Econ 6190: Econometrics I

Estimation

Chen Qiu

Cornell Economics

2024 Fall

Contents

Reference

- Hansen Ch. 10 and 11

1. Maximum Likelihood Estimation

Motivation

- Parameter estimation in complete probability models
  - Structural economic modeling
- Maximum likelihood estimation is very popular for these **parametric models**
- Advantage: wide applicability (many different data types); can handle complicated data and models
- Disadvantage: strong distributional assumption

Parametric model

- A **parametric model** for X is the assumption that X has a density or probability mass function $f(x|\theta)$ with **known** form of f but with **unknown** parameter vector $\theta \in \Theta$
- Example: Assume $X \sim \mathrm{N}(\mu, \sigma^2)$, which has density $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$. The parameters are $\mu \in \mathbb{R}, \sigma^2 > 0$
- In this course we focus on unconditional distributions: $f(x|\theta)$ does not depend on conditioning variables
- In many economic modeling, we focus on conditional distributions (next semester)

Correct specification

- **Definition**: A model is **correctly specified** when there is a **unique** parameter value $\theta_0 \in \Theta$ such that $f(x|\theta_0)$ coincides with the true density or pmf of X

  This parameter value θ_0 is called the true parameter value

  The parameter θ_0 is **unique** if there is *no* other θ such that $f(x|\theta_0) = f(x|\theta)$

- A model is **mis-specified** if there is *no* parameter value $\theta \in \Theta$ such that $f(x|\theta)$ coincides with the true density or pmf of X

Example

- Suppose true model is $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$
- The model is

$$f(x|p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = p \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2} + (1-p) \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2}$$

- The model is "correct" since it includes $f(x)$ as a special case
- However the "true" parameter is not unique, as they include

$$\begin{gathered}(p, 0, 1, 0, 1) \text{ for any } p \\ \left(1, 0, 1, \mu_2, \sigma_2^2\right) \text{ for any } \mu_2, \sigma_2^2 \\ \left(0, \mu_1, \sigma_1^2, 0, 1\right) \text{ for any } \mu_1, \sigma_1^2\end{gathered}$$

- Hence the model is not correctly specified

Likelihood

- The joint pdf or pmf of i.i.d $\{X_1, \ldots X_n\}$ given θ is a function

$$f(x_1, x_2, \ldots x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- **Definition**: The **likelihood function** is

$$L_n(\theta) = f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

- The likelihood function
  - is the joint pdf or pmf evaluated at the observed data
  - is viewed as function of θ
  - describes the compatibility of different values of θ with observed data

Maximum Likelihood Estimator (MLE)

- **Definition**: An maximum likelihood estimator $\hat{\theta}$ is the value that maximizes $L_n(\theta)$

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L_n(\theta)$$

or equivalently,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell_n(\theta)$$

where

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

is called the **log likelihood function**

Example: exponential distribution

- Suppose $f(x|\lambda) = \frac{1}{\lambda}\exp(-\frac{x}{\lambda}), x \geq 0, \lambda > 0$
- The log likelihood is

$$\ell_n(\lambda) = \sum_{i=1}^{n}\left(-\log\lambda - \frac{X_i}{\lambda}\right) = -n\log\lambda - n\frac{\bar{X}_n}{\lambda}$$

- FOC is

$$\frac{\partial}{\partial\lambda}\ell_n(\lambda) = -n\frac{1}{\lambda} + n\frac{\bar{X}_n}{\lambda^2}$$

- Setting $\frac{\partial}{\partial\lambda}\ell_n(\lambda)$ equal to zero yields $\hat{\lambda} = \bar{X}_n$
- $\hat{\lambda}$ is indeed a maximizer since

$$\frac{\partial^2}{\partial\lambda^2}\ell_n(\hat{\lambda}) = n\frac{1}{\hat{\lambda}^2} - 2n\frac{\bar{X}_n}{\hat{\lambda}^3} = -\frac{n}{\bar{X}_n^2} < 0$$

Likelihood analog principle

- Why does MLE make sense?
- Define **expected log likelihood function**

$$\ell(\theta) = \mathbb{E}[\log f(X|\theta)]$$

- **Theorem**: When the model is correctly specified, the true parameter θ_0 maximizes $\ell(\theta)$

- **Proof**: For each $\theta \neq \theta_0$

$$\ell(\theta) - \ell(\theta_0) = \mathbb{E}\left[\log\left(\frac{f(X|\theta)}{f(X|\theta_0)}\right)\right] < \log \mathbb{E}\left[\frac{f(X|\theta)}{f(X|\theta_0)}\right] \tag{1}$$

  where the inequality follows from Jensen's inequality and strict inequality holds since log is strictly concave and $\frac{f(X|\theta)}{f(X|\theta_0)}$ is not a constant

