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 $(Y_1, Y_2, \ldots, Y_p)$ 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

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

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

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

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

  where Σ is *symmetric* and *positive-definite*.
- Symmetric means that $\sigma_{jk} = \sigma_{kj}$ for any $j, k = 1, 2, \ldots, p$.
- This follows immediately from the fact that

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

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

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

$Var(\underset{1\times p}{a^T}\,\underset{p\times 1}{Y}) = Var(\underset{1\times 1}{u})$

  and equality holds only if $\mathbf{a} = \mathbf{0} = (0, 0, \ldots, 0)'$.
- This follows from the fact that $\mathrm{Var}(\mathbf{a}'\mathbf{Y}) = \mathrm{Var}(\mathbf{Y}'\mathbf{a}) = \mathbf{a}'\Sigma\mathbf{a}$.

The multivariate normal pdf

$\underset{\sim}{y} = \underset{\sim}{\theta} + \Sigma^{1/2}\underset{\sim}{z}$, where $z = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix}$

$\Sigma^{1/2}$: rescale
rotate

$z_i \overset{iid}{\sim} N(0,1)$

- If a p-dimensional random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)'$ has a multivariate normal with mean $\boldsymbol{\theta}$ and covariance Σ, $\mathrm{N}(\boldsymbol{\theta}, \Sigma)$, 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

Write $\Sigma = R^T A R = R^T D \cdot D^T R = (R^T D) \cdot (R^T D)^T$

Let $\Sigma^{1/2} = R^T D$

Now $y = \theta + \Sigma^{1/2} z = \theta + R^T D z$

- 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 $p \times p$ diagnomal 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 $i = 1, 2, \ldots, p$.

- 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 $A = \Sigma$.

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

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

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

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

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

Rescale then rotate!

- Now consider a "stretching" on $\boldsymbol{Z}$ followed by a rotation

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

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

- $D\boldsymbol{Z}$ rescales each element of $\boldsymbol{Z}$ by the corresponding factor $d_1, \ldots, d_p$.

Rescale then rotate!

- For any unit vector $\boldsymbol{a}$, $\boldsymbol{a}'\tilde{\boldsymbol{Z}}$ is the projection of $\tilde{\boldsymbol{Z}}$ along the direction of $\boldsymbol{a}$, i.e., a new coordinate along a new axis in the direction of $\boldsymbol{a}$. Thus $R'D\boldsymbol{Z}$ projects the rescaled vector $\tilde{\boldsymbol{Z}} = D\boldsymbol{Z}$ 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 $\boldsymbol{Z}$ and apply the rescaling and rotation.
- More generally, we can generate a $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ random vector by adding a location shift

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

Change-of-variable to get pdf of multivariate normal

$y = \theta + R' D z$

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

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

- Now plug in $\boldsymbol{z} = RD^{-1}(\mathbf{y} - \boldsymbol{\theta})$ and the Jacobian is $|D^{-1}R| = |D|^{-1} = |\Lambda|^{-1/2} = |\Sigma|^{-1/2}$, we get $p(y) = p(z(y)) \cdot \left|\frac{\partial z}{\partial y}\right|$

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

Under repeated sampling

- Now if we have n i.i.d. samples

$$\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \overset{\text{iid}}{\sim} \mathrm{N}(\boldsymbol{\theta}, \Sigma)$$

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

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

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 $\boldsymbol{\theta}$

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

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

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

  or equivalently,

$$\Sigma^{-1} \sim \text{Wishart}(\nu_0, \boldsymbol{S}_0^{\text{[redacted]}}).$$

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

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

- Thus

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

- 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

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

  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 multivarate 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?

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

  - 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

$\text{Gamma}(\alpha_1, \beta) + \text{Gamma}(\alpha_2, \beta)$
$= \text{Gamma}(\alpha_1 + \alpha_2, \beta)$

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

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

- This can also be written as Gamma$\left(\nu/2, \nu\sigma_0^2/2\right)$ by letting $\nu = k$ and $\sigma_0^2 = s_0^2/\nu$.
- 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 $\boldsymbol{Z}_1, \ldots \boldsymbol{Z}_k$ be p-dimensional random vectors such that

$$\boldsymbol{Z}_1 = \begin{pmatrix} z_{11} \\ z_{12} \\ \vdots \\ z_{1p} \end{pmatrix} \quad \boldsymbol{Z}_k = \begin{pmatrix} z_{k1} \\ z_{k2} \\ \vdots \\ z_{kp} \end{pmatrix}$$

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

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

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

Sum of squares of i.i.d. Gaussian variables w/ mean 0.

