Lecture 2: The Multivariate Normal Distribution
STATS305C: Applied Statistics III

Scott Linderman

April 6, 2023

Last Time...

▶ Course Overview

▶ Bayes' Rule

▶ Normal with unknown mean (normal prior)

▶ Normal with unknown precision (χ^2 prior)

▶ Normal with unknown variance (χ^{-2} prior)

▶ Normal with unknown mean and variance (NIX prior)

▶ Posterior marginals (Student's t distribution)

Today...

► Multivariate normal (MVN) distribution

  ► Generative story

  ► Marginal distributions

  ► Conditional distributions

  ► Linear Gaussian models

► The Wishart and inverse Wishart distributions

► Bayesian estimation with a normal-inverse-Wishart (NIW) prior

► Posterior marginals (multivariate Student's t distribution)

Reading: Bishop, Ch 2.3. See also: Murphy, Ch 2.3 and 3.2.4.

Generative Story

Start with a vector of standard normal random variates, $\mathbf{z} = [z_1, \ldots, z_D]^\top$ where $z_d \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

This is a D-dimensional random variable, but not a very interesting one. All the coordinates are independent! The joint density is,

$$p(\mathbf{z}) = \prod_{d=1}^{D} \mathcal{N}(z_d \mid 0, 1) \tag{1}$$

$$= \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_d^2} \tag{2}$$

$$= (2\pi)^{-\frac{D}{2}} \exp\left\{ -\frac{1}{2} \sum_{d=1}^{D} z_d^2 \right\} \tag{3}$$

$$= (2\pi)^{-\frac{D}{2}} \exp\left\{ -\frac{1}{2} \mathbf{z}^\top \mathbf{z} \right\} \tag{4}$$

$$\overset{\triangle}{=} \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}). \tag{5}$$

Question: What do the contours of this joint density look like in $D = 2$ dimensions?

Spherical Gaussian Density

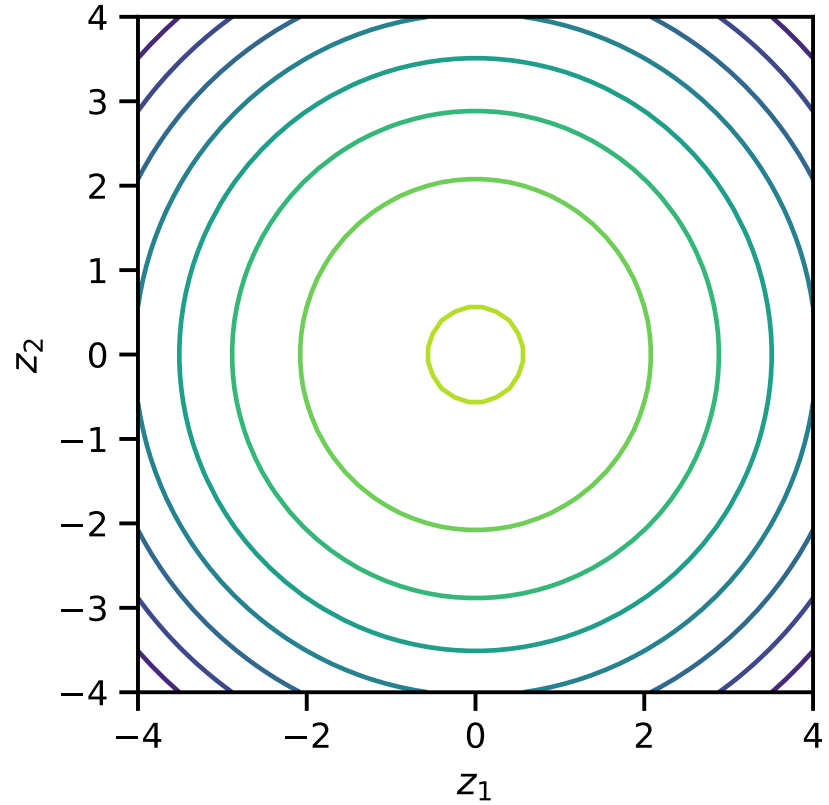
Figure: Contours of the pdf of a **spherical Gaussian** distribution, $\mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$.

Generative Story II

We can obtain more interesting joint distributions by transforming this random vector.

For example, let U be an orthogonal $D \times D$ matrix and $\Lambda = \text{diag}([\lambda_1, \ldots, \lambda_D])$ with $\lambda_d > 0$. Define the linearly transformed random variable $x = U\Lambda^{\frac{1}{2}}z$.

