Lecture 7: Conjugate Priors

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

February 1, 2018

Exponential families have conjugate priors

Examples:Bernoulli model with Beta prior

Examples:Multivariate normal with Normal-Inverse Wishart prior

Example: Poisson distribution

Reading B&S: 5.2, Hoff: 3.3,7.1–3

The posterior $p_{\theta|x_{1:n}}$ in an exponential family

- Exponential family model in canonical form

$$p_X(x;\theta) \;=\; c(x)e^{\theta^T x - \psi(\theta)} \tag{1}$$

Likelihood of sample $x_{1:n}$ with mean $\bar{x} = (\sum_i x_i)/n$

$$p_{x|\theta} \;=\; c(x)e^{x^T\theta - \psi(\theta)} \tag{2}$$

Prior for parameter θ

$$p_\theta(\theta;\mathit{parameters}) \quad \text{the parameters will be specified shortly} \tag{3}$$

By Bayes' rule

$$p_{\theta|x_{1:n}} \;\propto\; C(x_{1:n})e^{n(\bar{x}^T\theta - \psi(\theta)) + \ln p_\theta(\theta;\mathit{parameters})} \tag{4}$$

with $C(x_{1:n}) = \prod_i c(x_i)$.

- Let's look at the exponent

$$n(\underbrace{\bar{x}^T\theta}_{\text{bilinear}} - \underbrace{\psi(\theta)}_{\ln Z(\theta)}) + \ln p_\theta(\theta;\mathit{parameters}) \tag{5}$$

- First two terms look like an exponential family in θ. What would it take to make the posterior be an exponential family?
Answer: $\ln p_\theta(\theta;\text{parameters}) \;=\; \nu_0({\mu_0}^T\theta - \psi(\theta)) + \text{constant}(\nu_0,\mu_0)$.

The conjugate prior

- A prior

$$p_\theta(\theta; \nu_0, \mu_0) \propto \frac{1}{Z(\nu_0, \mu_0)} e^{\nu_0({\mu_0}^T\theta - \psi(\theta))} \tag{6}$$

is called **conjugate prior** for the exponential family defined by (1)

- The normalization constant is

$$Z(\nu_0, \mu_0) = e^{\phi(\nu_0, \mu_0)} = \int_\Theta e^{\nu_0({\mu_0}^T\theta - \psi(\theta))} d\theta \tag{7}$$

- The posterior is now

$$p_{\theta|x_{1:n}} \propto e^{n(\bar{x}^T\theta - \psi(\theta)) + \nu_0({\mu_0}^T\theta - \psi(\theta))} = e^{(n+\nu_0)\left(\frac{n\bar{x}^T + \nu_0{\mu_0}^T}{n+\nu_0}\theta - \psi(\theta)\right) - \phi(\nu_0, \mu_0))} \tag{8}$$

with **hyper-parameters** $\nu = n + \nu_0$, $\mu = \frac{n\bar{x}^T + \nu_0{\mu_0}^T}{n+\nu_0}$ Exercise Why did the factor $C(x_{1:n})$ disappear?

- Hence, the ν parameter behaves like an **equivalent sample size** and the μ parameter like a **mean value parameter** and $\nu\mu$ like a **equivalent sufficient statistic**
- When $n \gg \nu_0$ the influence of the prior becomes neglijible, while for $n \ll \nu_0$, the prior sets the model mean of near μ_0

Bernoulli model with Beta prior

- See Lecture 6.

Multivariate Normal

- The multivariate normal distribution in p dimensions is

$$Normal(x, \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{9}$$

with $x, \mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ positive definite.

