

# STAT 576 Bayesian Analysis

## Lecture 3: Bayesian Inference II

Chencheng Cai

Washington State University

# Recap: Single Parameter Bayesian Inference

- ▶ Bayesian Inference Procedure:
  - ▶ Name a prior
  - ▶ Get the posterior (proportional notation)
  - ▶ Point estimators: MAP, posterior mean, etc..
  - ▶ Credible interval: QBI, HDR.
  - ▶ Prediction for new observations.

# Recap: Single Parameter Bayesian Inference

- ▶ Bayesian Inference Procedure:
  - ▶ Name a prior
  - ▶ Get the posterior (proportional notation)
  - ▶ Point estimators: MAP, posterior mean, etc..
  - ▶ Credible interval: QBI, HDR.
  - ▶ Prediction for new observations.
- ▶ Prior Elicitation:
  - ▶ Conjugate Prior
  - ▶ Uninformative Prior / Jeffreys Prior
  - ▶ (Improper Prior Distribution)

# Recap: Single Parameter Bayesian Inference

- ▶ Bayesian Inference Procedure:
  - ▶ Name a prior
  - ▶ Get the posterior (proportional notation)
  - ▶ Point estimators: MAP, posterior mean, etc..
  - ▶ Credible interval: QBI, HDR.
  - ▶ Prediction for new observations.
- ▶ Prior Elicitation:
  - ▶ Conjugate Prior
  - ▶ Uninformative Prior / Jeffreys Prior
  - ▶ (Improper Prior Distribution)
- ▶ Important Examples:
  - ▶ Normal with known variance:  $p(\theta) \propto 1$  (conj. prior: Normal)
  - ▶ Normal with known mean:  $p(\sigma^2) \propto (\sigma^2)^{-1}$  (conj. prior: inv-Gamma)

## Nuisance Parameter

- ▶ **Nuisance** parameters are parameters that are unknown and of no interest.

## Nuisance Parameter

- ▶ **Nuisance** parameters are parameters that are unknown and of no interest.
- ▶ Suppose the unknown parameter is  $\theta = (\theta_1, \theta_2)$ .
- ▶ A well-defined observation model gives

$$y \mid \theta_1, \theta_2$$

- ▶ A Bayesian inference needs to define a prior for both  $\theta_1$  and  $\theta_2$ :  $p(\theta_1, \theta_2)$
- ▶ Then the **joint** posterior is obtained by

$$p(\theta_1, \theta_2 \mid y) \propto p(\theta_1, \theta_2)p(y \mid \theta_1, \theta_2)$$

## Nuisance Parameter

- ▶ **Nuisance** parameters are parameters that are unknown and of no interest.
- ▶ Suppose the unknown parameter is  $\theta = (\theta_1, \theta_2)$ .
- ▶ A well-defined observation model gives

$$y \mid \theta_1, \theta_2$$

- ▶ A Bayesian inference needs to define a prior for both  $\theta_1$  and  $\theta_2$ :  $p(\theta_1, \theta_2)$
- ▶ Then the **joint** posterior is obtained by

$$p(\theta_1, \theta_2 \mid y) \propto p(\theta_1, \theta_2)p(y \mid \theta_1, \theta_2)$$

- ▶ If we are only interested in  $\theta_1$ , we need to get the **marginal** posterior for  $\theta_1$ :

$$p(\theta_1 \mid y) = \int p(\theta_1, \theta_2 \mid y) d\mu(\theta_2)$$

## Nuisance Parameter

- An important observation for the marginal posterior is

$$p(\theta_1 \mid y) \propto \int p(\theta_1 \mid \theta_2, y) p(\theta_2 \mid y) d\mu(\theta_2)$$



# Nuisance Parameter

- ▶ An important observation for the marginal posterior is

$$p(\theta_1 \mid y) \propto \int p(\theta_1 \mid \theta_2, y) p(\theta_2 \mid y) d\mu(\theta_2)$$

- ▶ First observation:
  - ▶ In order to draw samples from  $p(\theta_1 \mid y)$
  - ▶ We may first draw  $\theta_2$  from  $p(\theta_2 \mid y)$  (if it is much easier)
  - ▶ Then draw  $\theta_1$  from  $p(\theta_1 \mid \theta_2, y)$  with  $\theta_2$  drawn in the first step.

# Nuisance Parameter

- ▶ An important observation for the marginal posterior is

$$p(\theta_1 | y) \propto \int p(\theta_1 | \theta_2, y) p(\theta_2 | y) d\mu(\theta_2)$$

- ▶ First observation:
  - ▶ In order to draw samples from  $p(\theta_1 | y)$
  - ▶ We may first draw  $\theta_2$  from  $p(\theta_2 | y)$  (if it is much easier)
  - ▶ Then draw  $\theta_1$  from  $p(\theta_1 | \theta_2, y)$  with  $\theta_2$  drawn in the first step.
- ▶ Second observation:
  - ▶ In order to construct a conjugate joint prior
  - ▶ We may find a conjugate prior for the conditional observation model:

$$p(y | \theta_1, \theta_2)$$

with fixed  $\theta_2$

- ▶ Then find a conjugate prior for the marginal observation model:

$$p(y | \theta_2) = \int p(y | \theta_1, \theta_2) p(\theta_1 | \theta_2) d\mu(\theta_1)$$

## Normal with Unknown Mean and Variance

- Suppose we observe

$$y_1, \dots, y_n \sim \mathcal{N}(\mu, \sigma^2), \quad i.i.d.$$

with unknown  $\mu$  and  $\sigma^2$ .

## Normal with Unknown Mean and Variance

- Suppose we observe

$$y_1, \dots, y_n \sim \mathcal{N}(\mu, \sigma^2), \quad i.i.d.$$

with unknown  $\mu$  and  $\sigma^2$ .

- The observation model is

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2}} \exp \left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\} = (\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2} \right\}$$

## Normal with Unknown Mean and Variance

- Suppose we observe

$$y_1, \dots, y_n \sim \mathcal{N}(\mu, \sigma^2), \quad i.i.d.$$

with unknown  $\mu$  and  $\sigma^2$ .

- The observation model is

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2}} \exp \left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\} = (\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2} \right\}$$

- Notice that

$$\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$$

## Normal with Unknown Mean and Variance

- Suppose we observe

$$y_1, \dots, y_n \sim \mathcal{N}(\mu, \sigma^2), \quad i.i.d.$$

with unknown  $\mu$  and  $\sigma^2$ .

- The observation model is

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto \prod_{i=1}^n \frac{1}{\sqrt{\sigma^2}} \exp \left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\} = (\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2} \right\}$$

- Notice that

$$\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$$

- Therefore, we write (with  $s^2 = (n-1)^{-1} \sum_i (y_i - \bar{y})^2$  the sample variance)

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

# Normal with Unkonwn Mean and Variance

- ▶ The score function is

$$\nabla \ell(\mu, \sigma^2) = \left( \begin{array}{c} -\frac{n(\mu - \bar{y})}{\sigma^2} \\ \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2(\sigma^2)^2} - \frac{n}{2\sigma^2} \end{array} \right)$$

- ▶ The Fisher's information ( $2 \times 2$  **matrix**) is

$$\mathcal{I}(\mu, \sigma^2) = -\mathbb{E}_{\mu, \sigma^2}[\Delta \ell(\mu, \sigma)] = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

- ▶ The estimations of  $\mu$  and of  $\sigma^2$  are independent.