Exercise: Compute the mean $\mathbb{E}[x]$ and covariance $\text{Cov}[x]$.

$$\mathbb{E}[x] = 0$$

$$\text{Cov}[x] = \text{Cov}\left(U\Lambda^{1/2}z\right) = U\Lambda^{1/2}\Lambda^{1/2}U^\top = U\Lambda U^\top$$

$$\triangleq \Sigma$$

Generative Story III

Exercise: Explain each step in this derivation of the density:

$$p(\boldsymbol{x}) = p(\boldsymbol{z}) \left| \frac{d\boldsymbol{z}}{d\boldsymbol{x}} \right| \qquad \text{Change of measure} \qquad (6)$$

$$= p(\Lambda^{-\frac{1}{2}} \boldsymbol{U}^{\top} \boldsymbol{x}) |\Lambda^{-\frac{1}{2}} \boldsymbol{U}^{\top}| \qquad (7)$$

$$= (2\pi)^{-\frac{D}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{x}^{\top} \boldsymbol{U} \Lambda^{-1} \boldsymbol{U}^{\top} \boldsymbol{x} \right\} |\Lambda^{-\frac{1}{2}}| |\boldsymbol{U}^{\top}| \qquad (8)$$

$$= (2\pi)^{-\frac{D}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{x}^{\top} \boldsymbol{U} \Lambda^{-1} \boldsymbol{U}^{\top} \boldsymbol{x} \right\} |\Lambda|^{-\frac{1}{2}} \qquad |\boldsymbol{U}^{\top}| = 1 \qquad (9)$$

$$= (2\pi)^{-\frac{D}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{x}^{\top} \Sigma^{-1} \boldsymbol{x} \right\} |\Sigma|^{-\frac{1}{2}} \qquad (10)$$

where $\Sigma = \boldsymbol{U} \Lambda \boldsymbol{U}^{\top}$.

$$x = U \Lambda^{1/2} z$$

$$z = \Lambda^{-1/2} U^{-1} x$$

$$= \Lambda^{-1/2} U^{\top} x$$

$$= \mathcal{N}(x \mid 0, \Sigma)$$

Generative Story IV

Last but not least, add a translation so that $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{z} + \boldsymbol{\mu}$ for $\boldsymbol{\mu} \in \mathbb{R}^D$.

Question: How does this change the mean and covariance of $\boldsymbol{x}$?

Following the same argument as above,

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \triangleq \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{11}$$

This is the pdf of the **multivariate normal** distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$. Again, $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\top}$.

It depends on $\boldsymbol{x}$ through the **squared Mahalanobis distance** between $\boldsymbol{x}$ and $\boldsymbol{\mu}$,

$$\Delta^2 = (\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{12}$$

Question: What do the contours of this density look like?
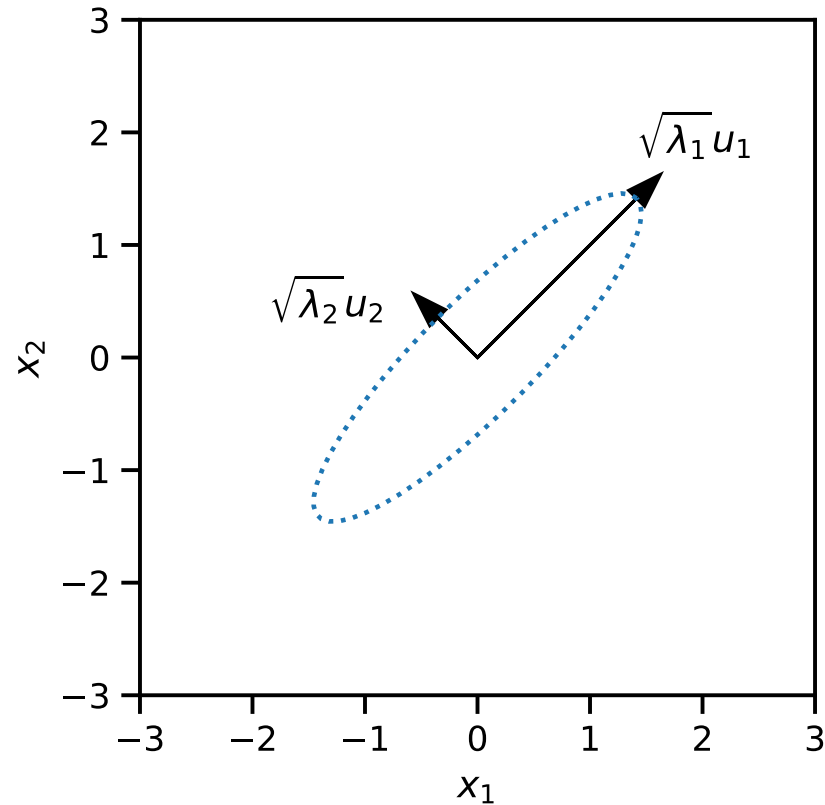
~~Spherical~~ Gaussian Density

Figure: Contours of the pdf of a **multivariate normal** distribution, $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
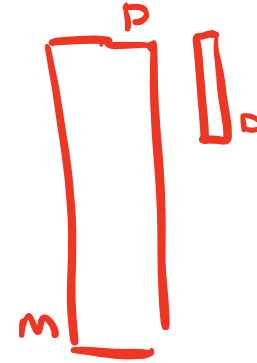
Generative Story V

We introduced the MVN by scaling z, applying a change of basis, and adding a translation.

