Agenda

- Examples where estimates of the mean and of the covariance matrix are used:
  - Comparing two samples
  - Markowitz problem in finance
- Maximum Likelihood Estimation for the Multivariate Normal Distribution
Examples
Example 1 – Comparing Two Samples

- Clinical study of subjects sampled from two populations:
  - Normal
  - Alzheimer’s disease
- Interested in developing a method of detection based on functional MRI.
Example 1 – Alzheimer’s Disease

- Disease process is very slow and can begin several years (even decades) before clinical diagnosis.
- Recent studies have examined the use of functional MRI to detect impairment of cognitive function due to AD.
Example 1 – Hypothetical Study

- Subject instructed to perform some task while lying in an MRI scanner.
- fMRI produces an image of the pattern of activation in the subjects brain during the task.
- Trial is repeated over multiple subjects from both groups resulting in data matrices $X$ and $Y$ corresponding to the two groups.
Example 1 – Question

- Is the mean pattern of activation equal across the two populations?
- If we knew the population means $\mu_A$ and $\mu_B$, we could just compare them directly.
Example 1 – Inference based on the sample

- Instead we have samples from the populations and estimates based on the sample means:

\[ \hat{\mu}_A = \bar{X} \]
\[ \hat{\mu}_B = \bar{Y} \]

- Is the observed difference due to sampling error?

\[ \hat{\mu}_A - \hat{\mu}_B = \bar{X} - \bar{Y} \]
Example 1 – Variance of the Difference

- We assume the samples are i.i.d. – subjects are independent, a sampled from the same populations.

\[
\text{Var}(\mu_A - \mu_B) = \text{Var}(\mu_A) + \text{Var}(\mu_B) \\
= \frac{1}{n_A} \Sigma_A + \frac{1}{n_B} \Sigma_B
\]

- Variance is a measure of our uncertainty about the observed difference.

- Unfortunately, \( \Sigma_A \) and \( \Sigma_B \) are unknown!
Example 2 – Portfolio Optimization

- You have the opportunity to invest in $p$ risky assets over one period of time.
- Your portfolio is represented by a $p$–vector of weights $\mathbf{w}$ corresponding to the proportion of the portfolio invested in each asset.
- You may go long or short on each asset, but your weights must sum to 1:

$$\langle 1_p, \mathbf{w} \rangle = \sum_{i=1}^{p} w_i = 1.$$
The rate of returns on the assets is given by the p-variate random vector

\[ X = (X_1, X_2, \ldots, X_p). \]

After one period your portfolio’s return is

\[ \langle X, w \rangle = \sum_{i=1}^{p} w_i X_i. \]
Example 2 – Portfolio Optimization (contd)

- Ideally, the mean returns are known and represented by
  \[ \mu = \mathbb{E}(X), \]
  and the covariance between the returns on the assets is also known and represented by
  \[ \Sigma = \text{Var}(X). \]

- Then your portfolio’s expected return is
  \[ \mathbb{E} \langle \mathbf{w}, X \rangle = \langle \mathbf{w}, \mu \rangle. \]
Example 2 – Risk

- The rate of returns is random. There is a risk that your return will be much less than you expect.
- We can use variance as a measure of risk. The variance of your return is

$$\text{Var}(\langle \mathbf{w}, \mathbf{X} \rangle) = \mathbf{w}^T \Sigma \mathbf{w}.$$
Example 2 – Markowitz Problem

- You want high return, but also low risk.
- Markowitz:
  - Minimize variance of the portfolio return subject to an expected level of return $\mu_p$. 
Minimize

\[ \frac{1}{2} \mathbf{w}^T \Sigma \mathbf{w} \]

subject to \( \langle \mu, \mathbf{w} \rangle = \mu_p \) and \( \langle 1_p, \mathbf{w} \rangle = 1 \)

▶ In practice:
  ▶ Mean rates of return \( \mu \) are unknown.
  ▶ Covariance matrix of the assets \( \Sigma \) also unknown. You want high return, but also low risk.
  ▶ Replace \( \mu \) and \( \Sigma \) by estimates.
  ▶ Danger: Possible that risk is underestimated – especially when \( p \) is large.
Multivariate Normal MLE
Suppose that the rows of $X$ represent a random sample from a multivariate Normal population (distribution).