## Normal — Uninformative Prior

- ▶ Attemp 1:

- ▶ Since estimating  $\mu$  and  $\sigma^2$  are independent, recall the Uninformative prior:

Normal with Known Variance :  $p(\mu) \propto 1$

Normal with Known Mean :  $p(\sigma^2) \propto 1/\sigma^2$

- ▶ By independence, we construct the following joint prior:

$$p(\mu, \sigma^2) \propto 1/\sigma^2$$

- ▶ The above prior is uniform in  $(\mu, \log \sigma^2)$ .



## Normal — Uninformative Prior

- ▶ Attemp 1:

- ▶ Since estimating  $\mu$  and  $\sigma^2$  are independent, recall the Uninformative prior:

Normal with Known Variance :  $p(\mu) \propto 1$

Normal with Known Mean :  $p(\sigma^2) \propto 1/\sigma^2$

- ▶ By independence, we construct the following joint prior:

$$p(\mu, \sigma^2) \propto 1/\sigma^2$$

- ▶ The above prior is uniform in  $(\mu, \log \sigma^2)$ .

- ▶ Attemp 2:

- ▶ With Jeffreys prior, we define the prior using the Fisher's information by

$$p(\mu, \sigma^2) \propto \sqrt{|\mathcal{I}|} \propto 1/\sigma^3$$

- ▶ The prior is uniform in  $(\mu/\sigma, \log \sigma^2)$ .

## Normal — Uninformative Prior

- ▶ Attemp 1:

- ▶ Since estimating  $\mu$  and  $\sigma^2$  are independent, recall the Uninformative prior:

Normal with Known Variance :  $p(\mu) \propto 1$

Normal with Known Mean :  $p(\sigma^2) \propto 1/\sigma^2$

- ▶ By independence, we construct the following joint prior:

$$p(\mu, \sigma^2) \propto 1/\sigma^2$$

- ▶ The above prior is uniform in  $(\mu, \log \sigma^2)$ .

- ▶ Attemp 2:

- ▶ With Jeffreys prior, we define the prior using the Fisher's information by

$$p(\mu, \sigma^2) \propto \sqrt{|\mathcal{I}|} \propto 1/\sigma^3$$

- ▶ The prior is uniform in  $(\mu/\sigma, \log \sigma^2)$ .

- ▶ **Only the second one is uninformative.**

# Uninformative Prior

- ▶ Jeffreys prior for multiparameter case:

$$p(\theta_1, \dots, \theta_k) \propto \sqrt{|\mathcal{I}(\theta_1, \dots, \theta_k)|}$$

# Uninformative Prior

- ▶ Jeffreys prior for multiparameter case:

$$p(\theta_1, \dots, \theta_k) \propto \sqrt{|\mathcal{I}(\theta_1, \dots, \theta_k)|}$$

- ▶ Reasoning:

- ▶ We assign uniform prior  $p(\theta) \propto 1$  for the case that

$$\mathcal{I}(\theta) \propto \mathbf{I}$$

- ▶ For any bijective continuous mapping  $\lambda = g(\theta)$ , we have

$$\mathcal{I}(\lambda) = \left( \frac{\partial \theta}{\partial \lambda} \right)^T \mathcal{I}(\theta) \left( \frac{\partial \theta}{\partial \lambda} \right)$$

- ▶ This corresponds to the change-of-variable of  $p(\theta)$  to  $\lambda$ :

$$p(\lambda) = p(\theta) \left| \frac{\partial \theta}{\partial \lambda} \right| \propto \sqrt{|\mathcal{I}(\lambda)|}$$

## Normal — Uninformative Prior

- Recall the observation model:

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

- Now we choose the Jeffreys prior as  $p(\mu, \sigma^2) \propto 1/\sigma^3$ .

## Normal — Uninformative Prior

- Recall the observation model:

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

- Now we choose the Jeffreys prior as  $p(\mu, \sigma^2) \propto 1/\sigma^3$ .
- The joint posterior is

$$p(\mu, \sigma^2 \mid y_1, \dots, y_n) \propto (\sigma^2)^{-(n+3)/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

## Normal — Uninformative Prior

- Recall the observation model:

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

- Now we choose the Jeffreys prior as  $p(\mu, \sigma^2) \propto 1/\sigma^3$ .
- The joint posterior is

$$p(\mu, \sigma^2 \mid y_1, \dots, y_n) \propto (\sigma^2)^{-(n+3)/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

- The conditional posterior for  $\mu$  is

$$p(\mu \mid \sigma^2, y_1, \dots, y_n) \sim \mathcal{N}(\bar{y}, \sigma^2/n)$$

## Normal — Uninformative Prior

- ▶ Recall the observation model:

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

- ▶ Now we choose the Jeffreys prior as  $p(\mu, \sigma^2) \propto 1/\sigma^3$ .
- ▶ The joint posterior is

$$p(\mu, \sigma^2 \mid y_1, \dots, y_n) \propto (\sigma^2)^{-(n+3)/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

- ▶ The conditional posterior for  $\mu$  is

$$p(\mu \mid \sigma^2, y_1, \dots, y_n) \sim \mathcal{N}(\bar{y}, \sigma^2/n)$$

- ▶ The conditional posterior for  $\sigma^2$  is

$$p(\sigma^2 \mid \mu, y_1, \dots, y_n) \sim \text{Inv-Gamma}((n+1)/2, [(n-1)s^2 + n(\bar{y} - \mu)^2]/2)$$



## Normal — Uninformative Prior

- The marginal posterior for  $\sigma^2$ :

$$p(\sigma^2 \mid y_1, \dots, y_n) \propto \int p(\mu, \sigma^2 \mid y_1, \dots, y_n) d\mu \propto (\sigma^2)^{-(n+2)/2} \exp \left\{ -\frac{(n-1)s^2}{2\sigma^2} \right\}$$

## Normal — Uninformative Prior

- The marginal posterior for  $\sigma^2$ :

$$p(\sigma^2 \mid y_1, \dots, y_n) \propto \int p(\mu, \sigma^2 \mid y_1, \dots, y_n) d\mu \propto (\sigma^2)^{-(n+2)/2} \exp \left\{ -\frac{(n-1)s^2}{2\sigma^2} \right\}$$

- Or we can take

$$p(\sigma^2 \mid y_1, \dots, y_n) \propto \frac{p(\mu, \sigma^2 \mid y_1, \dots, y_n)}{p(\mu \mid \sigma^2 y_1, \dots, y_n)} \propto (\sigma^2)^{-(n+2)/2} \exp \left\{ -\frac{(n-1)s^2}{2\sigma^2} \right\}$$