However, we could have applied *any* linear transformation,

$$x = Az + \mu. \tag{13}$$

for $A \in \mathbb{R}^{M \times D}$ and $\mu \in \mathbb{R}^M$. Then,

$$x \sim \mathcal{N}(\mu, \Sigma). \tag{14}$$

with $\Sigma = AA^\top$.

Question: What if $M > D$?

Generative Story VI

Finally, we can go the other direction as well: to sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, first compute **any square root** of the covariance matrix $\boldsymbol{\Sigma}^{\frac{1}{2}} \in \mathbb{R}^{D \times D}$ such that $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}^{\frac{1}{2}})(\boldsymbol{\Sigma}^{\frac{1}{2}})^{\top}$, then set,

$$\boldsymbol{x} = \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{z} + \boldsymbol{\mu}, \tag{15}$$

where $\boldsymbol{z} \in \mathbb{R}^{D}$ is a vector of standard normal random variates.

As before, we could use the eigendecomposition $\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}$ where $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\top}$. Since the covariance is positive semidefinite, we could also use the **Cholesky decomposition**, $\boldsymbol{\Sigma}^{\frac{1}{2}} = \mathrm{chol}(\boldsymbol{\Sigma})$, in which case the square root is lower triangular.

Linear Transformations of Gaussians

The multivariate normal distribution is closed under linear transformations,

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \Rightarrow \quad Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top). \tag{16}$$

Marginal Distributions

Let $\boldsymbol{x} \in \mathbb{R}^D$ be a multivariate normal random vector. Partition the vector and the MVN parameters into two subsets,

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}. \tag{17}$$

The symmetry of $\boldsymbol{\Sigma}$ implies that Σ_{aa} and Σ_{bb} are symmetric, whereas $\Sigma_{ab} = \Sigma_{ba}^{\top}$.

Now consider the linear transformation $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{bmatrix}$ so that $\boldsymbol{Ax} = \boldsymbol{x}_a$.

The linearity property implies that the **marginal distributions** of the multivariate normal are also multivariate normal,

$$p(\boldsymbol{x}_a) = \mathcal{N}\left(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\top}\right) = \mathcal{N}\left(\boldsymbol{\mu}_a, \Sigma_{aa}\right). \tag{18}$$

Simply extract the corresponding blocks of the mean and covariance matrix to get the marginal.

Conditional Distributions

Let $\boldsymbol{x} \in \mathbb{R}^D$ be a multivariate normal random vector. Partition the vector and the MVN parameters into two subsets,

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}. \tag{19}$$

The symmetry of $\boldsymbol{\Sigma}$ implies that Σ_{aa} and Σ_{bb} are symmetric, whereas $\Sigma_{ab} = \Sigma_{ba}^{\top}$.

We can write the inverse of $\boldsymbol{\Sigma}$ in block form as well,

$$\boldsymbol{\Sigma}^{-1} \triangleq \boldsymbol{\Lambda} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}. \tag{20}$$

It too is symmetric, and its blocks involve the **Schur complement**,

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} \tag{21}$$

$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}. \tag{22}$$

Conditional Distributions II

What is $p(\mathbf{x}_a \mid \mathbf{x}_b)$? By Bayes' rule,

$$p(\mathbf{x}_a \mid \mathbf{x}_b) \propto p(\mathbf{x}_a, \mathbf{x}_b) = \mathcal{N}(\mathbf{x} \mid \mu, \Sigma) \tag{23}$$

$$\propto \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu)\right\} \tag{24}$$

$$\propto \exp\left\{-\frac{1}{2}(\mathbf{x}_a - \mu_a)^\top \Lambda_{aa}(\mathbf{x}_a - \mu_a) - (\mathbf{x}_a - \mu_a)^\top \Lambda_{ab}(\mathbf{x}_b - \mu_b)\right\} \tag{25}$$

$$\propto \exp\left\{-\frac{1}{2}\mathbf{x}_a^\top J_{a|b}\mathbf{x}_a + \mathbf{x}_a^\top h_{a|b}\right\} \tag{26}$$

where $J_{a|b} = \Lambda_{aa}$ and $h_{a|b} = \Lambda_{aa}\mu_a - \Lambda_{ab}(\mathbf{x}_b - \mu_b)$.

