Maximum Likelihood Theory

Max Turgeon

STAT 4690–Applied Multivariate Analysis

Sufficient Statistics i

- We saw in the previous lecture that the multivariate normal distribution is completely determined by its mean vector $\mu \in \mathbb{R}^p$ and its covariance matrix Σ.
- Therefore, given a sample $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim N_p(\mu, \Sigma)$ $(n > p)$, we only need to estimate (μ, Σ).
    - Obvious candidates: sample mean $\bar{\mathbf{Y}}$ and sample covariance S_n.

Sufficient Statistics ii

- Write down the *likelihood*:

$$
\begin{aligned}
L &= \prod_{i=1}^n \left(\frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{y}_i - \mu)^T \Sigma^{-1} (\mathbf{y}_i - \mu) \right) \right) \\
&= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \mu)^T \Sigma^{-1} (\mathbf{y}_i - \mu) \right)
\end{aligned}
$$

- If we take the (natural) logarithm of L and drop any term that does not depend on (μ, Σ), we get

$$
\ell = -\frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \mu)^T \Sigma^{-1} (\mathbf{y}_i - \mu).
$$

Sufficient Statistics iii

- If we can re-express the second summand in terms of $\bar{\mathbf{Y}}$ and S_n, by the Fisher-Neyman factorization theorem, we will then know that $(\bar{\mathbf{Y}}, S_n)$ is jointly **sufficient** for (μ, Σ).
- First, we have

Sufficient Statistics iv

$$\begin{aligned}
\sum_{i=1}^{n}(\mathbf{y}_i-\mu)(\mathbf{y}_i-\mu)^T &= \sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}}+\bar{\mathbf{y}}-\mu)(\mathbf{y}_i-\bar{\mathbf{y}}+\bar{\mathbf{y}}-\mu)^T \\
&= \sum_{i=1}^{n}\Big((\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})^T + (\mathbf{y}_i-\bar{\mathbf{y}})(\bar{\mathbf{y}}-\mu)^T \\
&\quad + (\bar{\mathbf{y}}-\mu)(\mathbf{y}_i-\bar{\mathbf{y}})^T + (\bar{\mathbf{y}}-\mu)(\bar{\mathbf{y}}-\mu)^T\Big) \\
&= \sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})^T + n(\bar{\mathbf{y}}-\mu)(\bar{\mathbf{y}}-\mu)^T \\
&= (n-1)S_n + n(\bar{\mathbf{y}}-\mu)(\bar{\mathbf{y}}-\mu)^T.
\end{aligned}$$

Sufficient Statistics v

- Next, using the fact that $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$, we have

Sufficient Statistics vi

$$\begin{aligned}
\sum_{i=1}^n (\mathbf{y}_i - \mu)^T \Sigma^{-1} (\mathbf{y}_i - \mu) &= \mathrm{tr}\left(\sum_{i=1}^n (\mathbf{y}_i - \mu)^T \Sigma^{-1} (\mathbf{y}_i - \mu) \right) \\
&= \mathrm{tr}\left(\sum_{i=1}^n \Sigma^{-1} (\mathbf{y}_i - \mu)(\mathbf{y}_i - \mu)^T \right) \\
&= \mathrm{tr}\left(\Sigma^{-1} \sum_{i=1}^n (\mathbf{y}_i - \mu)(\mathbf{y}_i - \mu)^T \right) \\
&= (n-1)\mathrm{tr}\left(\Sigma^{-1} S_n \right) \\
&\qquad + n\mathrm{tr}\left(\Sigma^{-1} (\bar{\mathbf{y}} - \mu)(\bar{\mathbf{y}} - \mu)^T \right) \\
&= (n-1)\mathrm{tr}\left(\Sigma^{-1} S_n \right) \\
&\qquad + n(\bar{\mathbf{y}} - \mu)^T \Sigma^{-1} (\bar{\mathbf{y}} - \mu).
\end{aligned}$$