- Let the true density of the data be $f(x)$
- Since $f(x|\theta_0) = f(x)$ and $f(x|\theta)$ is a valid density

$$\mathbb{E}\left[\frac{f(X|\theta)}{f(X|\theta_0)}\right] = \int \frac{f(x|\theta)}{f(x|\theta_0)} f(x)dx = \int f(x|\theta)dx = 1 \tag{2}$$

- Conclusion follows by combining (1) and (2)

Evaluation of estimators

- Likelihood function of parametric models provides a way of evaluating their estimators
- Recall $\ell(\theta) = \mathbb{E}[\log f(X|\theta)]$ is the expected log likelihood
- Introduce some terminology
  - log-likelihood at single observation X and true parameter θ_0:
$$\log f(X|\theta_0)$$
  - **Efficient Score**:
$$S = \frac{\partial}{\partial\theta} \log f(X|\theta_0)$$
  - **Fisher Information**
$$\mathscr{F}_{\theta_0} = \mathbb{E}SS'$$

Property of efficient score

- **Theorem**: Assume model is correctly specified, the support of X does not depend on θ, and θ_0 lies in the interior of Θ. Then $\mathbb{E}S = 0$ and $\text{var}(S) = \mathscr{F}_{\theta_0}$

- Proof: By Leibniz rule

$$\begin{aligned}\mathbb{E}S &= \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right] \\ &= \frac{\partial}{\partial\theta}\mathbb{E}\left[\log f(X|\theta_0)\right] \\ &= \frac{\partial}{\partial\theta}\ell(\theta_0) \\ &= 0\end{aligned}$$

  where the last equality holds as θ_0 maximizes $\ell(\theta)$ and θ_0 is in the interior of Θ

- Then $\text{var}(S) = \mathbb{E}\left[(S - \mathbb{E}[S])(S - \mathbb{E}[S])'\right] = \mathbb{E}\left[SS'\right] = \mathscr{F}_{\theta_0}$

Property of Fisher information

- **Theorem [Information Matrix Equality]**

$$\underbrace{\mathbb{E}\left[\frac{\partial \log f(X|\theta_0)}{\partial\theta}\frac{\partial \log f(X|\theta_0)}{\partial\theta'}\right]}_{\text{Fisher information}} = \underbrace{-\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(X|\theta_0)\right]}_{\text{curvature of } \ell(\theta_0)}.$$

That is,

$$\mathscr{F}_{\theta_0} = \mathscr{H}_{\theta_0}$$

where

$$\mathscr{H}_{\theta_0} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(X|\theta_0)\right] = -\frac{\partial^2}{\partial\theta\partial\theta'}\mathbb{E}[\log f(X|\theta_0)] = -\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta_0)$$

is called the **Expected Hessian**

Remarks

- Fisher information is identical to the the curvature of expected log likelihood
- useful for simplifying formula for the asymptotic variance of MLE
- Proof left for homework

Cramér-Rao Lower Bound

- **Theorem**: Assume model is correctly specified, the support of X does not depend on θ, and θ_0 lies in the interior of Θ. If $\tilde{\theta}$ is an unbiased estimator of θ then

$$\text{var}(\tilde{\theta}) \geq (n\mathscr{F}_{\theta_0})^{-1}$$

  $(n\mathscr{F}_{\theta})^{-1}$ is called **Cramér-Rao Lower Bound** (**CRL**)

  An estimator $\tilde{\theta}$ is **Cramér-Rao efficient** if it is unbiased and $\text{var}(\tilde{\theta}) = (n\mathscr{F}_{\theta_0})^{-1}$

- If $\text{var}(\tilde{\theta})$ is a matrix, $\text{var}(\tilde{\theta}) \geq (n\mathscr{F}_{\theta_0})^{-1}$ means

$$\text{var}(\tilde{\theta}) - (n\mathscr{F}_{\theta_0})^{-1} \text{ is positive semidefinite}$$

- Intuition: More curvature of the expected log likelihood $\Rightarrow$ more information $\Rightarrow$ smaller variance bound