This is the **information form** of a Gaussian density on $\mathbf{x}_a$.

Conditional Distributions III

Completing the square (recall Lecture 1), we find that

$$p(\mathbf{x}_a \mid \mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a \mid \boldsymbol{\mu}_{a|b}, \Sigma_{a|b}), \tag{27}$$

where

$$\Sigma_{a|b} = J_{a|b}^{-1} \tag{28}$$

$$= \Lambda_{aa}^{-1} \tag{29}$$

$$= \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba} \tag{30}$$

and

$$\boldsymbol{\mu}_{a|b} = J_{a|b}^{-1} \mathbf{h}_{a|b} \tag{31}$$

$$= \boldsymbol{\mu}_a - \Lambda_{aa}^{-1} \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) \tag{32}$$

$$= \boldsymbol{\mu}_a + \Sigma_{ab} \Sigma_{bb}^{-1} (\mathbf{x}_b - \boldsymbol{\mu}_b). \tag{33}$$

.

Linear Gaussian Models

Now suppose $x \sim \mathcal{N}(b, Q)$ and $y \mid x \sim \mathcal{N}(Cx + d, R)$. What is the joint distribution $p(x, y)$?

Reparameterize the joint model as,

$$x = b + Q^{\frac{1}{2}} z_x \tag{34}$$

$$y = Cx + d + R^{\frac{1}{2}} z_y \tag{35}$$

$$= Cb + CQ^{\frac{1}{2}} z_x + d + R^{\frac{1}{2}} z_y \tag{36}$$

where z_x and z_y are independent standard normal vectors.

Now combine them into a single vector and rearrange terms,

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ Cb + d \end{bmatrix} + \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ CQ^{\frac{1}{2}} & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} z_x \\ z_y \end{bmatrix} \tag{37}$$

Linear Gaussian Models II

Thus, x and y are jointly Gaussian,

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}(\mu, \Sigma) \tag{38}$$

where

$$\mu = \begin{bmatrix} b \\ Cb + d \end{bmatrix} \tag{39}$$

and

$$\Sigma = \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ CQ^{\frac{1}{2}} & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} Q^{\frac{1}{2}} & Q^{\frac{1}{2}}C^{\top} \\ 0 & R^{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} Q & QC^{\top} \\ CQ & CQC^{\top} + R \end{bmatrix}. \tag{40}$$

Question: What is the marginal distribution of y?

$$y \sim \mathcal{N}(Cb+d, \; CQC^{\top} + R)$$

Maximum Likelihood Estimation

Exercise: The log likelihood is,

$$X_n \stackrel{iid}{\sim} N(\mu, \Sigma)$$

$$\mathcal{L}(\mu, \Sigma) = \sum_{n=1}^{N} \log p(x_n \mid \mu, \Sigma). \tag{41}$$

Take gradients and set them to zero to obtain the the maximum likelihood estimates,

$$\mu_{\text{ML}}, \Sigma_{\text{ML}} = \arg\max \mathcal{L}(\mu, \Sigma). \tag{42}$$

Show that,

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{43}$$

$$\Sigma_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\text{ML}})(x_n - \mu_{\text{ML}})^{\top}. \tag{44}$$

Bayesian Estimation: Multivariate Normal with Unknown Mean

Like last time, let's start with a simple Bayesian model of Gaussian data with a known covariance but an unknown mean:

$$\mu \sim \mathcal{N}(\mu_0, \Sigma_0) \tag{45}$$

$$x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Sigma). \tag{46}$$

Goal: Infer the posterior distribution $p(\mu \mid X, \eta)$ where $\eta = (\mu_0, \Sigma_0, \Sigma)$.

$$p(\mu \mid X, \eta) \propto \mathcal{N}(\mu \mid \mu_0, \Sigma_0) \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \Sigma) \tag{47}$$

$$\propto \exp\left\{-\frac{1}{2}(\mu - \mu_0)^\top \Sigma_0^{-1}(\mu - \mu_0)\right\} \prod_{n=1}^{N} \exp\left\{-\frac{1}{2}(x_n - \mu)^\top \Sigma^{-1}(x_n - \mu)\right\} \tag{48}$$

$$\propto \exp\left\{-\frac{1}{2}\mu^\top J_N \mu + \mu^\top h_N\right\} \tag{49}$$

where $J_N = \Sigma_0^{-1} + N\Sigma^{-1}$ and $h_N = \Sigma_0^{-1}\mu_0 + \sum_{n=1}^{N} \Sigma^{-1} x_n$.