- Rows of $X$ are i.i.d. $\text{MVN}_p(\mu, \Sigma)$

- Fitting a multivariate Normal distribution is equivalent to estimating $(\mu, \Sigma)$. 
Sometimes the data can be transformed so that its distribution is approximately Multivariate Normal.

- log transformation
- Box–Cox transformation
Maximum Likelihood Estimation

- Abbreviated MLE
- Principle for estimation proposed by Fisher.
- “Choose the value of the parameter that makes the observed data most likely.”
Parametric Estimation

- Suppose data are drawn i.i.d. from a multivariate distribution with density belonging to the model (family of densities):

\[ f_\theta(\cdot) \in \{ f_\theta(\cdot) : \theta \in \Theta \} \]

- Goal: Estimate \( \theta_0 \)
Maximum Likelihood Estimation

- “Choose the value of the parameter that makes the observed data most likely”

\[ \hat{\theta} = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta \mid X_1, \ldots, X_n) \]

- Taking the logarithm of the likelihood:

\[ \hat{\theta} = \arg \max_{\theta \in \Theta} \log \mathcal{L}(\theta \mid X_1, \ldots, X_n) = \arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \log f_{\theta}(X_i). \]
Normal MLE

Assume that the mean and covariance matrix are both unknown. Then the parameter space is

$$\Theta = \mathbb{R}^p \times \{A \in \mathbb{R}^{p \times p} : A > 0\}.$$ 

Product space of $p$–vectors and $p \times p$ positive definite matrices.
Normal Log-likelihood

\[
\log \mathcal{L}(\theta \mid X) = \prod_{i=1}^{n} \left\{ -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log(\det(\Sigma)) - \frac{1}{2} (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) \right\}.
\]
MLE of the Mean

- Maximizing the log-likelihood over \( \mu \) is equivalent to maximizing the following function over \( \mu \):

\[
\frac{1}{2} \text{Tr}[\Sigma^{-1}(\bar{X} - \mu)(\bar{X} - \mu)^T]
\]

- Its gradient with respect to \( \mu \) is

\[
n\Sigma^{-1}(\bar{X} - \mu)
\]

- Setting the gradient equal to 0 we have

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.
\]
MLE of the Covariance Matrix

- We may reparameterize by replacing $\Sigma$ by its inverse: $\Omega = \Sigma^{-1}$
- This provides a 1-to-1 transformation of the parameter space, so the MLE is preserved.
MLE of the Covariance Matrix (contd)

\[
\max_{\mu \in \mathbb{R}^p, \Omega > 0} \log L(\mu, \Omega \mid X) = \max_{\Omega > 0} \left(-\frac{np}{2} \log(2\pi) + \frac{n}{2} \log |\Omega| - \frac{n}{2} \text{Tr}(\Omega S_n) - \frac{n}{2} \text{Tr}(\Omega (\bar{X} - \mu)(\bar{X} - \mu)^T)\right)
\]

▶ The gradient of the objective function wrt \(\Omega\) is

\[
\frac{n}{2} (2M - \text{diag}(M))
\]

where \(M = \Sigma - S_n - (\bar{X} - \mu)^T (\bar{X} - \mu)\).

▶ Setting it equal to 0 gives the MLE:

letting \(\hat{\mu} = \bar{X}\) and

\[
\hat{\Sigma}^{-1} = \hat{\Omega} = S_n^{-1} \Rightarrow \hat{\Sigma} = S_n.
\]
Notes

- See Section 4.2.2 p103 of Mardia, Kent & Bibby for derivation of the gradient of the log determinant function: $\log(\det(A))$
- See Theorem 4.7.1 of Mardia, Kent & Bibby for joint concavity of the log-likelihood function of MVN—proof left to student
Summary

- MVN parameter estimates often used as input for other problems.
- MLE for the Multivariate Normal Distribution is just the sample mean and the sample covariance matrix (when it is nonsingular).