## Normal — Uninformative Prior

- The marginal posterior for  $\sigma^2$ :

$$p(\sigma^2 \mid y_1, \dots, y_n) \propto \int p(\mu, \sigma^2 \mid y_1, \dots, y_n) d\mu \propto (\sigma^2)^{-(n+2)/2} \exp \left\{ -\frac{(n-1)s^2}{2\sigma^2} \right\}$$

- Or we can take

$$p(\sigma^2 \mid y_1, \dots, y_n) \propto \frac{p(\mu, \sigma^2 \mid y_1, \dots, y_n)}{p(\mu \mid \sigma^2 y_1, \dots, y_n)} \propto (\sigma^2)^{-(n+2)/2} \exp \left\{ -\frac{(n-1)s^2}{2\sigma^2} \right\}$$

- Therefore,

$$p(\sigma^2 \mid y_1, \dots, y_n) \sim \text{InvGamma}(n/2, (n-1)s^2/2) \sim \text{Scaled-Inv-}\chi^2(n, s^2)$$

## Normal — Uninformative Prior

- ▶ The marginal posterior for  $\sigma^2$ :

$$p(\sigma^2 \mid y_1, \dots, y_n) \propto \int p(\mu, \sigma^2 \mid y_1, \dots, y_n) d\mu \propto (\sigma^2)^{-(n+2)/2} \exp \left\{ -\frac{(n-1)s^2}{2\sigma^2} \right\}$$

- ▶ Or we can take

$$p(\sigma^2 \mid y_1, \dots, y_n) \propto \frac{p(\mu, \sigma^2 \mid y_1, \dots, y_n)}{p(\mu \mid \sigma^2 y_1, \dots, y_n)} \propto (\sigma^2)^{-(n+2)/2} \exp \left\{ -\frac{(n-1)s^2}{2\sigma^2} \right\}$$

- ▶ Therefore,

$$p(\sigma^2 \mid y_1, \dots, y_n) \sim \text{InvGamma}(n/2, (n-1)s^2/2) \sim \text{Scaled-Inv-}\chi^2(n, s^2)$$

- ▶ The densities:

$$\text{InvGamma}(\alpha, \beta) \propto x^{-\alpha-1} e^{-\beta/x}, \quad \text{Scaled-Inv-}\chi^2(\nu, \tau^2) \propto x^{-\nu/2-1} e^{-\nu\tau^2/(2x)}$$

## Normal — Uninformative Prior

- The marginal posterior for  $\mu$  is:

$$\begin{aligned} p(\mu \mid y_1, \dots, y_n) &\propto \frac{p(\mu, \sigma^2 \mid y_1, \dots, y_n)}{p(\sigma^2 \mid \mu^2, y_1, \dots, y_n)} \\ &\propto [(n-1)s^2 + n(\bar{y} - \mu)^2]^{-(n+1)/2} \\ &\propto \left[ 1 + \frac{n(\bar{y} - \mu)^2}{(n-1)s^2} \right]^{-(n+1)/2} \end{aligned}$$

## Normal — Uninformative Prior

- The marginal posterior for  $\mu$  is:

$$\begin{aligned} p(\mu \mid y_1, \dots, y_n) &\propto \frac{p(\mu, \sigma^2 \mid y_1, \dots, y_n)}{p(\sigma^2 \mid \mu^2, y_1, \dots, y_n)} \\ &\propto [(n-1)s^2 + n(\bar{y} - \mu)^2]^{-(n+1)/2} \\ &\propto \left[ 1 + \frac{n(\bar{y} - \mu)^2}{(n-1)s^2} \right]^{-(n+1)/2} \end{aligned}$$

- It follows a noncentral scaled t distribution  $t_n(\bar{y}, (n-1)s^2/n^2)$ .

## Normal — Uninformative Prior

- ▶ The marginal posterior for  $\mu$  is:

$$\begin{aligned} p(\mu \mid y_1, \dots, y_n) &\propto \frac{p(\mu, \sigma^2 \mid y_1, \dots, y_n)}{p(\sigma^2 \mid \mu^2, y_1, \dots, y_n)} \\ &\propto [(n-1)s^2 + n(\bar{y} - \mu)^2]^{-(n+1)/2} \\ &\propto \left[ 1 + \frac{n(\bar{y} - \mu)^2}{(n-1)s^2} \right]^{-(n+1)/2} \end{aligned}$$

- ▶ It follows a noncentral scaled t distribution  $t_n(\bar{y}, (n-1)s^2/n^2)$ .
- ▶ The kernel:

$$t_\nu(\mu, \tau^2) \propto \left[ 1 + \frac{(x - \mu)^2}{\nu \tau^2} \right]^{-(\nu+1)/2}$$

## Normal — Conjugate Prior

- Recall the observation model:

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$



## Normal — Conjugate Prior

- Recall the observation model:

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

- We need some prior is the following form:

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-\alpha} \exp \left\{ -\frac{\beta + \gamma(\mu - \delta)^2}{2\sigma^2} \right\}$$

for some hyperparameters  $(\alpha, \beta, \gamma, \delta)$ .

## Normal — Conjugate Prior

- ▶ Recall the observation model:

$$p(y_1, \dots, y_n \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{2\sigma^2} \right\}$$

- ▶ We need some prior is the following form:

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-\alpha} \exp \left\{ -\frac{\beta + \gamma(\mu - \delta)^2}{2\sigma^2} \right\}$$

for some hyperparameters  $(\alpha, \beta, \gamma, \delta)$ .

- ▶ We observe:
  - ▶  $\mu \mid \sigma^2 \sim \mathcal{N}(\delta, \sigma^2/\gamma)$
  - ▶  $\sigma^2 \mid \mu \sim \text{InvGamma}(\alpha - 1, (\beta + \gamma(\mu - \delta)^2)/2)$
  - ▶  $\sigma^2 \sim \text{InvGamma}(\alpha - 3/2, \beta/2)$
  - ▶  $\mu \sim t_{2\alpha-3}(\delta, \beta/(\gamma(2\alpha - 3)))$

## Normal — Conjugate Prior

- ▶ We found the following combination most convenient:

$$\sigma^2 \sim \text{InvGamma}, \quad \mu \mid \sigma^2 \sim \text{Normal}$$

## Normal — Conjugate Prior

- ▶ We found the following combination most convenient:

$$\sigma^2 \sim \text{InvGamma}, \quad \mu \mid \sigma^2 \sim \text{Normal}$$

- ▶ With a bit change of notation, we define the prior as

$$\sigma^2 \sim \text{InvGamma} \left( \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2} \right) \sim \text{Scaled-Inv-}\chi^2(\nu_0, \sigma_0^2), \quad \mu \mid \sigma^2 \sim \mathcal{N} \left( \mu_0, \frac{\sigma^2}{\kappa_0} \right)$$

## Normal — Conjugate Prior

- ▶ We found the following combination most convenient:

$$\sigma^2 \sim \text{InvGamma}, \quad \mu \mid \sigma^2 \sim \text{Normal}$$

- ▶ With a bit change of notation, we define the prior as

$$\sigma^2 \sim \text{InvGamma} \left( \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2} \right) \sim \text{Scaled-Inv-}\chi^2(\nu_0, \sigma_0^2), \quad \mu \mid \sigma^2 \sim \mathcal{N} \left( \mu_0, \frac{\sigma^2}{\kappa_0} \right)$$

