

STA 602 - Intro to Bayesian Statistics

Lecture 12

Li Ma

Duke University

Multivariate sampling models

- ▶ We have seen one example of a distribution supported on a two-dimensional space: the bivariate normal.
- ▶ Of course this model can be used as a sampling model for observations as well as a prior model for a two-dimensional parameters.
- ▶ For example, in educational assessments, one may measure the performance on p different subjects (English, math, ...) on a collection of students.
- ▶ The scores of each student is a vector (Y_1, Y_2, \dots, Y_p) that can be modeled as drawn from some probability distribution supported on the p -dimensional space.

Example: Measuring air pollutants

- ▶ Suppose instead of measuring a single pollutant, we are measuring two (or even more) different pollutants each day.
- ▶ So each data point is of the form

$$\mathbf{Y} = (Y_1, Y_2)'.$$

- ▶ Depending on what those pollutants are, they may or may not rise or all in a correlated manner.
- ▶ For example, we may want to model each observation as a from a bivariate normal

$$\mathbf{Y} \sim N(\boldsymbol{\theta}, \Sigma)$$

where

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

where $\theta_1 = EY_1$, $\theta_2 = EY_2$, $\sigma_{11} = \text{Var} Y_1$, $\sigma_{22} = \text{Var} Y_2$,
 $\sigma_{12} = \sigma_{21} = \text{Cov}(Y_1, Y_2) = \rho_{12} \sqrt{\sigma_{11} \sigma_{22}}$ with $\rho_{12} = \text{Corr}(Y_1, Y_2)$.

- More generally, if there are p different measurements. Then

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

where Σ is *symmetric* and *positive-definite*.

- Symmetric means that $\sigma_{jk} = \sigma_{kj}$ for any $j, k = 1, 2, \dots, p$.
- This follows immediately from the fact that

$$\text{Cov}(Y_j, Y_k) = \text{Cov}(Y_k, Y_j).$$

- Positive-definiteness means that for any $\mathbf{x} \in \mathbb{R}^p$, we must have

$$\mathbf{a}'\Sigma\mathbf{a} \geq 0$$

and equality holds only if $\mathbf{a} = \mathbf{0} = (0, 0, \dots, 0)'$.

- This follows from the fact that $\text{Var}(\mathbf{a}'\mathbf{Y}) = \text{Var}(\mathbf{Y}'\mathbf{a}) = \mathbf{a}'\Sigma\mathbf{a}$.

The multivariate normal pdf

- ▶ If a p -dimensional random vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_p)'$ has a multivariate normal with mean $\boldsymbol{\theta}$ and covariance Σ , $N(\boldsymbol{\theta}, \Sigma)$, then its pdf is given by

$$p(\mathbf{y}|\boldsymbol{\theta}, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\theta})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\theta})}.$$

- ▶ Note how this generalizes the pdf of a univariate normal.

A little geometry

- Recall the *eigen-decomposition* of a symmetric matrix A

$$A = R' \Lambda R.$$

where A is a $p \times p$ symmetric matrix, R is an $p \times p$ orthonormal matrix (i.e., its columns are unit vectors that are pairwise orthogonal), and Λ is a $p \times p$ diagonal matrix.

- In particular, if A is positive-definite if and only if all diagonal elements of Λ is (strictly) positive. In particular we can write $\Lambda = D^2$ where D is $p \times p$ diagonal with $D_{ii} = \sqrt{\Lambda_{ii}}$ for $i = 1, 2, \dots, p$.
- The diagonal values of Λ are the so-called *eigen-values* of A and the columns of R is the so-called *eigen-vectors* of A .

How to generate a multivariate normal

- Now apply the above to a covariance matrix $A = \Sigma$.

$$\Sigma = R'D^2R$$

where $R = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p)$ and $D = \text{diag}(d_1, d_2, \dots, d_p)$ where all $d_i > 0$.

- Consider a random vector in \mathbb{R}^p with independent standard normal margins. That is

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix}$$

where $Z_i \stackrel{\text{iid}}{\sim} \text{N}(0, 1)$.

Rescale then rotate!

- ▶ Now consider a “stretching” on \mathbf{Z} followed by a rotation

$$\mathbf{Y} = \Sigma^{1/2} \mathbf{Z} = R' D \mathbf{Z} = \begin{pmatrix} \mathbf{a}'_1 \\ \vdots \\ \mathbf{a}'_p \end{pmatrix} \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{pmatrix} \mathbf{Z} = \begin{pmatrix} \mathbf{a}'_1 \tilde{\mathbf{Z}} \\ \vdots \\ \mathbf{a}'_p \tilde{\mathbf{Z}} \end{pmatrix}.$$

where $\tilde{\mathbf{Z}} = D\mathbf{Z} = (d_1 Z_1, d_2 Z_2, \dots, d_p Z_p)'$.

