

## USEFUL PROPERTIES OF THE MULTIVARIATE NORMAL\*

### 3.1. Conditionals and marginals

For Bayesian analysis it is very useful to understand how to write joint, marginal, and conditional distributions for the multivariate normal.

Given a vector  $x \in \mathbb{R}^p$  the multivariate normal density is

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right).$$

Now split the vector into two parts

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \text{of size } \begin{bmatrix} q \times 1 \\ (p-q) \times 1 \end{bmatrix},$$

and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \text{of size } \begin{bmatrix} q \times q & q \times (p-q) \\ (p-q) \times q & (p-q) \times (p-q) \end{bmatrix}.$$

We now state the joint and marginal distributions

$$x_1 \sim N(\mu_1, \Sigma_{11}), \quad x_2 \sim N(\mu_2, \Sigma_{22}), \quad x \sim N(\mu, \Sigma),$$

and the conditional density

$$x_1 | x_2 \sim N\left(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}\right).$$

The same idea holds for other sizes of partitions.

### 3.2. Conjugate priors

#### 3.2.1. Univariate normals

3.2.1.1. *Fixed variance, random mean.* We consider the parameter  $\sigma^2$  fixed so we are interested in the conjugate prior for  $\mu$ :

$$\pi(\mu | \mu_0, \sigma^2) \propto \frac{1}{\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right),$$

where  $\mu_0$  and  $\sigma^2$  are hyper-parameters for the prior distribution (when we don't have informative prior knowledge we typically consider  $\mu_0 = 0$  and  $\sigma^2$  large).

The posterior distribution for  $x_1, \dots, x_n$  with a univariate normal likelihood and the above prior will be

$$\text{Post}(\mu \mid x_1, \dots, x_n) \sim \text{N}\left(\frac{\sigma_0^2}{\frac{\sigma^2}{n} + \sigma_0^2} \bar{x} + \frac{\sigma^2}{\frac{\sigma^2}{n} + \sigma_0^2} \mu_0, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right).$$

3.2.1.2. *Fixed mean, random variance.* We will formulate this setting with two parameterizations of the scale parameter: (1) the variance  $\sigma^2$ , (2) the precision  $\tau = \frac{1}{\sigma^2}$ .

The two conjugate distributions are the Gamma and the inverse Gamma (really they are the same distribution, just reparameterized)

$$\text{IG}(\alpha, \beta) : f(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} \exp(-\beta(\sigma^2)^{-1}), \quad \text{Ga}(\alpha, \beta) : f(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp(-\beta\tau).$$

The posterior distribution of  $\sigma^2$  is

$$\sigma^2 \mid x_1, \dots, x_n \sim \text{IG}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \mu)^2\right).$$

The posterior distribution of  $\tau$  is not surprisingly

$$\tau \mid x_1, \dots, x_n \sim \text{Ga}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \mu)^2\right).$$

3.2.1.3. *Random mean, random variance.* We now put the previous priors together in what is called a Bayesian hierarchical model:

$$\begin{aligned} x_i \mid \mu, \tau &\stackrel{iid}{\sim} \text{N}(\mu, (\tau)^{-1}) \\ \mu \mid \tau &\sim \text{N}(\mu_0, (\kappa_0 \tau)^{-1}) \\ \tau &\sim \text{Ga}(\alpha, \beta). \end{aligned}$$

For the above likelihood and priors the posterior distribution for the mean and precision is

$$\begin{aligned} \mu \mid \tau, x_1, \dots, x_n &\sim \text{N}\left(\frac{\mu_0 \kappa_0 + n \bar{x}}{n + \kappa_0}, (\tau(n + \kappa_0))^{-1}\right) \\ \tau \mid x_1, \dots, x_n &\sim \text{Ga}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \bar{x})^2 + \frac{n}{n+1} \frac{(\bar{x} - x_i)^2}{2}\right). \end{aligned}$$

### 3.2.2. Multivariate normal

Given a vector  $x \in \mathbb{R}^p$  the multivariate normal density is

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right).$$

We will work with the precision matrix instead of the covariance and we will consider the following Bayesian hierarchical model:

$$\begin{aligned} x_i \mid \mu, \Lambda &\stackrel{iid}{\sim} \text{N}(\mu, (\Lambda)^{-1}) \\ \mu \mid \Lambda &\sim \text{N}(\mu_0, (\kappa_0 \Lambda)^{-1}) \\ \Lambda &\sim \text{Wi}(\Lambda_0, n_0), \end{aligned}$$

the precision matrix is modeled using the Wishart distribution

$$f(\Lambda; V, n) = \frac{|\Lambda|^{(n-d-1)/2} \exp(-.5\text{tr}(\Lambda V^{-1}))}{2^{nd/2} |V|^{n/2} \Gamma_d(n/2)}.$$

For the above likelihood and priors the posterior distribution for the mean and precision is

$$\begin{aligned}\mu \mid \Lambda, x_1, \dots, x_n &\sim \text{N}\left(\frac{\mu_0 \kappa_0 + n \bar{x}}{n + \kappa_0}, (\Lambda(n + \kappa_0))^{-1}\right) \\ \Lambda \mid x_1, \dots, x_n &\sim \text{Wi}\left(n_0 + \frac{n}{2}, \Lambda_0 + \frac{1}{2} \left[ \bar{\Sigma} + \frac{\kappa_0}{\kappa_0 + n} (\bar{x} - \mu_0)(\bar{x} - \mu_0)^T \right]\right).\end{aligned}$$



## LECTURE 4

### A Bayesian approach to linear regression

The main motivations behind a Bayesian formalism for inference are a coherent approach to modeling uncertainty as well as an axiomatic framework for inference. We will reformulate multivariate linear regression from a Bayesian formulation in this section.

