COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Recall: Multivariate Normal Distribution

Notation: $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$, where:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \boldsymbol{\mu}_{\boldsymbol{X}} = E(\boldsymbol{X}) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}, \boldsymbol{\Sigma}_{\boldsymbol{X}} = \sigma^2(\boldsymbol{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}$$

where $\mathrm{cov}\left(X_i, X_j\right) = \sigma_{ij},\ i \neq j$

Multivariate normal density function:

$$f(\boldsymbol{x}) = (2\pi)^{-p/2} \left| \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1/2} \right| e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_X)' \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} (\boldsymbol{X}-\boldsymbol{\mu}_X)}$$

Multivariate Normal Model

Assume that we have multivariate observations

$$\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \mid \boldsymbol{\theta}, \boldsymbol{\Sigma} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$$

Then the likelihood is

$$f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n | \boldsymbol{\theta}, \boldsymbol{\Sigma})$$

$$= \prod_{i=1}^{n} (2\pi)^{-\frac{p}{2}} (\det \boldsymbol{\Sigma})^{-½} e^{-½(\boldsymbol{x}_i - \boldsymbol{\theta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta})}$$

$$= (2\pi)^{-\frac{np}{2}} (\det \boldsymbol{\Sigma})^{-\frac{n}{2}} e^{-½ \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})' \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta})}$$

$$\propto e^{-½(\sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_i + n\boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i)}$$

$$\propto e^{-\frac{1}{2} \sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_i - \frac{1}{2} \boldsymbol{\theta}' \boldsymbol{A}_1 \boldsymbol{\theta} + \boldsymbol{\theta}' \boldsymbol{b}_1}$$

where $\boldsymbol{A}_1 = n\boldsymbol{\Sigma}^{-1}, \boldsymbol{b}_1 = n\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{x}}$

Prior for the mean vector

Let

$$\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$

Then

$$\pi(\theta) = (2\pi)^{-\frac{p}{2}} (\det \boldsymbol{\Lambda}_0)^{-½} e^{-½(\boldsymbol{\theta}-\boldsymbol{\mu}_0)' \boldsymbol{\Lambda}_0{}^{-1} (\boldsymbol{\theta}-\boldsymbol{\mu}_0)}$$
$$= (2\pi)^{-\frac{p}{2}} (\det \boldsymbol{\Lambda}_0)^{-½} e^{-½\boldsymbol{\theta}' \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\mu}_0' \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\theta}' \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0}$$
$$\propto e^{-½\boldsymbol{\theta}' \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\theta} + \boldsymbol{\theta}' \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0}$$
$$= e^{-½\boldsymbol{\theta}' \boldsymbol{A}_0 \boldsymbol{\theta} + \boldsymbol{\theta}' \boldsymbol{b}_0}$$

where $\boldsymbol{A}_0 = \boldsymbol{\Lambda}_0^{-1}, \boldsymbol{b}_0 = \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0$.

Conditional Posterior of $\boldsymbol{\theta}$

The conditional posterior of $\boldsymbol{\theta}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\Sigma}$ is

$$f(\boldsymbol{\theta}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\Sigma}) \propto e^{-½\boldsymbol{\theta}'\boldsymbol{A}_0\boldsymbol{\theta}+\boldsymbol{\theta}'\boldsymbol{b}_0} \times e^{-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{x}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i - \frac{1}{2}\boldsymbol{\theta}'\boldsymbol{A}_1\boldsymbol{\theta}+\boldsymbol{\theta}'\boldsymbol{b}_1}$$

$$\propto e^{-½\boldsymbol{\theta}'\boldsymbol{A}_n\boldsymbol{\theta}+\boldsymbol{\theta}'\boldsymbol{b}_n}$$

where

$$\boldsymbol{A}_n = \boldsymbol{A}_0 + \boldsymbol{A}_1 = \boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1}$$

$$\boldsymbol{b}_n = b_0 + b_1 = \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{x}}$$

The only distribution with such form of the density is the multivariate normal. Therefore,

$$\boldsymbol{\theta}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\Sigma} \sim \mathrm{N}_p(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n)$$

where

$$\boldsymbol{\mu}_n = \boldsymbol{A}_n^{-1}\boldsymbol{b}_n = (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{x}})$$

$$\boldsymbol{\Lambda}_n = \boldsymbol{A}_n^{-1} = (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}$$

Notice the analogy with the univariate case!