- ▶ $D\mathbf{Z}$ rescales each element of \mathbf{Z} by the corresponding factor d_1, \dots, d_p .

Rescale then rotate!

- ▶ For any unit vector \mathbf{a} , $\mathbf{a}'\tilde{\mathbf{Z}}$ is the projection of $\tilde{\mathbf{Z}}$ along the direction of \mathbf{a} , i.e., a new coordinate along a new axis in the direction of \mathbf{a} . Thus $R'D\mathbf{Z}$ projects the rescaled vector $\tilde{\mathbf{Z}} = D\mathbf{Z}$ onto the columns of R . That is, producing the new coordinates under the new axes (the columns of R .)
- ▶ Note that the covariance matrix of \mathbf{Y} is

$$R'D \cdot I \cdot DR = R'D^2R = \Sigma.$$

- ▶ Therefore, one can generate a multivariate normal with covariance Σ by generating \mathbf{Z} and apply the rescaling and rotation.
- ▶ More generally, we can generate a $N(\boldsymbol{\mu}, \Sigma)$ random vector by adding a location shift

$$\mathbf{Y} = \boldsymbol{\mu} + R'D\mathbf{Z}.$$

Change-of-variable to get pdf of multivariate normal

- ▶ The joint pdf of \mathbf{Y} then comes from the change-of-variable formula applied to the joint pdf of \mathbf{Z} .
- ▶ The joint pdf of \mathbf{Z} is

$$p(\mathbf{z}) = (2\pi)^{-p/2} e^{-\frac{\mathbf{z}'\mathbf{z}}{2}}.$$

- ▶ Now plug in $\mathbf{z} = R\mathbf{D}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and the Jacobian is $|D^{-1}R| = |D|^{-1} = |\Lambda|^{-1/2} = |\Sigma|^{-1/2}$, we get

$$\begin{aligned} p(\mathbf{y}) &= (2\pi)^{-p/2} e^{-\frac{(\mathbf{y}-\boldsymbol{\mu})'R'D^{-1}\cdot D^{-1}R(\mathbf{y}-\boldsymbol{\mu})}{2}} \cdot |\Sigma|^{-1/2} \\ &= (2\pi)^{-p/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})}. \end{aligned}$$

Under repeated sampling

- ▶ Now if we have n i.i.d. samples

$$\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\theta}, \Sigma)$$

where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{ip})'$.

- ▶ The joint pdf of n i.i.d. random vectors $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ (where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{ip})$) from $\mathcal{N}(\boldsymbol{\theta}, \Sigma)$ is given by

$$\begin{aligned} p(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n | \boldsymbol{\theta}, \Sigma) &= \prod_{i=1}^n p(\mathbf{y}_i | \boldsymbol{\theta}, \Sigma) \\ &= \prod_{i=1}^n (2\pi)^{-p/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\theta})' \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\theta})} \\ &= (2\pi)^{-np/2} |\Sigma|^{-n/2} e^{-\frac{1}{2} \sum_i (\mathbf{y}_i - \boldsymbol{\theta})' \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\theta})}. \end{aligned}$$

Bayesian inference on multivariate normal sampling models

- ▶ The general recipe for Bayesian inference remains:
 - ▶ Specify a prior on the unknown parameters ($\boldsymbol{\theta}, \Sigma$).
 - ▶ Apply Bayes theorem to identify the posterior.
 - ▶ When an analytic form of the joint posterior is unavailable, design a strategy to sample from the posterior (e.g., MCMC).
- ▶ In particular, if we can identify all of the full conditionals if possible, and apply Gibbs sampling.