Bayesian Estimation: Multivariate Normal with Unknown Mean II

Completing the square,

$$p(\boldsymbol{\mu} \mid \boldsymbol{X}, \boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \tag{50}$$

where

$$\boldsymbol{\Sigma}_N = \boldsymbol{J}_N^{-1} = (\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1})^{-1} \tag{51}$$

$$\boldsymbol{\mu}_N = \boldsymbol{J}_N^{-1} \boldsymbol{h}_N = \boldsymbol{\Sigma}_N \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_n \right) \tag{52}$$

Question: What does the posterior converge to in the uninformative limit $\boldsymbol{\Sigma}_0 \to \infty$?

$$\mu_N \to \frac{1}{N} \sum_n x_n \qquad \Sigma_N \to \Sigma/$$

Multivariate Normal with Unknown *Precision*

Now imagine the mean is known but not the covariance. Like last time, we will actually start by parameterizing the model in terms of the **precision**, i.e. inverse covariance, $\Lambda = \Sigma^{-1}$.

In the univariate case, we used a scaled χ^2 distribution for the prior, which was defined as the average of squared standard normal variates. We will take an analogous approach here.

Let $\Lambda = \sum_{i=1}^{\nu_0} z_i z_i^\top$ where $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Lambda_0)$ with covariance Λ_0.

Then, Λ follows a **Wishart distribution**[*] with ν_0 degrees of freedom and scale Λ_0, written,

$$\Lambda \sim \mathrm{W}(\nu_0, \Lambda_0). \tag{53}$$

Its expected value is $\mathbb{E}[\Lambda] = \nu_0 \Lambda_0$.

---

[*] There is an unfortunate asymmetry between this definition and the definition of the scaled χ^2 distribution from the previous lecture. Whereas we defined the scaled χ^2 as the *average* of squared Gaussian random variables, here we've defined the Wishart to be the sum of "squared" (really the outer product of) multivariate normal random vectors. This is to be consistent with the textbook definitions of the Wishart distribution, even though I find it more convenient to parameterize the distribution in terms of its mean.

The Wishart Distribution

Let $\mathscr{S}_D$ denote the set of $D \times D$ positive definite matrices, and let $\mathbf{\Lambda} \in \mathscr{S}_D$ be a random variable.

The Wishart distribution is a distribution on the set of positive definite matrices. Its pdf is,

$$\mathrm{W}(\mathbf{\Lambda} \mid \nu_0, \mathbf{\Lambda}_0) = \frac{1}{2^{\frac{\nu_0 D}{2}} |\mathbf{\Lambda}_0|^{\frac{\nu_0}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)} |\mathbf{\Lambda}|^{\frac{\nu_0 - D - 1}{2}} e^{-\frac{1}{2} \mathrm{Tr}(\mathbf{\Lambda}_0^{-1} \mathbf{\Lambda})} \tag{54}$$

where $\nu_0 > 0$ specifies the **degrees of freedom** and $\mathbf{\Lambda}_0 \in \mathscr{S}_D$ is the **scale**. Γ_D is the multivariate gamma function.

The mean of the Wishart distribution is $\nu_0 \mathbf{\Lambda}_0$ and the mode is $(\nu_0 - D - 1) \mathbf{\Lambda}_0$ for $\nu_0 \geq (D + 1)$.

In the univariate case where $D = 1$, we have $\mathrm{W}(\lambda \mid \nu_0, \nu_0^{-1} \lambda_0) = \chi^2(\nu_0, \lambda_0)$.

