Ch. 7 : Multivariate normal

- So far, our data have been univariate, i.e., $Y_i \in \mathbb{R}$ for all samples $Y_1, \dots, Y_n$.
- In many applications, we collect multiple measurements for each sample, e.g., height, weight, and age of people in a population.

$$\Rightarrow \vec{Y}_i = (Y_{i1}, \dots, Y_{id}) \in \mathbb{R}^d \quad \forall i = 1, \dots, n.$$

Preliminaries

- A d-dimensional random vector is a vector $X = (X_1, \dots, X_d)$ where each X_j is a random variable.
- The <u>mean</u> of X is the d-dimensional vector

$$\mathbb{E}[X] = \left(\mathbb{E}[X_1], \dots, \mathbb{E}[X_d]\right) \in \mathbb{R}^d$$

- The <u>covariance</u> of X is the $d \times d$ matrix

$$\mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathrm{Var}(X_d) \end{pmatrix}$$

where $\text{Cov}(X_j, X_k) = \mathbb{E}[X_j X_k] - \mathbb{E}[X_j]\mathbb{E}[X_k]$ is the

<u>covariance</u> between X_j and X_k.

- Note that $\text{Cov}(X_j, X_k) = \text{Cov}(X_k, X_j) \Rightarrow \text{Cov}(X)$ is symmetric
- $\text{Cov}(X_j, X_j) = \text{Var}(X_j)$

- <u>Correlation</u> is a normalized version of $\text{Cov}(X_j, X_k)$:

$$\text{Cor}(X_j, X_k) = \frac{\text{Cov}(X_j, X_k)}{\sqrt{\text{Var}(X_j)\,\text{Var}(X_k)}} \in [-1, 1].$$

Multivariate normal distribution

- One of the most common models for multivariate data is the multivariate normal aka multivariate Gaussian distribution.
- Specifically, a r.v. $Y = (Y_1, \ldots, Y_d)$ taking values in $\mathbb{R}^d$ has a multivariate normal distribution w/ mean $\theta \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$ if its density has the form

$$p(y|\theta,\Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(y-\theta)^T \Sigma^{-1} (y-\theta)}$$

where

- $y = (y_1, \cdots, y_d)^T \in \mathbb{R}^d$
- $\theta = (\theta_1, \cdots, \theta_d)^T \in \mathbb{R}^d$
- $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$ is a symmetric, invertible $d \times d$ matrix

- $\det(\Sigma)$ is the determinant of Σ
- Σ^{-1} is the inverse of Σ, i.e.

$$\Sigma\Sigma^{-1} = \Sigma^{-1}\Sigma = I_d = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

- x^T is the transpose of $x \in \mathbb{R}^d$

- Convenient property of multivariate normal $N(\theta, \Sigma)$ is that its marginals are univariate normal:

$$Y = (Y_1, \dots, Y_d) \sim N_d(\theta, \Sigma)$$

$$\Rightarrow Y_j \sim N_1(\theta_j, \sigma_j^2) \quad \forall j = 1, \dots, d.$$

2-dimensional example

- Study in which people (i) take a test, (ii) receive some instruction, (iii) take another test. So

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \text{Score on test I} \\ \text{Score on test II} \end{pmatrix}$$

- Let $\theta_1 = \mathbb{E}[Y_1]$ and $\theta_2 = \mathbb{E}[Y_2]$. So

$$\mathbb{E}[Y] = \begin{pmatrix} \mathbb{E}[Y_1] \\ \mathbb{E}[Y_2] \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \theta$$