Choices for priors for $(\boldsymbol{\theta}, \Sigma)$

- ▶ Motivated by the univariate normal example, let us consider placing independent priors on $\boldsymbol{\theta}$ and Σ , and aim for achieving semi-conjugacy (simple conjugate full conditionals).

$$p(\boldsymbol{\theta}, \Sigma) = p(\boldsymbol{\theta})p(\Sigma).$$

- ▶ Generalizing the univariate Gaussian model, we adopt
 - ▶ A multivariate prior on $\boldsymbol{\theta}$

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \Lambda_0).$$

- ▶ An *inverse-Wishart* (multivariate version of inverse-Gamma) for Σ

$$\Sigma \sim \text{inverse-Wishart}(\nu_0, \mathbf{S}_0),$$

or equivalently,

$$\Sigma^{-1} \sim \text{Wishart}(\nu_0, \mathbf{S}_0^{-1}).$$

The full conditional of $\boldsymbol{\theta}$

- ▶ The prior density on $\boldsymbol{\theta}$

$$\begin{aligned} p(\boldsymbol{\theta}) &= (2\pi)^{-p/2} |\Lambda_0|^{-1/2} e^{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)' \Lambda_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)} \\ &\propto e^{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)' \Lambda_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)} \\ &\propto e^{-\frac{1}{2}(\boldsymbol{\theta}' \Lambda_0^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}' \Lambda_0^{-1} \boldsymbol{\mu}_0)}. \end{aligned}$$

- ▶ Thus the full conditional of $\boldsymbol{\theta}$ is

$$\begin{aligned} p(\boldsymbol{\theta} | \mathbf{y}, \Sigma) &\propto p(\boldsymbol{\theta}, \Sigma, \mathbf{y}) \\ &= p(\mathbf{y} | \boldsymbol{\theta}, \Sigma) p(\boldsymbol{\theta}) p(\Sigma) \\ &\propto p(\mathbf{y} | \boldsymbol{\theta}, \Sigma) p(\boldsymbol{\theta}). \end{aligned}$$

The full conditional of $\boldsymbol{\theta}$

- Now notice that the likelihood part

$$\begin{aligned} p(\mathbf{y} | \boldsymbol{\theta}, \Sigma) &\propto e^{-\frac{1}{2} \sum_i (\mathbf{y}_i - \boldsymbol{\theta})' \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\theta})} \\ &\propto e^{-\frac{1}{2} [\sum_{i=1}^n (\boldsymbol{\theta}' \Sigma^{-1} \boldsymbol{\theta} - 2 \boldsymbol{\theta}' \Sigma^{-1} \mathbf{y}_i)]} \\ &\propto e^{-\frac{1}{2} [\boldsymbol{\theta}' (n \Sigma^{-1}) \boldsymbol{\theta} - 2 \boldsymbol{\theta}' (n \Sigma^{-1}) \bar{\mathbf{y}}]} \end{aligned}$$

- Thus

$$\begin{aligned} p(\boldsymbol{\theta} | \mathbf{y}, \Sigma) &\propto e^{-\frac{1}{2} [\boldsymbol{\theta}' (n \Sigma^{-1}) \boldsymbol{\theta} - 2 \boldsymbol{\theta}' (n \Sigma^{-1}) \bar{\mathbf{y}}]} \cdot e^{-\frac{1}{2} (\boldsymbol{\theta}' \Lambda_0^{-1} \boldsymbol{\theta} - 2 \boldsymbol{\theta}' \Lambda_0^{-1} \boldsymbol{\mu}_0)} \\ &= e^{-\frac{1}{2} [\boldsymbol{\theta}' (n \Sigma^{-1}) \boldsymbol{\theta} - 2 \boldsymbol{\theta}' (n \Sigma^{-1}) \bar{\mathbf{y}} + \boldsymbol{\theta}' \Lambda_0^{-1} \boldsymbol{\theta} - 2 \boldsymbol{\theta}' \Lambda_0^{-1} \boldsymbol{\mu}_0]} \\ &= e^{-\frac{1}{2} [\boldsymbol{\theta}' (n \Sigma^{-1} + \Lambda_0^{-1}) \boldsymbol{\theta} - 2 \boldsymbol{\theta}' (n \Sigma^{-1} \bar{\mathbf{y}} + \Lambda_0^{-1} \boldsymbol{\mu}_0)]}. \end{aligned}$$

- ▶ Let

$$\boldsymbol{\mu}_n = \Lambda_n(\Lambda_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\mathbf{y}}) \quad \text{and} \quad \Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}.$$

- ▶ We have

$$p(\boldsymbol{\theta} | \mathbf{y}, \Sigma) \propto e^{-\frac{1}{2}(\boldsymbol{\theta}'\Lambda_n^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}'\Lambda_n^{-1}\boldsymbol{\mu}_n)} \propto e^{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_n)'\Lambda_n^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_n)}.$$

Compare this with the prior $p(\boldsymbol{\theta})$.

