \frac{\gamma^{\alpha}\tau^{\alpha-1}e^{-\gamma\tau}}{\Gamma(\alpha)} \prod_{j=1}^{n} \frac{e^{y_j \log(\lambda_j)}e^{-e^{\log(\lambda_j)}}}{y_j!} \left(\frac{\tau}{2\pi}\right)^{-\frac{1}{2}} e^{-\frac{\tau(\log(\lambda_j)-\mathbf{X}_j\boldsymbol{\beta})^2}{2}}.$$

- The component of the joint distribution that is a function of $\log(\lambda_j)$ is,

$$\frac{e^{y_j \log(\lambda_j)}e^{-e^{\log(\lambda_j)}}}{y_j!} e^{-\frac{\tau(\log(\lambda_j)-\mathbf{X}_j\boldsymbol{\beta})^2}{2}}$$

- This implies the conditional posterior for $\log(\lambda_j)$ would be dependent on $\boldsymbol{\beta}$, y_j and $\mathbf{X}_j$. However the kernel is not in a form where we would recognise the distribution of the posterior. Therefore we will need to use a Metropolis-Hastings step to update this.

Data augmentation: an example

- We will code this example in R. The data consists of 84 lymphocyte counts. These counts were collected from patients on one of 7 dosage levels. The cell log-counts for the patients was also recorded. This information can be downloaded from LMS as `lymphocyte.csv`.
- As already said, Gibbs sampling will be used to sample from the posteriors of $\tau, \boldsymbol{\beta}$
- We will sample from the posterior of $\log(\lambda_j)$ using a Metropolis step, with proposed conditional distribution $J(\log(\lambda_j)^{(t)} | \log(\lambda_j)^{(t-1)}) = \mathcal{N}(\log(\lambda_j)^{(t-1)}, 2.4^2\sigma_j^2)$, where $\sigma_j^2 = 1/(y_j + 0.01)$.

Choosing the parameters of the proposed conditional distribution

- The choice of mean can be justified by a desire for symmetry.

$$\begin{aligned} J(\log(\lambda_j)^{(t)} | \log(\lambda_j)^{(t-1)}) = \mathcal{N}(\log(\lambda_j)^{(t-1)}, \sigma_j^2) &= (2\pi\sigma_j^2)^{-1/2} e^{-\frac{(\log(\lambda_j)^{(t)} - \log(\lambda_j)^{(t-1)})^2}{2\sigma_j^2}} \\ &= \mathcal{N}(\log(\lambda_j)^{(t)}, \sigma_j^2) \\ &= J(\log(\lambda_j)^{(t-1)} | \log(\lambda_j)^{(t)}) \end{aligned}$$

- The justification for the variance chosen for the Metropolis step is as follows,
  - The variance of a univariate function, $\mathrm{Var}(f(x))$ is approximately $f'(x)^2 \mathrm{Var}(x)$.
  - Given a Poisson likelihood for x, we know $E(x) = \mathrm{Var}(x) = \lambda$. Hence an estimator for $\log(\lambda_j)$ is $\log(y_j)$, with $\mathrm{Var}(\log(y_j)) \approx (1/\lambda_j)^2 \lambda_j = 1/\lambda_j$. Since we do not know λ_j, we substitute it with y_j, adding an offset to deal with zero counts.

Inappropriate data augmentation

- As shown in this and previous lectures, data augmentation can be a useful tool for simplifying the process of sampling from the posterior.
- However, any data augmentation strategy must still utilise the hierarchical priors specified.
  - In the lymphocyte example, where observations are Poisson distributed, we know a conjugate prior is Gamma, $\text{Ga}(\alpha, \gamma)$. If you sampled $\lambda_j; 1, \ldots, n$ from the resulting Gamma posterior(s) $\text{Ga}(y_j + \alpha, 1 + \gamma)$, and then sampled $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\lambda_1, \ldots \lambda_n, \tau)$ and τ from $p(\tau|\lambda_1, \ldots \lambda_n, \boldsymbol{\beta})$, this would be inappropriate.
  - Why?

Inappropriate data augmentation

- Let's think about what is being proposed. It looks like we are proposing to cycle between,
  - $\lambda_j \sim p(\lambda_j|y_j); j = 1, \ldots n$.
  - $\boldsymbol{\beta} \sim p(\boldsymbol{\beta}|\lambda_1, \ldots \lambda_n, \tau)$
  - $\tau \sim p(\tau|\lambda_1, \ldots \lambda_n, \boldsymbol{\beta})$
- We know from choosing the link function $\log(\boldsymbol{\lambda}) = \mathbf{X}\beta + \epsilon$, we remove the dependency on $y_1, \ldots y_n$ in the conditional posterior for $\boldsymbol{\beta}, \tau$.
- We know that in the correct augmentation, the conditional posterior of $\log(\lambda_j)$ was dependent on $y_j, \boldsymbol{\beta}$, whereas now we have λ_j dependent on y_j alone.
- However, we know that by the laws of probability, we can write,

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}|y_1, \ldots y_n) = p(\boldsymbol{\beta}|\boldsymbol{\lambda}, y_1, \ldots y_n)p(\boldsymbol{\lambda}|y_1, \ldots y_n)$$

Inappropriate data augmentation

- So the question now becomes is if we marginalise out β to obtain the marginal posterior of λ_j, would we get a posterior $\text{Ga}(y_j + \alpha, 1 + \gamma)$?

Inappropriate data augmentation

- So the question now becomes is if we marginalise out $\boldsymbol{\beta}$ to obtain the marginal posterior of λ_j, would we get a posterior $\text{Ga}(y_j + \alpha, 1 + \gamma)$?
  - The answer is we would not.

- The reason is the Poisson-lognormal model implies a prior for λ_j conditional on $\boldsymbol{\beta}$, which the $\text{Ga}(\alpha, \gamma)$ prior does not take into account.

- What do you think will be the impact of ignoring some of the structure in the prior specifications?

Inappropriate data augmentation

- So the question now becomes is if we marginalise out $\boldsymbol{\beta}$ to obtain the marginal posterior of λ_j, would we get a posterior $\text{Ga}(y_j + \alpha, 1 + \gamma)$?
  - The answer is we would not.
- The reason is the Poisson-lognormal model implies a prior for λ_j conditional on $\boldsymbol{\beta}$, which the $\text{Ga}(\alpha, \gamma)$ prior does not take into account.
- What do you think will be the impact of ignoring some of the structure in the prior specifications?
  - Less precise inference (link to the reason why Gibbs sampling works).
- Note: The distribution $\text{Ga}(y_j + \alpha, 1 + \gamma)$ could be a good proposed conditional distribution in a Metropolis-Hastings algorithm.