- ▶ This prior is called **Normal-Inverse-Gamma** distribution or **Normal-Inverse- $\chi^2$**  distribution with density:

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-(\nu_0+3)/2} \exp \left\{ -\frac{\nu_0 \sigma_0^2 + \kappa_0 (\mu - \mu_0)^2}{2\sigma^2} \right\}$$

- ▶ N-Inv-Gamma  $\left( \mu_0, \kappa_0, \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2} \right)$  or N-Inv- $\chi^2$   $(\mu_0, \kappa_0, \nu_0, \sigma_0^2)$

## Normal — Conjugate Prior

- ▶ We found the following combination most convenient:

$$\sigma^2 \sim \text{InvGamma}, \quad \mu \mid \sigma^2 \sim \text{Normal}$$

- ▶ With a bit change of notation, we define the prior as

$$\sigma^2 \sim \text{InvGamma} \left( \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2} \right) \sim \text{Scaled-Inv-}\chi^2(\nu_0, \sigma_0^2), \quad \mu \mid \sigma^2 \sim \mathcal{N} \left( \mu_0, \frac{\sigma^2}{\kappa_0} \right)$$

- ▶ This prior is called **Normal-Inverse-Gamma** distribution or **Normal-Inverse- $\chi^2$**  distribution with density:

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-(\nu_0+3)/2} \exp \left\{ -\frac{\nu_0 \sigma_0^2 + \kappa_0 (\mu - \mu_0)^2}{2\sigma^2} \right\}$$

- ▶ N-Inv-Gamma  $\left( \mu_0, \kappa_0, \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2} \right)$  or N-Inv- $\chi^2 \left( \mu_0, \kappa_0, \nu_0, \sigma_0^2 \right)$
- ▶ The Jeffreys prior corresponds to  $\mu_0 = 0 = \kappa_0 = 0 = \nu_0 = 0 = \sigma_0 = 0$

## Normal — Conjugate Prior

The posterior is

$$p(\mu, \sigma^2 \mid y)$$

$$\propto (\sigma^2)^{-(\nu_0+n+3)/2} \exp \left\{ -\frac{\nu_0\sigma_0^2 + (n-1)s^2 + \kappa_0(\mu - \mu_0)^2 + n(\mu - \bar{y})^2}{2\sigma^2} \right\}$$

$$\propto (\sigma^2)^{-(\nu_0+n+3)/2} \exp \left\{ -\frac{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{n+\kappa_0}(\mu_0 - \bar{y})^2 + (\kappa_0 + n) \left( \mu - \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n} \right)^2}{2\sigma^2} \right\}$$

which is N-Inv-Gamma  $\left( \mu_n, \kappa_n, \frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2} \right)$  with

$$\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{n+\kappa_0}(\mu_0 - \bar{y})^2$$

## Normal — Conjugate Prior

$$p(\mu, \sigma^2 \mid y) \sim \text{N-Inv-}\Gamma \left( \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \kappa_0 + n, \frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + (n - 1) s^2 + \frac{n \kappa_0}{n + \kappa_0} (\mu_0 - \bar{y})^2}{2} \right)$$

- Now recall our previous discussion on the marginal/conditional distributions.



## Normal — Conjugate Prior

$$p(\mu, \sigma^2 \mid y) \sim \text{N-Inv-}\Gamma \left( \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \kappa_0 + n, \frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{n+\kappa_0}(\mu_0 - \bar{y})^2}{2} \right)$$

- ▶ Now recall our previous discussion on the marginal/conditional distributions.
- ▶ conditional posterior of  $\mu$ :

$$p(\mu \mid \sigma^2, y) \sim \mathcal{N} \left( \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \frac{\sigma^2}{\kappa_0 + n} \right)$$

## Normal — Conjugate Prior

$$p(\mu, \sigma^2 \mid y) \sim \text{N-Inv-}\Gamma \left( \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \kappa_0 + n, \frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{n+\kappa_0}(\mu_0 - \bar{y})^2}{2} \right)$$

- ▶ Now recall our previous discussion on the marginal/conditional distributions.
- ▶ conditional posterior of  $\mu$ :

$$p(\mu \mid \sigma^2, y) \sim \mathcal{N} \left( \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \frac{\sigma^2}{\kappa_0 + n} \right)$$

- ▶ conditional posterior of  $\sigma^2$ :

$$p(\sigma^2 \mid \mu, y) \sim \text{InvGamma} \left( \frac{\nu_0 + n + 1}{2}, \frac{\nu_0 \sigma_0^2 + (n-1)s^2 + \kappa_0(\mu - \mu_0)^2 + n(\mu - \bar{y})^2}{2} \right)$$

## Normal — Conjugate Prior

$$p(\mu, \sigma^2 \mid y) \sim \text{N-Inv-}\Gamma \left( \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \kappa_0 + n, \frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + (n - 1) s^2 + \frac{n \kappa_0}{n + \kappa_0} (\mu_0 - \bar{y})^2}{2} \right)$$

► marginal posterior of  $\sigma^2$ :

$$p(\sigma^2 \mid y) \sim \text{InvGamma} \left( \frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + (n - 1) s^2 + \frac{n \kappa_0}{n + \kappa_0} (\mu_0 - \bar{y})^2}{2} \right)$$

## Normal — Conjugate Prior

$$p(\mu, \sigma^2 \mid y) \sim \text{N-Inv-}\Gamma \left( \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \kappa_0 + n, \frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + (n - 1) s^2 + \frac{n \kappa_0}{n + \kappa_0} (\mu_0 - \bar{y})^2}{2} \right)$$

- marginal posterior of  $\sigma^2$ :

$$p(\sigma^2 \mid y) \sim \text{InvGamma} \left( \frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + (n - 1) s^2 + \frac{n \kappa_0}{n + \kappa_0} (\mu_0 - \bar{y})^2}{2} \right)$$

- marginal posterior of  $\mu$ :

$$p(\mu \mid y) \sim t_{\nu_0 + n} \left( \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_0 + n}, \frac{\nu_0 \sigma_0^2 + (n - 1) s^2 + \frac{n \kappa_0}{n + \kappa_0} (\mu_0 - \bar{y})^2}{(\nu_0 + n)(\kappa_0 + n)} \right)$$

## Recap

Normal-Inverse-Gamma( $\mu, \lambda, \alpha, \beta$ ):

$$p(x, \sigma^2) \propto (\sigma^2)^{-\alpha-3/2} \exp \left\{ -\frac{2\beta + \lambda(x - \mu)^2}{2\sigma^2} \right\}$$

## Recap

Normal-Inverse-Gamma( $\mu, \lambda, \alpha, \beta$ ):

$$p(x, \sigma^2) \propto (\sigma^2)^{-\alpha-3/2} \exp \left\{ -\frac{2\beta + \lambda(x - \mu)^2}{2\sigma^2} \right\}$$