- ▶ Just like in the univariate Gaussian model, the posterior mean is a weighted average of sample mean $\bar{\mathbf{y}}$ and the prior mean $\boldsymbol{\mu}_0$, and the posterior precision (i.e., inverse of covariance) is the sum of the prior precision and sample precision.

Prior choice for Σ

- ▶ Again, drawing the analogy from the univariate Gaussian model, we expect that a multivariate generalization of the Gamma prior on the precision matrix might lead to a conjugate full conditional for Σ .
- ▶ How to generalize the Gamma distribution?
- ▶ Note that whatever prior on Σ we adopt, it must guarantee that Σ is always a covariance matrix—that is, it is symmetric and positive-definite.
- ▶ How?
 - ▶ One could use the eigen-decomposition, randomly generate D and R .
 - ▶ Alternatively, generalize the Gamma distribution. (To get semi-conjugacy.)

Gamma and scaled χ^2 -square

- ▶ Recall that the χ^2 distribution with 1 degree of freedom is also $\text{Gamma}(1/2, 1/2)$. It is the distribution of Z^2 where $Z \sim \text{N}(0, 1)$.
- ▶ Also, χ^2 with k degrees of freedom is $\text{Gamma}(k/2, 1/2)$. It is the distribution of $\sum_{i=1}^k Z_i^2$ where $Z_i \stackrel{\text{iid}}{\sim} \text{N}(0, 1)$.
- ▶ Now consider $Z_i \stackrel{\text{iid}}{\sim} \text{N}(0, s_0^{-2})$, then

$$\sum_{i=1}^k Z_i^2 \sim \text{Gamma}(k/2, s_0^2/2).$$

- ▶ This can also be written as $\text{Gamma}(v/2, v\sigma_0^2/2)$ by letting $v = k$ and $\sigma_0^2 = s_0^2/v$.
- ▶ This gives a reparametrization of Gamma distribution with two free parameters.
- ▶ Hence we can think of the Gamma distribution as that of the sum of squares of Gaussian variables with 0 mean.

Generalization to a multivariate distribution

- ▶ Now consider the distribution of the “sum of squares” of multivariate Gaussian variables with $\mathbf{0}$ mean.
- ▶ Let $\mathbf{Z}_1, \dots, \mathbf{Z}_k$ be p -dimensional random vectors such that

$$\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_k \stackrel{\text{iid}}{\sim} \mathbf{N}(\mathbf{0}, \mathbf{S}_0^{-1}).$$

- ▶ Let $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_k)$ be a $p \times k$ matrix whose columns are the \mathbf{Z}_i 's.
- ▶ Then let's consider the “sum of squares” matrix

$$\Sigma^{-1} = \mathbf{Z}\mathbf{Z}' = \sum_{i=1}^k \mathbf{Z}_i \mathbf{Z}_i'.$$

- Note that

$$\mathbf{Z}_i \mathbf{Z}_i' = \begin{pmatrix} Z_{i1} \\ Z_{i2} \\ \vdots \\ Z_{ip} \end{pmatrix} (Z_{i1} \quad Z_{i2} \quad \cdots Z_{ip}) = \begin{pmatrix} Z_{i1}^2 & Z_{i1}Z_{i2} & \cdots & Z_{i1}Z_{ip} \\ Z_{i2}Z_{i1} & Z_{i2}^2 & \cdots & Z_{i2}Z_{ip} \\ \vdots & \vdots & \cdots & \vdots \\ Z_{ip}Z_{i1} & Z_{ip}Z_{i2} & \cdots & Z_{ip}^2 \end{pmatrix}.$$

That is, $\mathbf{Z}_i \mathbf{Z}_i'$ is a $p \times p$ matrix whose (j, k) th element is $Z_{ij}Z_{ik}$.

- Therefore, $\Sigma^{-1} = \mathbf{Z}\mathbf{Z}' = \sum_{i=1}^k \mathbf{Z}_i \mathbf{Z}_i'$ is a $p \times p$ matrix whose (j, k) th element is

$$(\mathbf{Z}\mathbf{Z}')_{jk} = \sum_{i=1}^n Z_{ij}Z_{ik}.$$

Is Σ a covariance matrix

- ▶ This randomly generated matrix $\mathbf{Z}\mathbf{Z}'$ is
 - ▶ Symmetric.
 - ▶ Positive-definite (with probability 1) when $k > p - 1$, because for any $\mathbf{a} \in \mathbb{R}^p$,