Maximum Likelihood Estimators

- Going back to the log-likelihood, we get:
  $$\ell = -\frac{n}{2}\log|\Sigma| - \frac{(n-1)}{2}\mathrm{tr}\left(\Sigma^{-1}S_n\right) - \frac{n}{2}(\bar{\mathbf{y}}-\mu)^T\Sigma^{-1}(\bar{\mathbf{y}}-\mu).$$
- Since Σ^{-1} is positive definite, for Σ fixed, the log-likelihood is maximised at
  $$\hat{\mu} = \bar{\mathbf{y}}.$$
- With extra effort, it can be shown that $-\log|\Sigma| - \frac{(n-1)}{n}\mathrm{tr}\left(\Sigma^{-1}S_n\right)$ is maximised at
  $$\hat{\Sigma} = \frac{(n-1)}{n}S_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T.$$
- *In other words*: $(\bar{\mathbf{Y}}, \hat{\Sigma})$ are the **maximum likelihood estimators** for (μ, Σ).

Maximum Likelihood Estimators

- Since the multivariate normal density is "well-behaved", we can deduce the usual properties:
  - **Consistency**: $(\bar{\mathbf{Y}}, \hat{\Sigma})$ converges in probability to (μ, Σ).
  - **Efficiency**: Asymptotically, the covariance of $(\bar{\mathbf{Y}}, \hat{\Sigma})$ achieves the Cramér-Rao lower bound.
  - **Invariance**: For any transformation $(g(\mu), G(\Sigma))$ of (μ, Σ), its MLE is $(g(\bar{\mathbf{Y}}), G(\hat{\Sigma}))$.

Visualizing the likelihood

```r
library(mvtnorm)
set.seed(123)

n <- 50; p <- 2

mu <- c(1, 2)
Sigma <- matrix(c(1, 0.5, 0.5, 1), ncol = p)

Y <- rmvnorm(n, mean = mu, sigma = Sigma)
```

Visualizing the likelihood

```r
loglik <- function(mu, sigma, data = Y) {
  # Compute quantities
  y_bar <- colMeans(Y)
  Sn <- (n-1)*cov(Y)/n
  Sigma_inv <- solve(sigma)

  # Compute quadratic form
  quad_form <- drop(t(y_bar - mu) %*% Sigma_inv %*%
                      (y_bar - mu))

  -0.5*n*(log(det(sigma)) +
            sum(diag(Sigma_inv %*% Sn)) +
            quad_form)
}
```

```r
grid_xy <- expand.grid(seq(0.5, 1.5,
                           length.out = 32),
                       seq(1, 3,
                           length.out = 32))

contours <- purrr::map_df(seq_len(nrow(grid_xy)),
                          function(i) {
  # Where we will evaluate loglik
  mu_obs <- as.numeric(grid_xy[i,])
  # Evaluate at the pop covariance
  z <- loglik(mu_obs, sigma = Sigma)
  # Output data.frame
  data.frame(x = mu_obs[1],
             y = mu_obs[2],
             z = z)
})
```

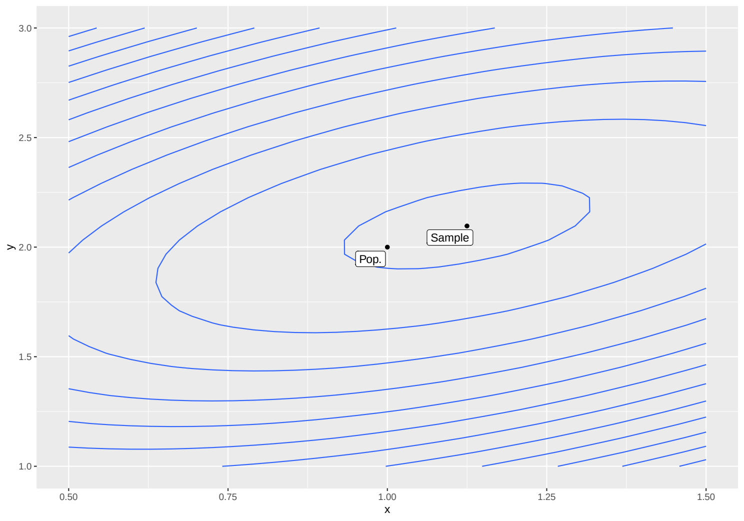
Visualizing the likelihood i

```r
library(tidyverse)
library(ggrepel)
# Create df with pop and sample means
data_means <- data.frame(x = c(mu[1], mean(Y[,1])),
                         y = c(mu[2], mean(Y[,2])),
                         label = c("Pop.", "Sample"))
```