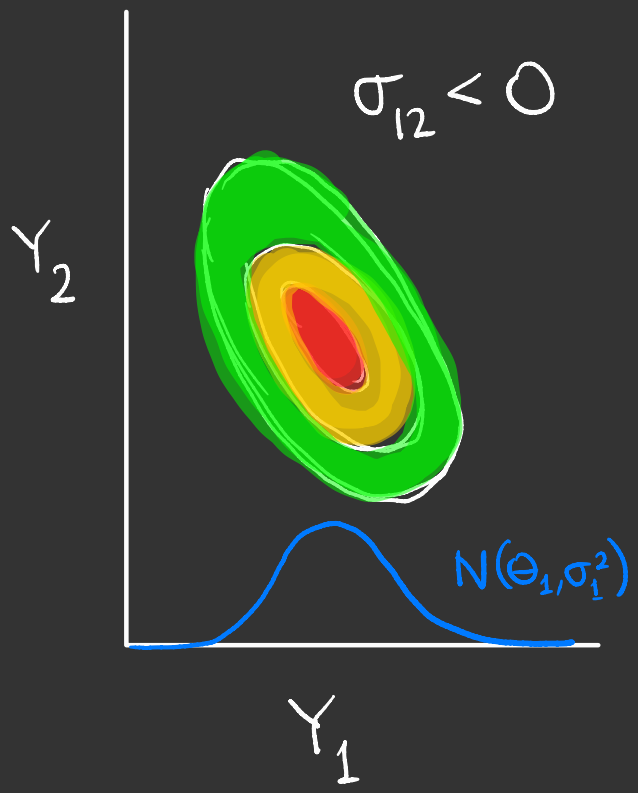
Average on test I

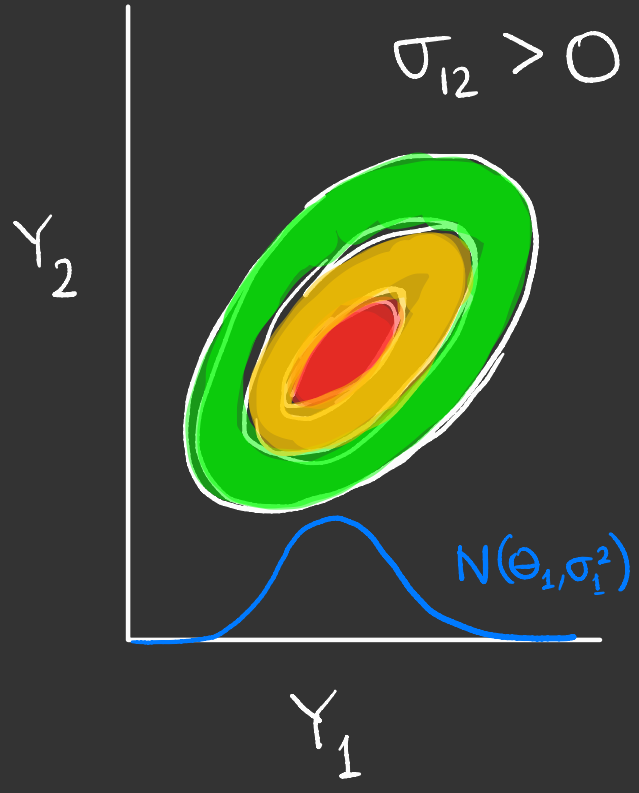
Average on test II

- Let $\sigma_j^2 = \text{Var}(Y_j)$ and $\sigma_{12} = \sigma_{21} = \text{Cov}(Y_1, Y_2)$.

$$\Rightarrow \Sigma = \text{Cov}(Y) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}.$$

  - σ_1^2 = variance of test I
  - σ_2^2 = variance of test II
  - σ_{12} = covariance between test I and II scores


$\sigma_{12} < 0$
Y_2
$N(\theta_1, \sigma_1^2)$
Y_1


$\sigma_{12} > 0$
Y_2
$N(\theta_1, \sigma_1^2)$
Y_1

Bayesian model

- Want to infer parameters Θ and Σ from data
  - $\Theta \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ (symmetric) $\Rightarrow d + \frac{d(d+1)}{2}$ parameters
  - Example. $d=3 \Rightarrow 3 + \frac{3(4)}{2} = 9$ parameters:

$$\Theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \cdot & \sigma_2^2 & \sigma_{23} \\ \cdot & \cdot & \sigma_3^2 \end{pmatrix}$$

In the 1-dimensional case, we saw

- $Y_1, \ldots, Y_n \mid \theta, \sigma^2 \overset{\text{iid}}{\sim} N_1(\theta, \sigma^2)$ ← Likelihood
- $\theta \sim N_1(\mu_0, \tau_0^2)$ ← Prior for θ
- $\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$ ← Prior for σ^2

yields full conditionals

- $\theta \mid \sigma^2, \vec{y}_n \sim N_1(\mu_n, \tau_n^2)$
- $\sigma^2 \mid \theta, \vec{y}_n \sim IG\left(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2\right)$

Note that the full conditionals $p(\theta \mid \sigma^2, \vec{y}_n)$ and $p(\sigma^2 \mid \theta, \vec{y}_n)$ belong to the same family as their respective priors. This is called <u>semi-conjugacy</u>.

In the d-dimensional case:

- $Y_1, \ldots, Y_n \mid \theta, \Sigma \overset{iid}{\sim} N_d(\theta, \Sigma)$ ← Likelihood
- $\theta \sim N_d(\mu_0, \Lambda_0)$ ← Prior for θ
- $\Sigma \sim IW(\nu_0, S_0^{-1})$ ← Prior for Σ

where IW is the inverse-Wishart distribution (more on this below)

As in the univariate case, the above model is semi-conjugate with full conditionals

- $\theta | \Sigma, \vec{y}_n \sim N_d(\mu_n, \Lambda_n)$
  - $\Lambda_n = \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)^{-1}$
  - $\mu_n = \Lambda_n\left(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}\right)$

- $\Sigma | \theta, \vec{y}_n \sim IW(\nu_n, S_n^{-1})$
  - $\nu_n = \nu_0 + n$
  - $S_n = S_0 + S_\theta$ where

$$S_\theta = \sum_{i=1}^{n} (y_i - \theta)(y_i - \theta)^T$$

- Full conditionals $\Longrightarrow$ Gibbs sampler

Details about the above model

- Covariance matrices are symmetric and positive definite, i.e., $x^T \Sigma x > 0 \quad \forall x \in \mathbb{R}^d \setminus \{0\}$.

  $\Rightarrow$ The prior on Σ must be a probability distribution on the space of pos. def. matrices

  $\Rightarrow$ The Wishart distribution is a convenient choice of distribution over pos. def. matrices. It is a multivariate generalization of the

gamma distribution.

$\Rightarrow$ We saw that if $\frac{1}{X} \sim \text{Gamma}(a, b)$, then $X \sim \text{InverseGamma}(a, b)$. Similarly, if $\Sigma^{-1} \sim \text{Wishart}(\nu_0, S_0)$, then $\Sigma \sim IW(\nu_0, S_0)$.

That is, if the <u>precision matrix</u> Σ^{-1} has a Wishart distribution, then the covariance Σ has an inverse Wishart distribution.

2-dimensional example continued