The Wishart distribution plays a key role in **frequentist statistics** as the distribution of the sample covariance matrix of mean-zero multivariate normal r.v.'s. In **Bayesian statistics**, it arises as the conjugate prior for the precision of a multivariate normal distribution.
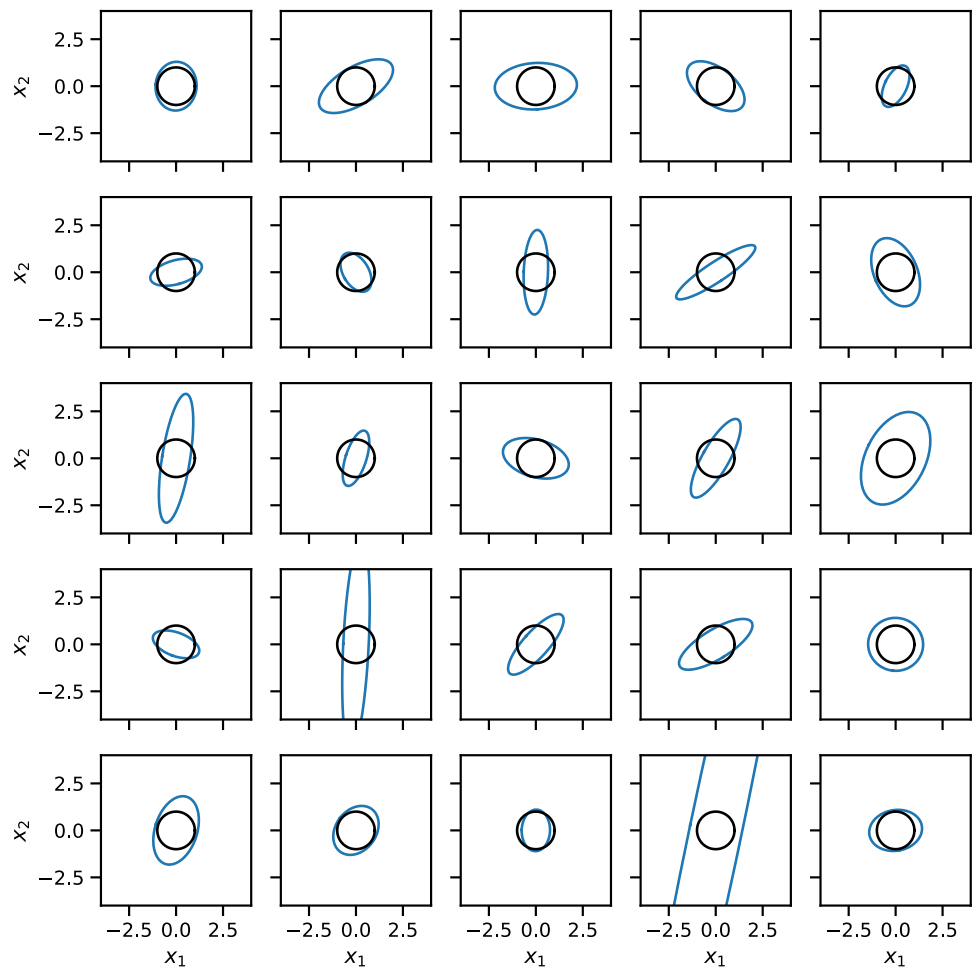
The Wishart Distribution

Figure: Visualizing Λ^{-1} where $\Lambda \sim W(\nu_0, \Lambda_0)$.

Multivariate Normal with Unknown Precision II

The Wishart distribution is a conjugate prior for the precision of a multivariate normal distribution.

$$\Lambda \sim W(\nu_0, \Lambda_0), \tag{55}$$

$$x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Lambda^{-1}). \tag{56}$$

Then, letting $\eta = (\mu, \nu_0, \Lambda_0)$,

$$p(\Lambda \mid X, \eta) \propto W(\Lambda \mid \nu_0, \Lambda_0) \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \Lambda^{-1}) \tag{57}$$

$$\propto |\Lambda|^{\frac{\nu_0 - D - 1}{2}} e^{-\frac{1}{2} \operatorname{Tr}(\Lambda_0^{-1} \Lambda)} \prod_{n=1}^{N} |\Lambda|^{\frac{1}{2}} e^{-\frac{1}{2}(x_n - \mu)^\top \Lambda (x_n - \mu)} \tag{58}$$

$$\propto |\Lambda|^{\frac{\nu_0 + N - D - 1}{2}} \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left(\left[\Lambda_0^{-1} + \sum_{n=1}^{N} (x_n - \mu)(x_n - \mu)^\top \right] \Lambda \right) \right\} \tag{59}$$

(handwritten annotation):
$$\operatorname{Tr}\left((x_n - \mu)^\top \Sigma^{-1} (x_n - \mu) \right)$$
$$= \operatorname{Tr}\left(\Sigma^{-1} (x_n - \mu)(x_n - \mu)^\top \right)$$

Multivariate Normal with Unknown Precision III

We recognize this as yet another Wishart distribution,

$$p(\Lambda \mid X, \eta) \propto \mathrm{W}(\Lambda \mid \nu_N, \Lambda_N), \tag{60}$$

where

$$\nu_N = \nu_0 + N \tag{61}$$

$$\Lambda_N = \left[\Lambda_0^{-1} + \sum_{n=1}^{N} (x_n - \mu)(x_n - \mu)^\top \right]^{-1} \tag{62}$$

Question: What is the posterior mean under the uninformative prior where $\nu_0 \to 0$ or in the large data limit where $N \to \infty$?