Visualizing the likelihood ii

```r
contours %>%
  ggplot(aes(x, y)) +
  geom_contour(aes(z = z)) +
  geom_point(data = data_means) +
  geom_label_repel(data = data_means,
                   aes(label = label))
```

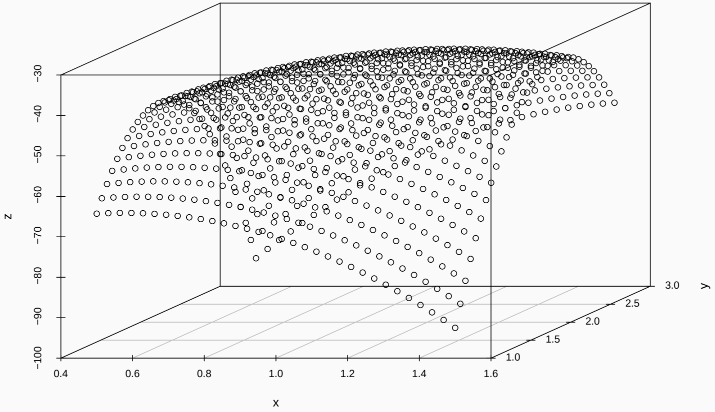
Visualizing the likelihood iii



Visualizing the likelihood iv

```r
library(scatterplot3d)
with(contours, scatterplot3d(x, y, z))
```

Visualizing the likelihood v



Sampling Distributions

- Recall the univariate case:
  - $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$;
  - $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$;
  - $\bar{X}$ and s^2 are independent.
- In the multivariate case, we have similar results:
  - $\bar{\mathbf{Y}} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$;
  - $(n-1)S_n = n\hat{\Sigma}$ follows a *Wishart* distribution with $n-1$ degrees of freedom;
  - $\bar{\mathbf{Y}}$ and S_n are independent.

Wishart Distribution

- Suppose $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \sim N_p(0, \Sigma)$ are independently distributed. Then we say that

$$W = \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i^T$$

  follows a *Wishart distribution* $W_n(\Sigma)$ with n degrees of freedom.
- Note that since $E(\mathbf{Z}_i \mathbf{Z}_i^T) = \Sigma$, we have $E(W) = n\Sigma$.
- From the previous slide: $\sum_{i=1}^{n} (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$ has the same distribution as $\sum_{i=1}^{n-1} \mathbf{Z}_i \mathbf{Z}_i^T$ for some choice of $\mathbf{Z}_1, \ldots, \mathbf{Z}_{n-1} \sim N_p(0, \Sigma)$.

Useful Properties

- If $W_1 \sim W_{n_1}(\Sigma)$ and $W_2 \sim W_{n_2}(\Sigma)$ are independent, then
$$W_1 + W_2 \sim W_{n_1+n_2}(\Sigma).$$
- If $W \sim W_n(\Sigma)$ and C is $q \times p$, then
$$CWC^T \sim W_n(C\Sigma C^T).$$

Density function

- Let Σ be a fixed $p \times p$ positive definite matrix. The density of the Wishart distribution with n degrees of freedom, with $n \geq p$, is given by

$$w_n(A; \Sigma) = \frac{|A|^{(n-p-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}A)\right)}{2^{np/2}\pi^{p(p-1)/4}|\Sigma|^{n/2}\prod_{i=1}^{p}\left(\frac{1}{2}(n-i+1)\right)},$$

where A is ranging over all $p \times p$ positive definite matrices.

Eigenvalue density function

- For a random matrix $A \sim W_n(I_p)$ with $n \geq p$, the joint distribution of its eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ has density

$$C_{n,p} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\right) \prod_{i=1}^{p} \lambda_i^{(n-p-1)/2} \prod i < j|\lambda_i - \lambda_j|,$$

  for some constant $C_{n,p}$.
- We will study this distribution in STAT 7200–Multivariate Analysis I.