* parameter k : degrees of freedom

* parameter S_0 : shape

Define $\Sigma^{-1} \sim \text{Wishart}(k, S_0)$

$\Sigma \sim \text{Inverse-Wishart}(k, S_0)$

- Note that

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

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

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

Is Σ a covariance matrix

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

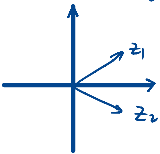
$$\boldsymbol{a}'\Sigma^{-1}\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{ZZ}'\boldsymbol{a} = \boldsymbol{a}'\left(\sum_{i=1}^{k}\boldsymbol{Z}_i\boldsymbol{Z}_i'\right)\boldsymbol{a} = \sum_{i=1}^{k}(\boldsymbol{a}'\boldsymbol{Z}_i)^2 \geq 0$$

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

$$\boldsymbol{a}'\boldsymbol{Z} = (\boldsymbol{a}'\boldsymbol{Z}_1, \boldsymbol{a}'\boldsymbol{Z}_2, \ldots, \boldsymbol{a}'\boldsymbol{Z}_k) = \boldsymbol{0}_k' = (0, 0, \ldots, 0).$$

a is orthogonal to $(z_1, \ldots, z_k)$ with probability 0.

if $p=2$, then



impossible to find an $a \in \mathbb{R}^2$ s.t $a^T(z_1\ z_2) = 0$.

- In words, the projection of $\boldsymbol{Z}_i$ onto the direction of $\boldsymbol{a}$ is 0, that is the vector $\boldsymbol{a}$ is orthogonal to each of the vector $\boldsymbol{Z}_i$ simultaneously for all $\boldsymbol{Z}_i$, which has 0 probability to occur when $k > p - 1$.
- To see this, if such an $\boldsymbol{a}$ exists, the row vectors of $\boldsymbol{Z}$ (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 $\Sigma^{-1} = \boldsymbol{Z}\boldsymbol{Z}'$ as the Wishart$(k, \boldsymbol{S}_0)$, distribution.
- By construction, we see that

Wishart

$$\mathrm{E}\left(\Sigma^{-1} \mid k, \boldsymbol{S}_0\right) = k\,\mathrm{E}\,\boldsymbol{Z}_i\boldsymbol{Z}_i' = k\boldsymbol{S}_0^{-1}$$

for $k > p-1$

and

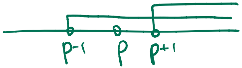
Inverse-Wishart

$$\mathrm{E}\left(\Sigma \mid k, \boldsymbol{S}_0\right) = \frac{1}{k-p-1}\boldsymbol{S}_0.$$

for $k > p+1$

S_0 kind of plays the role of prior mean

- 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 interpretatbility, reparametrize by setting

$$\Sigma_0 = \frac{1}{k-p-1} S_0 \qquad \Sigma_0^{-1} = (k-p-1)\boldsymbol{S}_0^{-1}.$$

  Then

$$\mathrm{E}\left(\Sigma \,|\, k, \boldsymbol{S}_0\right) = \Sigma_0.$$

  The distribution of Σ^{-1} is more tight around Σ_0^{-1} as k increases.
- To see this, recall that

$$\boldsymbol{Z}_i \overset{\text{iid}}{\sim} \mathrm{N}\left(\boldsymbol{0}, \frac{1}{k-p-1}\Sigma_0^{-1}\right)$$

  whereas

$$\Sigma^{-1} = \sum_i \boldsymbol{Z}_i \boldsymbol{Z}_i' = \frac{1}{k-p-1}\left(\sum_i \tilde{\boldsymbol{Z}}_i \tilde{\boldsymbol{Z}}_i'\right) = \underbrace{\frac{k}{k-p-1}}_{\downarrow\, k\to\infty \atop 1}\underbrace{\left(\frac{\sum_i \tilde{\boldsymbol{Z}}_i \tilde{\boldsymbol{Z}}_i'}{k}\right)}_{\downarrow \atop \text{center around } \Sigma_0^{-1}}$$

  where

$$\tilde{\boldsymbol{Z}}_i = \sqrt{k-p-1} \cdot \boldsymbol{Z}_i \sim \mathrm{N}(\boldsymbol{0}, \Sigma_0^{-1}).$$

- We can therefore use larger k for stronger prior belief around Σ_0^{-1}.

Inverse-Wishart distribution

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

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

  where $\mathrm{tr}(A) = \sum_j A_{jj}$ 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$(\nu_0, \boldsymbol{S}_0)$ where ν_0 is no longer required to be an integer.