$$\Lambda_N \to \left(\sum_n (x_n - \mu)(x_n - \mu) \right)^{-1}$$

$$\mathbb{E}[\Lambda_N] \to N \Lambda_N$$

Visualizing the Posterior
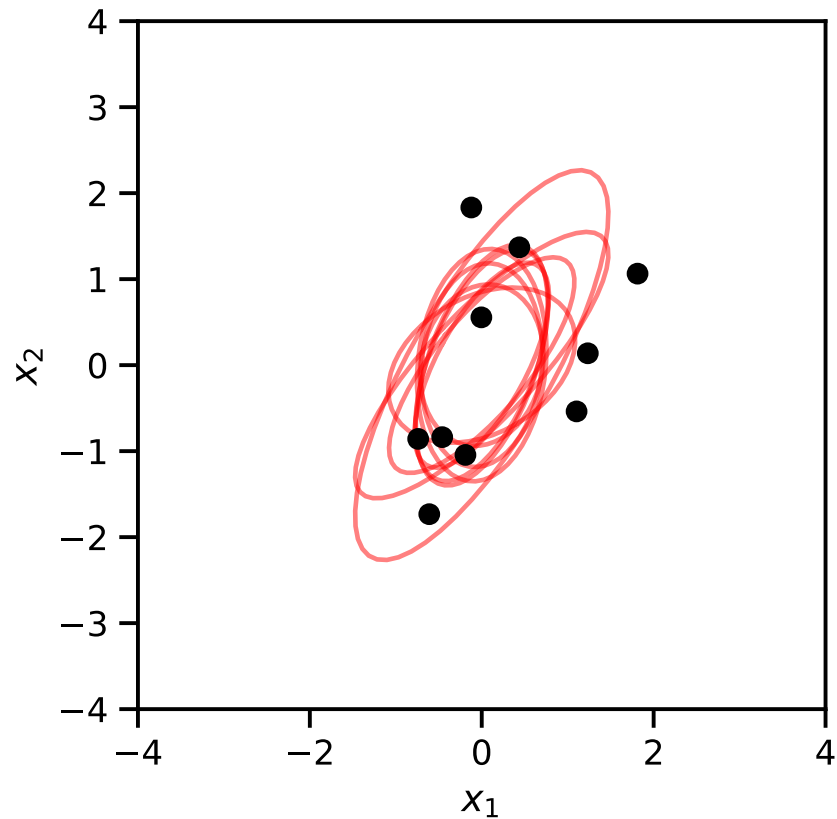
Figure: Visualizing $\mathbf{\Lambda}^{-1}$ under the posterior $\mathbf{\Lambda} \sim \mathrm{W}(\nu_N, \mathbf{\Lambda}_N)$.

Multivariate Normal with Unknown *Covariance*

As with the scaled inverse chi-squared distribution in the univariate case, we define the **inverse Wishart** distribution as,

$$\Lambda \sim W(\nu_0, \Lambda_0) \iff \Sigma = \Lambda^{-1} \sim IW(\nu_0, \Sigma_0). \tag{63}$$

where $\Sigma_0 = \Lambda_0^{-1}$.

It too is a distribution on the set of positive definite matrices. Its pdf is,

$$IW(\Sigma \mid \nu_0, \Sigma_0) = \frac{|\Sigma_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 D}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right)} |\Sigma|^{-\frac{\nu_0 + D + 1}{2}} e^{-\frac{1}{2} \operatorname{Tr}(\Sigma_0 \Sigma^{-1})} \tag{64}$$

where $\nu_0 > 0$ specifies the **degrees of freedom** and $\Sigma_0 \in \mathcal{S}_D$ is the **scale**.

Its mean is $\frac{\Sigma_0}{\nu - D - 1}$ for $\nu_0 \geq (D + 1)$ and its mode is $\frac{\Sigma_0}{\nu + D + 1}$.

In the univariate case where $D = 1$, we have $IW(\sigma^2 \mid \nu_0, \nu_0 \sigma_0^2) = \chi^{-2}(\nu_0, \sigma_0^2)$.

Multivariate Normal with Unknown *Covariance*

Exercise: Consider the model,

$$\Sigma \sim \text{IW}(\nu_0, \Sigma_0), \tag{65}$$

$$\boldsymbol{x}_n \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \Sigma). \tag{66}$$

Show that,

$$p(\Sigma \mid \boldsymbol{X}, \boldsymbol{\eta}) = \text{IW}(\nu_N, \Sigma_N) \tag{67}$$

where

$$\nu_N = \nu_0 + N \tag{68}$$

$$\Sigma_N = \Sigma_0 + \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top. \tag{69}$$

Multivariate Normal with Unknown Mean and Covariance

As you can probably guess, the normal-inverse-chi-squared distribution for the univariate case generalizes to the multivariate case too. Here, the conjugate prior for the mean and covariance is the **normal inverse Wishart** (NIW) distribution,

$$\mathrm{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Sigma}_0) = \mathrm{IW}(\boldsymbol{\Sigma} \mid \nu_0, \boldsymbol{\Sigma}_0)\, \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0). \tag{70}$$