► conditional  $x \mid \sigma^2$ :

$$x \mid \sigma^2 \sim \mathcal{N} \left( \mu, \frac{\sigma^2}{\lambda} \right)$$

## Recap

Normal-Inverse-Gamma( $\mu, \lambda, \alpha, \beta$ ):

$$p(x, \sigma^2) \propto (\sigma^2)^{-\alpha-3/2} \exp \left\{ -\frac{2\beta + \lambda(x - \mu)^2}{2\sigma^2} \right\}$$

► conditional  $x \mid \sigma^2$ :

$$x \mid \sigma^2 \sim \mathcal{N} \left( \mu, \frac{\sigma^2}{\lambda} \right)$$

► conditional  $\sigma^2 \mid x$ :

$$\sigma^2 \mid x \sim \text{Inv-Gamma} \left( \alpha + \frac{1}{2}, \beta + \frac{\lambda(x - \mu)^2}{2} \right)$$

## Recap

Normal-Inverse-Gamma( $\mu, \lambda, \alpha, \beta$ ):

$$p(x, \sigma^2) \propto (\sigma^2)^{-\alpha-3/2} \exp \left\{ -\frac{2\beta + \lambda(x - \mu)^2}{2\sigma^2} \right\}$$

► conditional  $x \mid \sigma^2$ :

$$x \mid \sigma^2 \sim \mathcal{N} \left( \mu, \frac{\sigma^2}{\lambda} \right)$$

► conditional  $\sigma^2 \mid x$ :

$$\sigma^2 \mid x \sim \text{Inv-Gamma} \left( \alpha + \frac{1}{2}, \beta + \frac{\lambda(x - \mu)^2}{2} \right)$$

► marginal  $x$

$$x \sim t_{2\alpha} \left( \mu, \frac{\beta}{\alpha\lambda} \right)$$



## Recap

Normal-Inverse-Gamma( $\mu, \lambda, \alpha, \beta$ ):

$$p(x, \sigma^2) \propto (\sigma^2)^{-\alpha-3/2} \exp \left\{ -\frac{2\beta + \lambda(x - \mu)^2}{2\sigma^2} \right\}$$

► conditional  $x \mid \sigma^2$ :

$$x \mid \sigma^2 \sim \mathcal{N} \left( \mu, \frac{\sigma^2}{\lambda} \right)$$

► conditional  $\sigma^2 \mid x$ :

$$\sigma^2 \mid x \sim \text{Inv-Gamma} \left( \alpha + \frac{1}{2}, \beta + \frac{\lambda(x - \mu)^2}{2} \right)$$

► marginal  $x$

$$x \sim t_{2\alpha} \left( \mu, \frac{\beta}{\alpha\lambda} \right)$$

► marginal  $\sigma^2$ :

# Multinomial

- **Categorical** distribution:  $y \in \{1, \dots, k\}$  with

$$\mathbb{P}(y = i \mid \boldsymbol{\theta}) = \theta_i \text{ for } i = 1, \dots, k.$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T$  and  $\sum_{i=1}^k \theta_i = 1$ .

# Multinomial

- **Categorical** distribution:  $y \in \{1, \dots, k\}$  with

$$\mathbb{P}(y = i \mid \boldsymbol{\theta}) = \theta_i \text{ for } i = 1, \dots, k.$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T$  and  $\sum_{i=1}^k \theta_i = 1$ .

- **Multinomial** distribution:  $\mathbf{y} \in \mathbb{Z}^k$  with

$$p(\mathbf{y} \mid n, \boldsymbol{\theta}) = \binom{n}{y_1, y_2, \dots, y_k} \prod_{i=1}^k \theta_i^{y_i}$$

for all  $\mathbf{y} = (y_1, \dots, y_k)^T$  such that  $\sum_{i=1}^k y_i = n$  and  $y_i \geq 0 \forall i$ .

# Multinomial

- **Categorical** distribution:  $y \in \{1, \dots, k\}$  with

$$\mathbb{P}(y = i \mid \boldsymbol{\theta}) = \theta_i \text{ for } i = 1, \dots, k.$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T$  and  $\sum_{i=1}^k \theta_i = 1$ .

- **Multinomial** distribution:  $\mathbf{y} \in \mathbb{Z}^k$  with

$$p(\mathbf{y} \mid n, \boldsymbol{\theta}) = \binom{n}{y_1, y_2, \dots, y_k} \prod_{i=1}^k \theta_i^{y_i}$$

for all  $\mathbf{y} = (y_1, \dots, y_k)^T$  such that  $\sum_{i=1}^k y_i = n$  and  $y_i \geq 0 \forall i$ .

- Generalized binomial coefficient:

$$\binom{n}{y_1, y_2, \dots, y_k} = \frac{n!}{y_1! y_2! \cdots y_k!}$$

# Multinomial

- **Categorical** distribution:  $y \in \{1, \dots, k\}$  with

$$\mathbb{P}(y = i \mid \boldsymbol{\theta}) = \theta_i \text{ for } i = 1, \dots, k.$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T$  and  $\sum_{i=1}^k \theta_i = 1$ .

- **Multinomial** distribution:  $\mathbf{y} \in \mathbb{Z}^k$  with

$$p(\mathbf{y} \mid n, \boldsymbol{\theta}) = \binom{n}{y_1, y_2, \dots, y_k} \prod_{i=1}^k \theta_i^{y_i}$$

for all  $\mathbf{y} = (y_1, \dots, y_k)^T$  such that  $\sum_{i=1}^k y_i = n$  and  $y_i \geq 0 \forall i$ .

- Generalized binomial coefficient:

$$\binom{n}{y_1, y_2, \dots, y_k} = \frac{n!}{y_1! y_2! \cdots y_k!}$$

- The categorical distribution is a generalization of Bernoulli distribution.
- The multinomial distribution is a generalization of the Binomial distribution.

# Multinomial

- ▶ Suppose we observe  $\mathbf{y}$  from a multinomial distribution with parameters  $n$  and  $\boldsymbol{\theta}$ .
- ▶ It is immediate that  $n = \sum_{i=1}^k y_i$ . Therefore, the only parameter of interest is  $\boldsymbol{\theta}$ .

# Multinomial

- ▶ Suppose we observe  $\mathbf{y}$  from a multinomial distribution with parameters  $n$  and  $\boldsymbol{\theta}$ .
- ▶ It is immediate that  $n = \sum_{i=1}^k y_i$ . Therefore, the only parameter of interest is  $\boldsymbol{\theta}$ .
- ▶ The likelihood function:

$$p(\mathbf{y} \mid \boldsymbol{\theta}) \propto \prod_{i=1}^k \theta_i^{y_i}$$

- ▶ The conjugate prior can be constructed by

$$p(\boldsymbol{\theta}) \propto \prod_{i=1}^k \theta_i^{\alpha_i - 1}$$

for some  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)^T$ .

- ▶ This prior distribution is known as **Dirichlet** distribution with parameter  $\boldsymbol{\alpha}$ .