Bayesian inference involves thinking in terms of probability distributions and conditional distributions. One important idea is that of a conjugate prior. Another tool we will use extensively in this class is the multivariate normal distribution and its properties.

#### 4.1. Conjugate priors

Given a likelihood function  $p(x | \theta)$  and a prior  $\pi(\theta)$  one can write the posterior as

$$p(\theta | x) = \frac{p(x | \theta)\pi(\theta)}{\int_{\theta'} p(x | \theta')\pi(\theta') d\theta'} = \frac{p(x, \theta)}{p(x)},$$

where  $p(x)$  is the marginal density for the data,  $p(x, \theta)$  is the joint density of the data and the parameter  $\theta$ .

The idea of a prior and likelihood being conjugate is that the prior and the posterior densities belong to the same family. We now state some examples to illustrate this idea.

**Beta, Binomial:** Consider the Binomial likelihood with  $n$  (the number of trials) fixed

$$f(x | p, n) = \binom{n}{x} p^x (1-p)^{n-x},$$

the parameter of interest (the probability of a success) is  $p \in [0, 1]$ . A natural prior distribution for  $p$  is the Beta distribution which has density

$$\pi(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}, \quad p \in (0, 1) \text{ and } \alpha, \beta > 0,$$

where  $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$  is a normalization constant. Given the prior and the likelihood densities the posterior density modulo normalizing constants will take the form

$$\begin{aligned} f(p | x) &\propto \left[ \binom{n}{x} \times \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \right] p^x (1-p)^{n-x} \times p^{\alpha-1} (1-p)^{\beta-1}, \\ &\propto p^{x+\alpha-1} (1-p)^{n-x+\beta-1}, \end{aligned}$$

which means that the posterior distribution of  $p$  is also a Beta with

$$p | x \sim \text{Beta}(\alpha + x, \beta + n - x).$$

**Normal, Normal:** Given a normal distribution with unknown mean the density for the likelihood is

$$f(x | \theta, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(x - \theta)^2\right),$$

and one can specify a normal prior

$$\pi(\theta; \theta_0, \tau_0^2) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \theta_0)^2\right),$$

with hyper-parameters  $\theta_0$  and  $\tau_0$ . The resulting posterior distribution will have the following density function

$$f(\theta | x) \propto \exp\left(-\frac{1}{2\sigma^2}(x - \theta)^2\right) \times \exp\left(-\frac{1}{2\tau_0^2}(\theta - \theta_0)^2\right),$$

which after completing squares and reordering can be written as

$$\theta | x \sim \text{N}(\theta_1, \tau_1^2), \quad \theta_1 = \frac{\frac{\theta_0}{\tau_0^2} + \frac{x}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \quad \tau_1^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}.$$

## 4.2. Bayesian linear regression

We start with the likelihood as

$$f(Y | \mathbf{X}, \beta, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right).$$

and the prior as

$$\pi(\beta) \propto \exp\left(-\frac{1}{2\tau_0^2}\beta^T\beta\right).$$

The density of the posterior is

$$\text{Post}(\beta | D) \propto \left[ \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right) \right] \times \frac{1}{(2\pi)^{p/2}\tau_0^{1/2}} \exp\left(-\frac{1}{2\tau_0^2}\beta^T\beta\right).$$

With a good bit of manipulation the above can be rewritten as a multivariate normal distribution

$$\beta | Y, \mathbf{X}, \sigma^2 \sim \text{N}_p(\mu_1, \Sigma_1)$$

with

$$\Sigma_1 = (\tau_0^{-2}\mathbf{I}_p + \sigma^{-2}\mathbf{X}^T\mathbf{X})^{-1}, \quad \mu_1 = \sigma^{-2}\Sigma_1 \mathbf{X}^T Y.$$

Note the similarities of the above distribution to the MAP estimator. Relate the mean of the above estimator to the MAP estimator.

Predictive distribution: Given data  $D = \{(x_i, y_i)\}_{i=1}^n$  and a new value  $x_*$  one would like to estimate  $y_*$ . This can be done using the posterior and is called the posterior predictive distribution

$$f(y_* | D, x_*, \sigma^2, \tau_0^2) = \int_{\mathbb{R}^p} f(y_* | x_*, \beta, \sigma^2) f(\beta | Y, \mathbf{X}, \sigma^2, \tau_0^2) d\beta,$$

where with some manipulation

$$y_* | D, x_*, \sigma^2, \tau_0^2 \sim \text{N}(\mu_*, \sigma_*^2),$$

where

$$\mu_* = \frac{1}{\sigma^2} \mathbf{1}^T \mathbf{X}^T Y x_*, \quad \sigma_*^2 = \sigma^2 + x_*^T \mathbf{1} x_*.$$