Exercise: Show that under this prior the posterior is,

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X}, \boldsymbol{\eta}) = \mathrm{NIW}(\boldsymbol{\mu}_N, \kappa_N, \nu_N, \boldsymbol{\Sigma}_N) \tag{71}$$

where

$$\nu_N = \nu_0 + N \tag{72}$$

$$\kappa_N = \kappa_0 + N \tag{73}$$

$$\boldsymbol{\mu}_N = \frac{1}{\kappa_N}\left(\kappa_0 \boldsymbol{\mu}_0 + \sum_{n=1}^{N} \boldsymbol{x}_n\right) \tag{74}$$

$$\boldsymbol{\Sigma}_N = \boldsymbol{\Sigma}_0 + \kappa_0 \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^\top + \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top - \kappa_N \boldsymbol{\mu}_N \boldsymbol{\mu}_N^\top \tag{75}$$
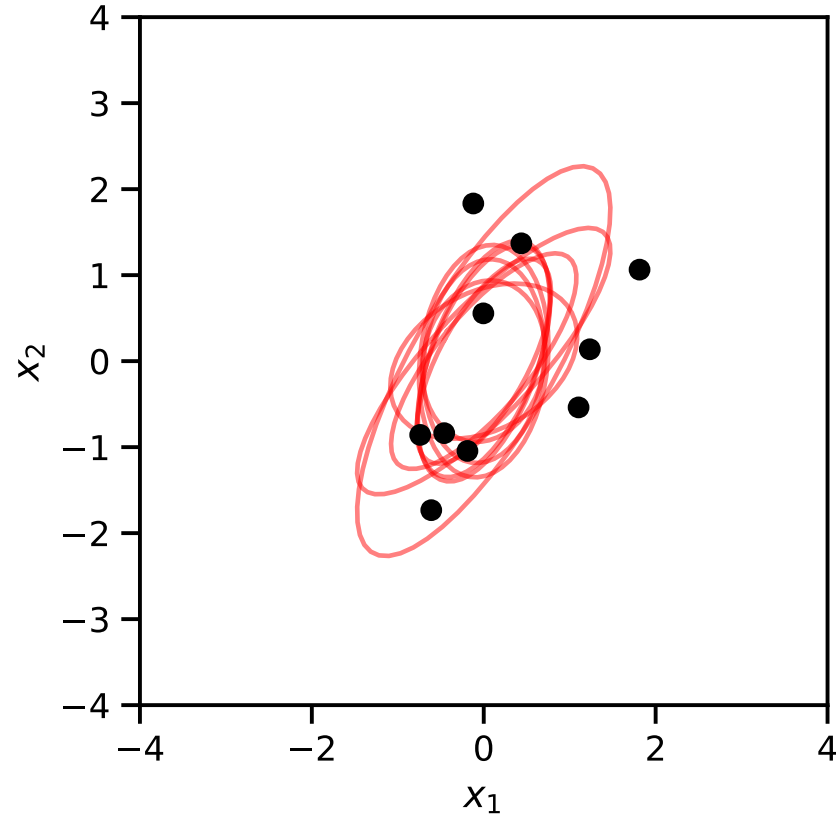
Visualizing the Posterior

Figure: Visualizing $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}^{-1}$ under the posterior $\boldsymbol{\mu}, \boldsymbol{\Lambda} \sim \mathrm{NW}(\boldsymbol{\mu}_N, \kappa_N, \nu_N, \boldsymbol{\Lambda}_N)$.

Posterior Marginals

And again, just like in the univariate case, we find that the posterior marginal of $\boldsymbol{\mu} \in \mathbb{R}^D$ is

$$p(\boldsymbol{\mu} \mid \boldsymbol{X}, \boldsymbol{\eta}) = \int p(\boldsymbol{\mu}, \Sigma \mid \boldsymbol{X}, \boldsymbol{\eta}) \, \mathrm{d}\Sigma \tag{76}$$

$$= \int \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_N, \Sigma/\kappa_N) \, \mathrm{IW}(\Sigma \mid \nu_N, \Sigma_N) \, \mathrm{d}\Sigma \tag{77}$$

$$= \mathrm{St}\left(\nu_N, \boldsymbol{\mu}_N, \frac{1}{\kappa_N(\nu_N - D + 1)} \Sigma_N \right) \tag{78}$$

where St denotes the **multivariate Student's t** distribution with density,

$$\mathrm{St}(\boldsymbol{x} \mid \nu, \boldsymbol{\mu}, \Sigma) = \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})}(\nu\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\left[1 + \frac{\Delta^2}{\nu} \right]^{-\frac{\nu+D}{2}} \tag{79}$$

where $\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})^2$ is the squared Mahalanobis distance.

The mean of the multivariate Student's t is $\boldsymbol{\mu}$ and the covariance is $\frac{\nu}{\nu+2}\Sigma$.