# Dirichlet Distribution

$$p(\theta_1, \dots, \theta_k \mid \alpha_1, \dots, \alpha_k) = \frac{1}{\mathbf{B}(\alpha_1, \dots, \alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$$

- The generalized Beta function:

$$\mathbf{B}(\alpha_1, \dots, \alpha_k) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\alpha_0)} \quad \text{with } \alpha_0 = \sum_{i=1}^k \alpha_i$$



# Dirichlet Distribution

$$p(\theta_1, \dots, \theta_k \mid \alpha_1, \dots, \alpha_k) = \frac{1}{\mathbf{B}(\alpha_1, \dots, \alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$$

- The generalized Beta function:

$$\mathbf{B}(\alpha_1, \dots, \alpha_k) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\alpha_0)} \quad \text{with } \alpha_0 = \sum_{i=1}^k \alpha_i$$

- The conditional distribution for  $\theta_1, \dots, \theta_m$  for  $m < k$ :

$$\theta_1, \dots, \theta_m \mid \theta_{m+1}, \dots, \theta_k \sim \text{Dir}(\alpha_1, \dots, \alpha_m) \times \left( 1 - \sum_{i=m+1}^k \theta_i \right)$$

# Dirichlet Distribution

$$p(\theta_1, \dots, \theta_k \mid \alpha_1, \dots, \alpha_k) = \frac{1}{\mathbf{B}(\alpha_1, \dots, \alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$$

- The generalized Beta function:

$$\mathbf{B}(\alpha_1, \dots, \alpha_k) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\alpha_0)} \quad \text{with } \alpha_0 = \sum_{i=1}^k \alpha_i$$

- The conditional distribution for  $\theta_1, \dots, \theta_m$  for  $m < k$ :

$$\theta_1, \dots, \theta_m \mid \theta_{m+1}, \dots, \theta_k \sim \text{Dir}(\alpha_1, \dots, \alpha_m) \times \left(1 - \sum_{i=m+1}^k \theta_i\right)$$

- The marginal distribution for  $\theta_1, \dots, \theta_m$  for  $m < k$ :

$$\theta_1, \dots, \theta_m, \left(1 - \sum_{i=m+1}^k \theta_i\right) \sim \text{Dir}\left(\alpha_1, \dots, \alpha_m, \sum_{i=m+1}^k \alpha_i\right)$$

# Dirichlet Distribution

$$p(\theta_1, \dots, \theta_k \mid \alpha_1, \dots, \alpha_k) = \frac{1}{\mathbf{B}(\alpha_1, \dots, \alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$$

# Dirichlet Distribution

$$p(\theta_1, \dots, \theta_k \mid \alpha_1, \dots, \alpha_k) = \frac{1}{\mathbf{B}(\alpha_1, \dots, \alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$$

- The conditional distribution for  $\theta_1$ :

$$\theta_1 \mid \theta_2, \dots, \theta_k = 1 - \sum_{i=2}^k \theta_i$$

# Dirichlet Distribution

$$p(\theta_1, \dots, \theta_k \mid \alpha_1, \dots, \alpha_k) = \frac{1}{\mathbf{B}(\alpha_1, \dots, \alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$$

- The conditional distribution for  $\theta_1$ :

$$\theta_1 \mid \theta_2, \dots, \theta_k = 1 - \sum_{i=2}^k \theta_i$$

- The marginal distribution for  $\theta_1$ :

$$\theta_1 \sim \text{Beta}(\alpha_1, \alpha_0 - \alpha_0)$$

# Multinomial

- Observation model:

$$p(\mathbf{y} \mid \boldsymbol{\theta}) \propto \prod_{i=1}^k \theta_i^{y_i}$$

- The prior distribution:

$$p(\mathbf{y} \mid \boldsymbol{\alpha}) \propto \prod_{i=1}^k \theta_i^{\alpha_i - 1} \sim \text{Dir}(\boldsymbol{\alpha})$$

# Multinomial

- Observation model:

$$p(\mathbf{y} \mid \boldsymbol{\theta}) \propto \prod_{i=1}^k \theta_i^{y_i}$$

- The prior distribution:

$$p(\mathbf{y} \mid \boldsymbol{\alpha}) \propto \prod_{i=1}^k \theta_i^{\alpha_i - 1} \sim \text{Dir}(\boldsymbol{\alpha})$$

- The posterior distribution:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto \prod_{i=1}^k \theta_i^{\alpha_i + y_i - 1} \sim \text{Dir}(\boldsymbol{\alpha} + \mathbf{y})$$

# Multinomial

Now we consider the uninformative prior.

► Notice that

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \theta_j} = -\frac{y_i}{\theta_i^2} \mathbb{I}\{i = j\}$$



# Multinomial

Now we consider the uninformative prior.

- Notice that

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = -\frac{y_i}{\theta_i^2} \mathbb{I}\{i = j\}$$

- The Fisher's information matrix is

$$\mathcal{I}(\boldsymbol{\theta}) = \text{diag} \left( \frac{n}{\theta_1}, \dots, \frac{n}{\theta_k} \right)$$

# Multinomial

Now we consider the uninformative prior.

- ▶ Notice that

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = -\frac{y_i}{\theta_i^2} \mathbb{I}\{i = j\}$$

- ▶ The Fisher's information matrix is

$$\mathcal{I}(\boldsymbol{\theta}) = \text{diag} \left( \frac{n}{\theta_1}, \dots, \frac{n}{\theta_k} \right)$$

- ▶ The Jeffreys prior is

$$p(\boldsymbol{\theta}) \propto \sqrt{|\mathcal{I}(\boldsymbol{\theta})|} \propto \prod_{i=1}^k \theta_i^{-1/2}$$

- ▶ which corresponds to  $\text{Dir}(1/2, 1/2, \dots, 1/2)$ .

# Multivariate Normal with Known Variance

Multivariate normal  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ :

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

► If we have  $\mathbf{y}_1, \dots, \mathbf{y}_n$  i.i.d. from  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , then

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right\}$$

## Multivariate Normal with Known Variance

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right\}$$

- ▶ Suppose we fix  $\boldsymbol{\Sigma}$ .
- ▶ The conjugate prior is

$$p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}) \propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

## Multivariate Normal with Known Variance

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right\}$$

- ▶ Suppose we fix  $\boldsymbol{\Sigma}$ .
- ▶ The conjugate prior is

$$p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}) \propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

- ▶ The posterior is

$$\begin{aligned} & p(\boldsymbol{\mu} \mid \mathbf{y}_1, \dots, \mathbf{y}_n, \boldsymbol{\Sigma}) \\ & \propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right\} \\ & \sim \mathcal{N} \left( (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1} (\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1} \bar{\mathbf{y}}), (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1} \right) \end{aligned}$$

# Multivariate Normal

Consider the general case with unknown mean and variance:

$$\begin{aligned} & p(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ & \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right\} \\ & \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \right\} \\ & \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S} \boldsymbol{\Sigma}^{-1}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \right\} \end{aligned}$$

with  $\mathbf{S} = \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$  the sum of squares matrix about the sample mean.