Posterior Predictive Distribution

If $\mathbf{\Sigma}$ is known, then it can be shown that

$$\boldsymbol{x}_{new}|\boldsymbol{x} \sim N_p(\boldsymbol{\mu}_n, \Sigma + \boldsymbol{\Lambda}_n)$$

Proof:

$$f(\boldsymbol{x}_{new}|\boldsymbol{x}) = \int f(\boldsymbol{x}_{new}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) f(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\Sigma}) d\boldsymbol{\mu}$$

Convince yourself that only a multivariate normal density can be the answer, and then find the mean and variance with tricks we used before.

Wishart Distribution

- It is a generalization to multidimensions of the Chi-Square distribution.
- The Wishart distribution is a sum of outer products of random vectors.
- It is a *random matrix* which is symmetric and positive definite.
- Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ be independent. Then the distribution of the $p \times p$ random matrix $\boldsymbol{M} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i{}'$ is said to have the Wishart distribution with df = n.
- It defines the distribution of the sample covariance matrix.
- Notation: $\boldsymbol{M} \sim W_p(\boldsymbol{\Sigma}, n)$

Obtaining samples in `R`:
To generate *n* random matrices, distributed according to the Wishart distribution with parameters Σ and df, $W_p(\Sigma, m)$, where m = df.
Use Function: rWishart(n, `df`, Σ)

Wishart Distribution Properties

- Let $\boldsymbol{M} \sim W_p(\boldsymbol{\Sigma}, n)$
- $E(\boldsymbol{M}) = n\boldsymbol{\Sigma}$
- $\boldsymbol{M} \sim \boldsymbol{A}\boldsymbol{W}_p(\boldsymbol{I}_p, n)\boldsymbol{A}'$, where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}'$ is the LU-decomposition
- Assume $n > p$ and $\boldsymbol{\Sigma}$ is invertible. Then the pdf of $\boldsymbol{M}$ is

$$f(\boldsymbol{m}, n, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{m}|^{\frac{n-p-1}{2}} e^{-½\,\mathrm{tr}\left(\boldsymbol{m}\boldsymbol{\Sigma}^{-1}\right)}}{2^{\frac{pn}{2}} \pi^{\frac{p(p-1)}{4}} |\boldsymbol{\Sigma}|^{\frac{n}{2}} \prod_{i=1}^{p} \Gamma\left(\frac{n+1-1}{2}\right)}$$

where the support is all symmetric positive definite matrices $\boldsymbol{m}$.

Aside: Positive Definite Matrix

Any variance matrix must be positive definite, which is the analogy of the univariate variance $\sigma^2 > 0$.

Definition: A symmetric $p \times p$ matrix $\boldsymbol{M}$ is positive definite iff

$$\boldsymbol{x}'\boldsymbol{M}\boldsymbol{x} > 0, \forall\, \boldsymbol{x} \in \mathbb{R}^{\boldsymbol{p}} \backslash 0$$

Definition: If M is a square $p \times p$ matrix, then $\boldsymbol{x} \neq 0$ is an eigenvector of M is there $\exists\, \lambda \in \mathbb{C}$ such that $\boldsymbol{M}\boldsymbol{x} = \lambda \boldsymbol{x}$

Theorem: If $\boldsymbol{M}$ is a symmetric matrix, then all its eigenvalues are real numbers.

Theorem: A symmetric matrix $\boldsymbol{M}$ is positive definite iff all its eigenvalues are positive.

Note: The requirement that the covariance matrix $\boldsymbol{\Sigma}$ is positive definite guarantees that all variances (on the diagonal) are positive and that all correlations are between -1 and 1.

Connection with sample covariance

Let $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n \in \mathbb{R}^p$ be some observations from a multivariate distribution. The sum of squares (SS) *matrix* is defined as

$$\mathrm{SS} = \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i' = \boldsymbol{Z}'\boldsymbol{Z}$$

where $\boldsymbol{Z}$ is an n×p matrix with i^{th} row $\boldsymbol{z}_i'$.