Proof

- Write $\mathbf{x} = (x_1, \dots x_n)'$, $\mathbf{X} = (X_1, \dots X_n)'$
- Write the joint density of $\mathbf{X}$ as $f(\mathbf{x}|\theta)$
- Since $\tilde{\theta}$ is an estimator, $\tilde{\theta} = \tilde{\theta}(\mathbf{X})$
- Since $\tilde{\theta}$ is unbiased, it must hold that

$$\theta = \mathbb{E}_\theta[\tilde{\theta}(\mathbf{X})] = \int \tilde{\theta}(\mathbf{x}) f(\mathbf{x}|\theta) d\mathbf{x}$$

for any θ. By taking derivative on both sides

$$\begin{aligned} I &= \int \tilde{\theta}(\mathbf{x}) \frac{\partial}{\partial \theta'} f(\mathbf{x}|\theta) d\mathbf{x} \\ &= \int \tilde{\theta}(\mathbf{x}) \left(\frac{\partial}{\partial \theta'} \log f(\mathbf{x}|\theta) \right) f(\mathbf{x}|\theta) d\mathbf{x} \end{aligned}$$

where I is identity matrix

- Evaluated at true value θ_0

$$
\begin{aligned}
I &= \int \tilde{\theta}(\mathbf{x}) \left(\frac{\partial}{\partial \theta'} \log f(\mathbf{x}|\theta_0) \right) f(\mathbf{x}|\theta_0) d\mathbf{x} \\
&= \mathbb{E}\left[\tilde{\theta}(\mathbf{X}) \left(\frac{\partial}{\partial \theta'} \log f(\mathbf{X}|\theta_0) \right) \right] \\
&= \mathbb{E}\left[\tilde{\theta}(\mathbf{X}) \left(\frac{\partial}{\partial \theta'} \log f(\mathbf{X}|\theta_0) \right) \right] - \underbrace{\mathbb{E}\left[\tilde{\theta}(\mathbf{X}) \right]}_{\theta_0} \underbrace{\mathbb{E}\left[\frac{\partial}{\partial \theta'} \log f(\mathbf{X}|\theta_0) \right]}_{0} \\
&= \mathrm{cov}\left(\tilde{\theta}(\mathbf{X}), \frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta_0) \right)
\end{aligned}
$$

where the third equality follows from

$$
\mathbb{E}\left[\left(\frac{\partial}{\partial \theta'} \log f(\mathbf{X}|\theta_0) \right) \right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta'} \log f(X_i|\theta_0) \right) \right] = n\mathbb{E}[S'] = 0
$$

- Thus (showing $\mathrm{var}(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta_0)) = n\mathscr{F}_\theta$ left for homework)

$$\mathrm{var}\begin{pmatrix} \tilde{\theta} \\ \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta_0) \end{pmatrix} = \begin{pmatrix} \mathrm{var}(\tilde{\theta}) & I \\ I & n\mathscr{F}_{\theta_0} \end{pmatrix}$$

- Since this matrix is positive semidefinite

$$A'\mathrm{var}\begin{pmatrix} \tilde{\theta} \\ \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta_0) \end{pmatrix} A \geq 0$$

  for any matrix A

- Picking $A = \left\{ \begin{array}{c} I \\ -(n\mathscr{F}_{\theta_0})^{-1} \end{array} \right\}$ yields

$$\mathrm{var}(\tilde{\theta}) - (n\mathscr{F}_{\theta_0})^{-1} \geq 0$$

Asymptotic property of MLE

- If θ_0 uniquely maximizes $\ell(\theta) = \mathbb{E} \log f(X|\theta)$ and some technical conditions hold so that

$$\frac{1}{n} \sum_{i=1}^{n} \log f(X_i|\theta) \xrightarrow{p} \mathbb{E} \log f(X|\theta)$$

uniformly for all $\theta \in \Theta$, then

$$\hat{\theta} \xrightarrow{p} \theta_0$$

- With more technical conditions, we can also show

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathscr{F}_{\theta_0}^{-1})$$

- Thus MLE estimator is: consistent, converging at rate $n^{-\frac{1}{2}}$, asymptotically normal and **asymptotically** Cramér-Rao efficient

Variance estimation

- The asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is $\mathscr{F}_{\theta_0}^{-1}$, which is unknown
- Since

$$\mathscr{F}_\theta = \mathbb{E}\left[\frac{\partial \log f(X|\theta_0)}{\partial \theta}\frac{\partial \log f(X|\theta_0)}{\partial \theta'}\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(X|\theta_0)\right]$$

we can estimate $\mathscr{F}_\theta^{-1}$ by either

$$\left\{-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\log f(X_i|\hat{\theta})\right\}^{-1}$$

or

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\hat{\theta})\frac{\partial}{\partial\theta'}\log f(X_i|\hat{\theta})\right\}^{-1}$$

2. Method of Moments

Introduction

- MLE is used for **parametric** models
- Method of Moments (MM) allows **semi-parametric** models: estimation of finite dimensional parameter when distribution is **non-parametric**
- A distribution is called **non-parametric** if it cannot be described by a finite list of parameters
- Example: Estimation of the mean $\theta = \mathbb{E}[X]$ when the distribution of X is unspecified