Remark When $p_X \propto e^{-\frac{1}{2}X^T AX + b^T X}$, $X \sim Normal(A^{-1}b, A^{-1})$

- Useful to separate prior $p_{\mu,\Sigma} = p_{\mu|\Sigma} p_\Sigma$

The conjugate prior on μ

$$p_\mu(\mu;\ \mu_0, \Lambda_0) \;=\; Normal(\mu; \mu_0, \Lambda_0) \tag{10}$$

- Data sufficient statistics $n, \bar{x}, S$, $S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$ the sample covariance matrix
- The posterior covariance $\Lambda^{-1} = {\Lambda_0}^{-1} + n\Sigma^{-1}$
- Cov^{-1} is called a **precision** matrix.
- The posterior mean $\mu = \Lambda^{-1}({\Lambda_0}^{-1}\mu_0 + n\bar{x})$
- Data affects posterior of μ only via $\bar{x}, n$

Defining prior over Σ

- **Idea** "prior parameter is a sufficient statistic". Hence, conjugate distribution should be the distribution of statistics from $Normal(0, S_0)$
  - Assume $z_{1:\nu_0} \sim$ i.i.d.$N(0, S_0)$, $z_i \in \mathbb{R}^p$.
  - Then $S_{\nu_0} = \sum_{i'=1}^{\nu_0} z_i z_i^T$ is a covariance matrix ($= \nu_0 \times$ sample covar of $z_{1:\nu_0}$), and it is non-singular w.p. 1 for $\nu_0 \geq p$.
  - The distribution of $S_{0\,\nu_0}$ is the **Wishart** distribution
- We set the conjugate prior for Σ^{-1} to be this distribution
- ...and we say Σ is distributed as the **Inverse Wishart**

Wishart and Inverse Wishart

- The Wishart distribution with ν_0 degrees of freedom, over $\mathbb{S}_p^+$ the group of positive definite $p \times p$ matrices

$$p_K(K; \nu_0, S_0) = \frac{1}{2^{\nu_0 p/2} \det S_0 \Gamma_p\left(\frac{\nu_0}{2}\right)} \det K^{(\nu_0 - p - 1)/2} e^{-\frac{1}{2} \mathrm{trace}\, S_0{}^{-1} K} \qquad (11)$$

with $\Gamma_p\left(\frac{\nu_0}{2}\right) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{\nu_0}{2} - \frac{j-1}{2}\right)$

- $E[\Sigma^{-1}] = \nu_0 S_0$
- $E[\Sigma] = \frac{1}{\nu_0 - p - 1} S_0{}^{-1}$
- Posterior parameters $\nu_0 + n$, $S_0{}^{-1} + nS(\mu)$
  - again, posterior parameters and sufficient statistics combine linearly
- Posterior expectation of $\Sigma = \frac{1}{\nu_0 + n - p - 1} [S_0 + nS(\mu)]^{-1}$

Univariate Normal and its conjugate prior

$$p_X(x;\mu,\sigma^2) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{12}$$

- Prior $p_{\mu|\sigma^2,\mu_0,\lambda_0} = Normal(\mu;\mu_0,\lambda_0)$
- Posterior $p_{\mu|x_{1:n},\sigma^2,\mu_0,\lambda_0} = Normal(\mu;\mu,\lambda)$
- Posterior mean $E[\mu] = \frac{1}{\lambda}\left(\frac{1}{\lambda_0}\mu_0 + n\bar{x}\right)$
  - $1/\lambda_0$ is equivalent sample size
- Posterior variance $\frac{1}{\lambda_0} \;=\; \frac{1}{\lambda_0} + n\frac{1}{\sigma^2}$
  - Precision increases with observing data

The Poisson distribution and the Gamma prior

- Poisson distribution $P_X(x) = \frac{1}{x!}e^{-\lambda}\lambda^x = \frac{1}{\Gamma(x-1)}e^{\theta x - e^\theta}$ with $\theta = \ln \lambda$.
- The conjugate prior is then

$$p_{\lambda|\mu} \propto e^{(\theta\mu - e^\theta)\nu} = e^{\theta(\nu\mu)}e^{-\nu e^\theta} \tag{13}$$

- Changing the variable back to λ we have. $d\theta = d\lambda/\lambda$ and

$$p_{\lambda|\nu,\mu} = \lambda^{\nu\mu}e^{-\nu\lambda}\frac{1}{\lambda} \propto gamma(\lambda; \mu\nu, \mu) \tag{14}$$

- Recall that the mean of $gamma(\alpha, \beta)$ is $\frac{\alpha}{\beta}$; hence, $E[\lambda] = \mu$.