Notes:

- If $n > p$ and the $\boldsymbol{z}_i$'s are *linearly independent*, then SS is a positive definite matrix.
- If we assume the population mean is 0, then SS/n is the MLE of the population covariance matrix.
- The Wishart distribution is constructed in this way and will be positive definite with probability 1 (the chance of linearly dependent observation vectors is 0).

Inverse Wishart Distribution

- In the univariate case, for a conjugate prior, we had to use IG distribution on the population variance σ^2.
- In the multivariate case we need the *Inverse Wishart distribution* as a prior on $\boldsymbol{\Sigma}$.
- We say that $\boldsymbol{M} \sim W_p^{-1}(\boldsymbol{\Sigma}, n)$, if $\boldsymbol{M}^{-1} \sim W_p(\boldsymbol{\Sigma}^{-1}, n)$
- Property: $\mathrm{E}(\boldsymbol{M}) = \frac{\boldsymbol{\Sigma}}{n-p-1}$
- Prior: $\pi(\boldsymbol{\Sigma}) \sim W_p^{-1}(\boldsymbol{S}_0, \nu_0)$ with pdf

$$f(\boldsymbol{\Sigma}) = \left[2^{\frac{\nu_0 p}{2}} \pi^{\frac{\binom{p}{2}}{2}} (\det \boldsymbol{S}_0)^{-\frac{\nu_0}{2}} \prod_{j=1}^{p} \Gamma\left(\frac{\nu_0 + 1 - j}{2}\right)\right]^{-1} \times$$

$$(\det \boldsymbol{\Sigma})^{-\frac{\nu_0+p+1}{2}} e^{-\frac{\mathrm{tr}\left(\boldsymbol{S}_0 \boldsymbol{\Sigma}^{-1}\right)}{2}}$$

Aside: Trace of a matrix

Definition: Trace of a square $p \times p$ matrix $\boldsymbol{M}$ is defined as

$$\mathrm{tr}(\boldsymbol{M}) = \sum_{j=1}^{p} m_{jj}$$

Properties:

- $\mathrm{tr}(\boldsymbol{A} + \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B})$
- $\mathrm{tr}(c\boldsymbol{A}) = c\,\mathrm{tr}(\boldsymbol{A})$
- $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$, where $\boldsymbol{A}$ is $m \times n$ and $\boldsymbol{B}$ is $n \times m$
- If $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are symmetric, then

$$\mathrm{tr}(\boldsymbol{ABC}) = \mathrm{tr}(\boldsymbol{CBA}) = \mathrm{tr}(\boldsymbol{CAB}) = \ldots$$

- Trace is equal to the sum of the eigenvalues.

Back to Multivariate Normal Model

Assume that we have multivariate observations

$$\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \mid \boldsymbol{\theta}, \boldsymbol{\Sigma} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$$

Recall the likelihood is

$$f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n | \boldsymbol{\theta}, \boldsymbol{\Sigma})$$
$$= (2\pi)^{-\frac{np}{2}} (\det \boldsymbol{\Sigma})^{-\frac{n}{2}} e^{-½ \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta})}$$
$$\propto (\det \boldsymbol{\Sigma})^{-\frac{n}{2}} e^{-½ tr\, (\boldsymbol{S}_\theta \boldsymbol{\Sigma}^{-1})}$$

where we used

$$\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta}) = \mathrm{tr} \left[\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\theta}) \right]$$
$$= tr\, (\boldsymbol{S}_\theta \boldsymbol{\Sigma}^{-1})$$

with $\boldsymbol{S}_\theta = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})' (\boldsymbol{x}_i - \boldsymbol{\theta})$.