# Multivariate Normal

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S} \boldsymbol{\Sigma}^{-1}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \right\}$$

## Multivariate Normal

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \right\}$$

- The conjugate prior would be

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0 + d + 2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}^{-1}) - \frac{\kappa_0}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

where  $d$  is the dimension of  $\boldsymbol{\mu}$ .



# Multivariate Normal

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S} \boldsymbol{\Sigma}^{-1}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \right\}$$

- ▶ The conjugate prior would be

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0 + d + 2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}^{-1}) - \frac{\kappa_0}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

where  $d$  is the dimension of  $\boldsymbol{\mu}$ .

- ▶ This is known as **Normal-Inverse-Wishart** distribution:  $\text{NIW}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda}_0)$ .

# Multivariate Normal

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S} \boldsymbol{\Sigma}^{-1}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \right\}$$

- ▶ The conjugate prior would be

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0 + d + 2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}^{-1}) - \frac{\kappa_0}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

where  $d$  is the dimension of  $\boldsymbol{\mu}$ .

- ▶ This is known as **Normal-Inverse-Wishart** distribution:  $\text{NIW}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda}_0)$ .
- ▶ It is constructed by:

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(\nu_0, \boldsymbol{\Lambda}_0), \quad \boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_0, \kappa_0^{-1} \boldsymbol{\Sigma})$$

# Multivariate Normal

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S} \boldsymbol{\Sigma}^{-1}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \right\}$$

- The conjugate prior would be

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu_0 + d + 2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}^{-1}) - \frac{\kappa_0}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

where  $d$  is the dimension of  $\boldsymbol{\mu}$ .

- This is known as **Normal-Inverse-Wishart** distribution:  $\text{NIW}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda}_0)$ .
- It is constructed by:

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(\nu_0, \boldsymbol{\Lambda}_0), \quad \boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_0, \kappa_0^{-1} \boldsymbol{\Sigma})$$

- The posterior is  $\text{NIW}(\boldsymbol{\mu}_n, \kappa_n, \nu_n, \boldsymbol{\Lambda}_n)$  with

$$\boldsymbol{\mu}_n = \frac{\kappa_0 \boldsymbol{\mu}_0 + n \bar{\mathbf{y}}}{\kappa_0 + n}, \quad \kappa_n = \kappa_0 + n, \quad \nu_n = \nu_0 + n, \quad \boldsymbol{\Lambda}_n = \boldsymbol{\Lambda}_0 + \mathbf{S} + \frac{\kappa_0 n}{\kappa_0 + n} (\boldsymbol{\mu}_0 - \bar{\mathbf{y}})(\boldsymbol{\mu}_0 - \bar{\mathbf{y}})^T.$$

## Normal-Inverse-Wishart Distribution

Consider a NIW( $\boldsymbol{\mu}_0, \kappa, \nu, \boldsymbol{\Lambda}$ ) distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1}) - \frac{\kappa}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

## Normal-Inverse-Wishart Distribution

Consider a NIW( $\boldsymbol{\mu}_0, \kappa, \nu, \boldsymbol{\Lambda}$ ) distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1}) - \frac{\kappa}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

► Conditional of  $\boldsymbol{\mu}$ :

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \propto \mathcal{N}(\boldsymbol{\mu}_0, \kappa^{-1} \boldsymbol{\Sigma})$$

# Normal-Inverse-Wishart Distribution

Consider a NIW( $\boldsymbol{\mu}_0, \kappa, \nu, \boldsymbol{\Lambda}$ ) distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1}) - \frac{\kappa}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

► Conditional of  $\boldsymbol{\mu}$ :

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \propto \mathcal{N}(\boldsymbol{\mu}_0, \kappa^{-1} \boldsymbol{\Sigma})$$

► Conditional of  $\boldsymbol{\Sigma}$ :

$$\boldsymbol{\Sigma} \mid \boldsymbol{\mu} \sim \text{Inv-Wishart}(\nu + 1, \boldsymbol{\Lambda} + \kappa(\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T)$$

# Normal-Inverse-Wishart Distribution

Consider a NIW( $\boldsymbol{\mu}_0, \kappa, \nu, \boldsymbol{\Lambda}$ ) distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1}) - \frac{\kappa}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

- Conditional of  $\boldsymbol{\mu}$ :

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \propto \mathcal{N}(\boldsymbol{\mu}_0, \kappa^{-1} \boldsymbol{\Sigma})$$

- Conditional of  $\boldsymbol{\Sigma}$ :

$$\boldsymbol{\Sigma} \mid \boldsymbol{\mu} \sim \text{Inv-Wishart}(\nu + 1, \boldsymbol{\Lambda} + \kappa(\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T)$$

- Marginal of  $\boldsymbol{\mu}$ :

$$\boldsymbol{\mu} \sim t_{\nu+1-d}(\boldsymbol{\mu}_0, (\nu\kappa)^{-1} \boldsymbol{\Lambda}) \quad (\text{multivariate t distribution})$$

## Normal-Inverse-Wishart Distribution

Consider a NIW( $\boldsymbol{\mu}_0, \kappa, \nu, \boldsymbol{\Lambda}$ ) distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+2)/2} \exp \left\{ -\frac{1}{2} \text{tr}(\boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1}) - \frac{\kappa}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right\}$$

- Conditional of  $\boldsymbol{\mu}$ :

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \propto \mathcal{N}(\boldsymbol{\mu}_0, \kappa^{-1} \boldsymbol{\Sigma})$$

- Conditional of  $\boldsymbol{\Sigma}$ :

$$\boldsymbol{\Sigma} \mid \boldsymbol{\mu} \sim \text{Inv-Wishart}(\nu + 1, \boldsymbol{\Lambda} + \kappa(\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T)$$

- Marginal of  $\boldsymbol{\mu}$ :

$$\boldsymbol{\mu} \sim t_{\nu+1-d}(\boldsymbol{\mu}_0, (\nu\kappa)^{-1} \boldsymbol{\Lambda}) \quad (\text{multivariate t distribution})$$

- Marginal of  $\boldsymbol{\Sigma}$ :

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(\nu, \boldsymbol{\Lambda})$$



## Multivariate Normal — Jeffreys Prior

The log-likelihood function is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \frac{n}{2}(\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{y}}) - \frac{n}{2} \log |\boldsymbol{\Sigma}|$$

## Multivariate Normal — Jeffreys Prior

The log-likelihood function is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \frac{n}{2}(\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{y}}) - \frac{n}{2} \log |\boldsymbol{\Sigma}|$$

- For Fisher's information matrix on  $\boldsymbol{\mu}$ , we have

$$\frac{\partial^2 \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}^T \partial \boldsymbol{\mu}} = -n\boldsymbol{\Sigma}^{-1}$$

## Multivariate Normal — Jeffreys Prior

The log-likelihood function is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \frac{n}{2}(\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{y}}) - \frac{n}{2} \log |\boldsymbol{\Sigma}|$$