$$\mathbf{a}'\Sigma\mathbf{a} = \mathbf{a}'\mathbf{Z}\mathbf{Z}'\mathbf{a} = \mathbf{a}'\left(\sum_{i=1}^k \mathbf{z}_i\mathbf{z}_i'\right)\mathbf{a} = \sum_{i=1}^k (\mathbf{a}'\mathbf{z}_i)^2 \geq 0$$

and it is 0 only when $\mathbf{a}'\mathbf{z}_i = 0$ for all i . That is if

$$\mathbf{a}'\mathbf{Z} = (\mathbf{a}'\mathbf{z}_1, \mathbf{a}'\mathbf{z}_2, \dots, \mathbf{a}'\mathbf{z}_k) = \mathbf{0}'_k = (0, 0, \dots, 0).$$

- ▶ In words, the projection of \mathbf{Z}_i onto the direction of \mathbf{a} is 0, that is the vector \mathbf{a} is orthogonal to each of the vector \mathbf{Z}_i simultaneously for all \mathbf{Z}_i , which has 0 probability to occur when $k > p - 1$.
- ▶ To see this, if such an \mathbf{a} exists, the row vectors of \mathbf{Z} (which is a $p \times k$ matrix) are linearly dependent. So the p row vectors (each in \mathbb{R}^k) will perfectly lie in a $(p - 1)$ -dimensional subspace of \mathbb{R}^k , which has probability 0.
 - ▶ Say, if $k = 2$ and $p = 2$, two random vectors in \mathbb{R}^2 (i.e., \mathbb{R}^k) has probability 0 to fall on a line (i.e., $(p - 1)$ -dimensional space).

Wishart distribution

- ▶ We define the distribution of the random matrix $\Sigma^{-1} = \mathbf{Z}\mathbf{Z}'$ as the **Wishart(k, \mathbf{S}_0), distribution.**
- ▶ By construction, we see that

$$\mathrm{E}(\Sigma^{-1} | k, \mathbf{S}_0) = k \mathrm{E} \mathbf{Z}_i \mathbf{Z}_i' = k \mathbf{S}_0^{-1}$$

and

$$\mathrm{E}(\Sigma | k, \mathbf{S}_0) = \frac{1}{k - p - 1} \mathbf{S}_0.$$

- ▶ Note that we need $k > p + 1$ to have a finite mean, while the distribution is defined for $k > p - 1$.

An equivalent reparametrization

- For additional interpretability, reparametrize by setting

$$\Sigma_0 = \frac{1}{k-p-1} \mathbf{S}_0 \quad \text{or} \quad \Sigma_0^{-1} = (k-p-1) \mathbf{S}_0^{-1}.$$

Then

$$\mathbb{E}(\Sigma | k, \mathbf{S}_0) = \Sigma_0.$$

The distribution of Σ^{-1} is more tight around Σ_0^{-1} as k increases.

- To see this, recall that

$$\mathbf{z}_i \stackrel{\text{iid}}{\sim} \mathcal{N}\left(\mathbf{0}, \frac{1}{k-p-1} \Sigma_0^{-1}\right)$$

whereas

$$\Sigma^{-1} = \sum_i \mathbf{z}_i \mathbf{z}_i' = \frac{1}{k-p-1} \left(\sum_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i' \right) = \frac{k}{k-p-1} \left(\frac{\sum_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i'}{k} \right)$$

where

$$\tilde{\mathbf{z}}_i = \sqrt{k-p-1} \cdot \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_0^{-1}).$$

- We can therefore use larger k for stronger prior belief around Σ_0 .

Inverse-Wishart distribution

- ▶ Correspondingly its inverse $\Sigma = (\mathbf{Z}\mathbf{Z}')^{-1}$ is said to have an **Inverse-Wishart**(k, \mathbf{S}_0).
- ▶ We can also show that its pdf is

$$p(\Sigma) \propto |\Sigma|^{-\frac{k+p+1}{2}} \cdot e^{-\frac{1}{2}\text{tr}(\mathbf{S}_0\Sigma^{-1})}$$

where $\text{tr}(A) = \sum_j A_{jj}$ for a square matrix A is its *trace*.

- ▶ The normalizing constant is complicated but available in close-form. (Refer to textbook for detailed formula.) In practice we usually don't need to memorize it.
- ▶ More generally, we can extend this definition to “fractional sums”. That is we replace the integer k with any $\nu_0 > 0$ (called *degrees of freedom*), which gives inverse-Wishart(ν_0, \mathbf{S}_0) where ν_0 is no longer required to be an integer.