Conditional Posterior of Σ

The conditional posterior of $\boldsymbol{\Sigma}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\theta}$ is

$$f(\boldsymbol{\Sigma}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\theta}) \propto \pi(\boldsymbol{\Sigma}) \times f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n|\boldsymbol{\theta}, \boldsymbol{\Sigma})$$

$$\propto (\det \boldsymbol{\Sigma})^{-\frac{\nu_0+p+1}{2}} e^{-\frac{\mathrm{tr}\left(\boldsymbol{S}_0\boldsymbol{\Sigma}^{-1}\right)}{2}} \times (\det \boldsymbol{\Sigma})^{-\frac{n}{2}} e^{-½tr\left(\boldsymbol{S}_\theta\boldsymbol{\Sigma}^{-1}\right)}$$

$$(\det \boldsymbol{\Sigma})^{-\frac{\nu_0+n+p+1}{2}} e^{-½tr\left((\boldsymbol{S}_0+\boldsymbol{S}_\theta)\boldsymbol{\Sigma}^{-1}\right)}$$

Therefore,

$$\boldsymbol{\Sigma}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\theta} \sim W_p^{-1}(\boldsymbol{S}_0 + \boldsymbol{S}_\theta, \nu_0 + n)$$

and

$$\mathrm{E}(\boldsymbol{\Sigma}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\theta}) = \frac{\boldsymbol{S}_0 + \boldsymbol{S}_\theta}{\nu_0 + n - p - 1}$$

More importantly, the fact that we know the distributions of each conditional posterior means we can develop a Gibbs sampler to obtain samples from the full joint posterior!

Exercise 1

Consider the following model

$$\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \mid \boldsymbol{\theta}, \boldsymbol{\Sigma} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$$

$$\pi(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto (\det \boldsymbol{\Sigma})^{-\frac{p+1}{2}}$$

This is known as the independence-Jeffreys prior (note that it is improper).

a) Derive the conditional posteriors of $\boldsymbol{\theta}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$

b) Under what condition is the joint posterior proper?

Hint:

$$f(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\theta}) \propto (\det \boldsymbol{\Sigma})^{-\frac{n+p+1}{2}} e^{-½tr\,(\boldsymbol{S}_\theta \boldsymbol{\Sigma}^{-1})}$$

$$\boldsymbol{S}_\theta = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\theta})(\boldsymbol{x}_i - \boldsymbol{\theta})' = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})' + n(\boldsymbol{\theta} - \overline{\boldsymbol{x}})(\boldsymbol{\theta} - \overline{\boldsymbol{x}})'$$

$$\Rightarrow f(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\theta}) \propto (\det \boldsymbol{\Sigma})^{-\frac{n+p+1}{2}} e^{-½tr\,(\boldsymbol{S}_{\overline{x}} \boldsymbol{\Sigma}^{-1})} e^{-½\, tr\left(n(\boldsymbol{\theta} - \overline{\boldsymbol{x}})(\boldsymbol{\theta} - \overline{\boldsymbol{x}})'\, \boldsymbol{\Sigma}^{-1}\right)}$$

Exercise 2

Consider the following model

$$\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \mid \boldsymbol{\theta}, \boldsymbol{\Sigma} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\theta} | \boldsymbol{\Sigma} \sim \mathrm{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0)$$
$$\boldsymbol{\Sigma} \sim W_p^{-1}(\Lambda_0, \nu_0)$$

a) Derive the conditional posterior of $\boldsymbol{\theta} | \boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{\Sigma}$
b) Derive the posterior of $\boldsymbol{\Sigma} | \boldsymbol{x}_1, \dots, \boldsymbol{x}_n$

Answers:
a) Multivariate normal
b) Inverse Wishart

Gibbs Sampler

We showed that:

$$\boldsymbol{\theta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Sigma} \sim \mathrm{N}_p(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n)$$
$$\boldsymbol{\Sigma}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\theta} \sim W_p^{-1}(\boldsymbol{S}_n, \nu_n)$$

where

$$\boldsymbol{\mu}_n = (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{x}})$$
$$\boldsymbol{\Lambda}_n = (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1})^{-1}$$
$$\boldsymbol{S}_{\boldsymbol{n}} = \boldsymbol{S}_0 + \boldsymbol{S}_\theta$$
$$\boldsymbol{S}_\theta = \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\theta})'(\boldsymbol{x}_i - \boldsymbol{\theta})$$
$$\nu_n = \nu_0 + n$$

Start with some starting value $\boldsymbol{\Sigma}^{(0)}$ (say, the sample covariance matrix) and iterate between the above two conditional posteriors to obtain a sequence $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\Sigma}^{(1)}), \ldots, (\boldsymbol{\theta}^{(S)}, \boldsymbol{\Sigma}^{(S)})$