- For Fisher's information matrix on  $\boldsymbol{\mu}$ , we have

$$\frac{\partial^2 \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}^T \partial \boldsymbol{\mu}} = -n \boldsymbol{\Sigma}^{-1}$$

- For Fisher's information on the interaction between  $\boldsymbol{\mu}$ ,  $\boldsymbol{\Sigma}$ , we first notice

$$\frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = \frac{1}{2} \boldsymbol{\Sigma}^{-1} \mathbf{S} \boldsymbol{\Sigma}^{-1} + \frac{n}{2} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) (\boldsymbol{\mu} - \bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} - \frac{n}{2} \boldsymbol{\Sigma}^{-1}$$

with its vectorized version:

$$\frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \text{vec}(\boldsymbol{\Sigma})} = \frac{1}{2} \text{vec}(\boldsymbol{\Sigma}^{-1} \mathbf{S} \boldsymbol{\Sigma}^{-1}) + \frac{n}{2} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \otimes \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{y}}) - \frac{n}{2} \text{vec}(\boldsymbol{\Sigma}^{-1})$$

## Multivariate Normal — Jeffreys Prior

- Then we have

$$\frac{\partial^2 \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}^T \partial \text{vec}(\boldsymbol{\Sigma})} = -\frac{n}{2} \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{\mu} - \bar{\mathbf{y}}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \otimes \boldsymbol{\Sigma}^{-1}$$

with its expectation as zero.

## Multivariate Normal — Jeffreys Prior

- ▶ Then we have

$$\frac{\partial^2 \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}^T \partial \text{vec}(\boldsymbol{\Sigma})} = -\frac{n}{2} \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{\mu} - \bar{\mathbf{y}}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \otimes \boldsymbol{\Sigma}^{-1}$$

with its expectation as zero.

- ▶ Furthermore, for  $\boldsymbol{\Sigma}$ , we have (ignoring  $d\boldsymbol{\mu}$ )

$$\begin{aligned} d \frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \text{vec}(\boldsymbol{\Sigma})} &= \frac{1}{2} \text{vec}(d\boldsymbol{\Sigma}^{-1} \mathbf{S}' \boldsymbol{\Sigma}^{-1}) + \frac{1}{2} \text{vec}(\boldsymbol{\Sigma}^{-1} \mathbf{S}' d\boldsymbol{\Sigma}^{-1}) - \frac{n}{2} \text{vec}(d\boldsymbol{\Sigma}^{-1}) \\ &= -\frac{1}{2} (\boldsymbol{\Sigma}^{-1} \mathbf{S}' \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) d\text{vec}(\boldsymbol{\Sigma}) - \frac{1}{2} (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \mathbf{S}' \boldsymbol{\Sigma}^{-1}) d\text{vec}(\boldsymbol{\Sigma}) \\ &\quad + \frac{n}{2} (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) d\text{vec}(\boldsymbol{\Sigma}) \end{aligned}$$

## Multivariate Normal — Jeffreys Prior

- ▶ Then we have

$$\frac{\partial^2 \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}^T \partial \text{vec}(\boldsymbol{\Sigma})} = -\frac{n}{2} \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{\mu} - \bar{\mathbf{y}}) - \frac{n}{2} (\boldsymbol{\mu} - \bar{\mathbf{y}}) \otimes \boldsymbol{\Sigma}^{-1}$$

with its expectation as zero.

- ▶ Furthermore, for  $\boldsymbol{\Sigma}$ , we have (ignoring  $d\boldsymbol{\mu}$ )

$$\begin{aligned} d \frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \text{vec}(\boldsymbol{\Sigma})} &= \frac{1}{2} \text{vec}(d\boldsymbol{\Sigma}^{-1} \mathbf{S}' \boldsymbol{\Sigma}^{-1}) + \frac{1}{2} \text{vec}(\boldsymbol{\Sigma}^{-1} \mathbf{S}' d\boldsymbol{\Sigma}^{-1}) - \frac{n}{2} \text{vec}(d\boldsymbol{\Sigma}^{-1}) \\ &= -\frac{1}{2} (\boldsymbol{\Sigma}^{-1} \mathbf{S}' \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) d\text{vec}(\boldsymbol{\Sigma}) - \frac{1}{2} (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \mathbf{S}' \boldsymbol{\Sigma}^{-1}) d\text{vec}(\boldsymbol{\Sigma}) \\ &\quad + \frac{n}{2} (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) d\text{vec}(\boldsymbol{\Sigma}) \end{aligned}$$

By noticing  $\mathbb{E}[\mathbf{S}'] = \mathbb{E}[\mathbf{S} + n(\boldsymbol{\mu} - \bar{\mathbf{y}})(\boldsymbol{\mu} - \bar{\mathbf{y}})^T] = n\boldsymbol{\Sigma}$ , we have

$$-\mathbb{E} \left[ \frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \text{vec}(\boldsymbol{\Sigma})^T \partial \text{vec}(\boldsymbol{\Sigma})} \right] = \frac{n}{2} \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}$$

## Multivariate Normal — Jeffreys Prior

- So the Fisher's information matrix is

$$\mathcal{I}(\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma})) = \begin{bmatrix} n\boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2}\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \end{bmatrix}$$

## Multivariate Normal — Jeffreys Prior

- So the Fisher's information matrix is

$$\mathcal{I}(\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma})) = \begin{bmatrix} n\boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2}\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \end{bmatrix}$$

- The Jeffreys prior is

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \sqrt{|\mathcal{I}(\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma}))|} \propto |\boldsymbol{\Sigma}|^{-(2d+1)/2}$$



## Multivariate Normal — Jeffreys Prior

- ▶ So the Fisher's information matrix is

$$\mathcal{I}(\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma})) = \begin{bmatrix} n\boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2}\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \end{bmatrix}$$

- ▶ The Jeffreys prior is

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \sqrt{|\mathcal{I}(\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma}))|} \propto |\boldsymbol{\Sigma}|^{-(2d+1)/2}$$

- ▶ Actually, this is **not** the case!!!
- ▶ Reason: variables in  $\mathcal{I}(\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma}))$  are not independent, because  $\boldsymbol{\Sigma}$  has to be symmetric!

## Multivariate Normal — Jeffreys Prior

- ▶ So the Fisher's information matrix is

$$\mathcal{I}(\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma})) = \begin{bmatrix} n\boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2}\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \end{bmatrix}$$

- ▶ The Jeffreys prior is

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \sqrt{|\mathcal{I}(\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma}))|} \propto |\boldsymbol{\Sigma}|^{-(2d+1)/2}$$

- ▶ Actually, this is **not** the case!!!
- ▶ Reason: variables in  $\mathcal{I}(\boldsymbol{\mu}, \text{vec}(\boldsymbol{\Sigma}))$  are not independent, because  $\boldsymbol{\Sigma}$  has to be symmetric!
- ▶ The correct information matrix should only contains the diagonal and upper triangle part of  $\boldsymbol{\Sigma}$ .
- ▶ The **correct** Jeffreys prior:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(d+2)/2}$$