Introduction to multiparameter models - part 3

Data analytics

Jerzy Baranowski

Multivariate normal model with known variance

Sometimes we have measurements that are related to each other in a known way

- Multivariate normal likelihood has a vector matrix form
  $y \mid \mu, \Sigma \sim \mathrm{Normal}(\mu, \Sigma)$

$$p(y_1, \ldots, y_n \mid \mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^{\mathsf{T}}\Sigma^{-1}(y_i - \mu)\right)$$

$$|\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S_0)\right)$$

With $S_0 = \sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^{\mathsf{T}}$

Sometimes it is simple

Conjugate prior for μ with known Σ is normal: $\mu \sim \mathrm{Normal}(\mu_0, \Lambda_0)$

- Posterior in such case is

$$
\begin{aligned}
p(\mu \mid y, \Sigma) &\propto \exp\left(-\frac{1}{2}(\mu - \mu_n)^{\mathsf{T}}\Lambda_n^{-1}(\mu - \mu_n)\right) \\
&= \mathrm{Normal}(\mu \mid \mu_n, \Lambda_n) \\
\mu_n &= (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 n\Sigma^{-1}\bar{y}) \\
\Lambda_n^{-1} &= \Lambda_0^{-1} + n\Sigma^{-1}
\end{aligned}
$$

Nuisance μ's can be marginalized without loss of normality

Marginal distributions are normal.

- Marginal distributions of subvectors of μ with known Σ, eg. $\mu^{(1)}$, is also multivariate normal, with mean vector equal to the appropriate subvector of the posterior mean vector μ_n and variance matrix equal to the appropriate submatrix of Λ_n

- Appropriate conditional distribution, assuming $\mu = (\mu^{(1)}, \mu^{(2)})$

$$\mu^{(1)} \,|\, \mu^{(2)}, y \sim \text{Normal}(\mu_n^{(1)} + \beta^{1|2}(\mu^{(2)} - \mu_n^{(2)}), \Lambda^{1|2})$$

$$\beta^{1|2} = \Lambda_n^{(12)} \left(\Lambda_n^{(22)}\right)^{-1}$$

$$\Lambda^{1|2} = \Lambda_n^{(11)} + \Lambda_n^{(12)} \left(\Lambda_n^{(22)}\right)^{-1} \Lambda_n^{(21)}$$

Posterior predictive distribution for known Σ

Surprise! It's also normal!

- We need to observe that the joint distribution $p(\tilde{y}, \mu \mid y) = \text{Normal}(y \mid \mu, \Sigma)\text{Normal}(\mu \mid \mu_n, \Lambda_n)$
- Because of that we can easily compute conditional expectation and variance i.e.

$$
\begin{aligned}
\mathrm{E}(\tilde{y} \mid y) &= \mathrm{E}(\mathrm{E}(\tilde{y} \mid \mu, y) \mid y) \\
&= \mathrm{E}(\mu \mid y) = \mu_n \\
\mathrm{var}(\tilde{y} \mid y) &= \mathrm{E}(\mathrm{var}(\tilde{y} \mid \mu, y) \mid y) + \mathrm{var}(\mathrm{E}(\tilde{y} \mid \mu, y) \mid y) \\
&= \mathrm{E}(\Sigma \mid y) + \mathrm{var}(\mu \mid y) = \Sigma + \Lambda_n
\end{aligned}
$$

Multivariate normal distribution with unknown mean and variance

Here it becomes difficult

- The conjugate prior distribution for (μ, Σ), the normal-inverse-Wishart, is parameterized in terms of hyperparameters $(\mu_0, \Lambda_0/\kappa_0, \nu_0, \Lambda_0)$:

$$p(\mu, \Sigma) \propto |\Sigma|^{\left(-\frac{\nu_0+d}{2}+1\right)} \exp\left(-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^{\mathsf{T}}\Sigma^{-1}(\mu-\mu_0)\right)$$

- Posteriors are of the same family. Noninformative priors are obtained changing number of degrees of freedom

- normal-inverse-Wishart is however a terrible prior, because its parameters are not interpretable and covariance matrices sampled from it are often close to singular.

Instead of giving prior for covariance matrix we can do it for correlation matrix

- This is better, because correlation matrix elements are in $[-1,1]$
- Covariance matrix Σ is related to correlation matrix Ω in the following way

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} \Omega \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}$$

$$\Omega = \begin{bmatrix} 1 & \frac{\sigma_{12}}{\sigma_1\sigma_2} & \frac{\sigma_{13}}{\sigma_1\sigma_3} \\ \frac{\sigma_{12}}{\sigma_1\sigma_2} & 1 & \frac{\sigma_{23}}{\sigma_2\sigma_3} \\ \frac{\sigma_{13}}{\sigma_1\sigma_3} & \frac{\sigma_{23}}{\sigma_2\sigma_3} & 1 \end{bmatrix}$$

LKJ Prior

Recent development - 2009 - Lewandowski-Kurowicka-Joe

- This is a certain generalization of Beta distribution, that fulfills the structural requirements of correlation matrix.

- This is a distribution over positive definite, symmetric matrices with unit diagonal parametrized by $\eta > 0$, with density

$$\mathrm{LkjCorr}(\Omega \,|\, \eta) \propto \det(\Omega)^{(\eta-1)}$$

- In practice we use $\eta \geq 1$, while

    - $\eta = 1$ then the density is uniform over correlation matrices

    - $\eta > 1$ identity matrix is a mode of density, sharper with rising η

LKJ prior for Cholesky factors

Numerical considerations

- There are issues of stability with classical form, we can however use the fact that every positive definite matrix has a Cholesky decomposition i.e.

$$\Omega = LL^\top$$

  where L is lower triangular matrix

- LKJ prior can be reformulated for Cholesky factors, giving density

$$\mathrm{LkjCholesky}(L \mid \eta) \propto |J| \det(LL^\top)^{\eta-1} = \prod_{k=2}^{K} L_{kk}^{K-k+2\eta-2}$$

Covariance estimation example