Multivariate means

- To start with, for random vector X, its mean $\mu = \mathbb{E}X$ can be estimated by MME
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- By CLT, if $\mathbb{E}\left\|X\right\|^2 < \infty$
$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathrm{N}(0, \Sigma)$$
where $\Sigma = \mathrm{var}[X]$
- Σ can be consistently estimated by sample covariance matrix
$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})'$$

Mean of transformed variable

- The mean of any transformation $g(X)$ is $\theta = \mathbb{E}[g(X)]$
- MME for θ is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

- By CLT, if $\mathbb{E} \|g(X)\|^2 < \infty$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{N}(0, V_\theta)$$

where $V_\theta = \mathrm{var}[g(X)]$

- V_θ can be consistently estimated by

$$\hat{V} = \frac{1}{n-1} \sum_{i=1}^{n} (g(X_i) - \hat{\theta})(g(X_i) - \hat{\theta})'$$

Example: moments

- The $m-$th moment of random variable X is $\mu'_m = \mathbb{E}X^m$
- Similarly, MME for μ_m is

$$\hat{\mu}'_m = \frac{1}{n}\sum_{i=1}^{n} X_i^m$$

- CLT yields its asymptotic distribution

Example: empirical distribution function

- The cdf of X is

$$F(x) = P\{X \leq x\} = \mathbb{E}[\mathbf{1}\{X \leq x\}]$$

- The MME for $F(x)$ is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}$$

- $F_n(x)$ is called the empirical distribution function
- We can show (homework)

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} \mathrm{N}\left(0, F(x)(1 - F(x))\right)$$

Smooth functions of moments

- Now let's be a bit general
- Suppose the parameter is

$$\beta = h(\theta), \text{ where } \theta = \mathbb{E}[g(X)]$$

  and X, g and h can all be vectors

- By plugging in MME $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$, β can be estimated by

$$\hat{\beta} = h(\hat{\theta})$$

- When h is continuously differentiable we call it **smooth**
- By applying delta method

$$\hat{\beta} - \beta \xrightarrow{d} \mathrm{N}(0, V_\beta)$$

where $V_\beta = \mathbf{H}' V_\theta \mathbf{H}$, $\mathbf{H}' = \frac{\partial}{\partial \theta'} h(\theta)$, $V_\theta = \mathrm{var}(g(X))$

- V_β can be consistently estimated by $\hat{V}_\beta = \hat{\mathbf{H}}' \hat{V}_\theta \hat{\mathbf{H}}$ where

$$\hat{\mathbf{H}}' = \frac{\partial}{\partial \theta'} h(\hat{\theta})$$

$$\hat{V}_\theta = \frac{1}{n-1} \sum_{i=1}^{n} (g(X_i) - \hat{\theta})(g(X_i) - \hat{\theta})'$$

Example: variance

- The variance of random variable X is

$$\begin{aligned}\sigma^2 &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2\end{aligned}$$

  a smooth function of uncentered first and second moment

- MME for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

- The asymptotic distribution of $\hat{\sigma}^2$ can be found by delta method

Moment equations

- In many problems, we can write moments as explicit functions of parameters
$$\mathbb{E}[m(X, \beta)] = 0$$
where parameter $\beta \in \mathbb{R}^k$ and $m(x, \beta)$ is a $k \times 1$ function

- For each β, the sample moment of $\mathbb{E}[m(X, \beta)]$ is
$$\frac{1}{n} \sum_{i=1}^{n} m(X_i, \beta)$$

- The MME $\hat{\beta}$ solves a system of k nonlinear equations
$$\frac{1}{n} \sum_{i=1}^{n} m(X_i, \hat{\beta}) = 0$$

Example: parametric models

- Classical way of defining MME
- Let $f(x|\beta)$ be a parametric density with parameter $\beta \in \mathbb{R}^m$
- The $k-$th moment of the model is

$$\mu_k(\beta) = \int x^k f(x|\beta) dx$$

a mapping from parameter space to $\mathbb{R}$

- Hence β satisfy

$$\mathbb{E} \begin{bmatrix} X - \mu_1(\beta) \\ X^2 - \mu_2(\beta) \\ \vdots \\ X^m - \mu_m(\beta) \end{bmatrix} = 0,$$

- We can set

$$m(x,\beta) = \begin{pmatrix} x - \mu_1(\beta) \\ x^2 - \mu_2(\beta) \\ \vdots \\ x^m - \mu_m(\beta) \end{pmatrix}$$

- MME $\hat{\beta}$ solves

$$\frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} X_i - \mu_1(\hat{\beta}) \\ X_i^2 - \mu_2(\hat{\beta}) \\ \vdots \\ X_i^m - \mu_m(\hat{\beta}) \end{bmatrix} = 0$$

Example: Euler equation in macro

- Consumer's utility function

$$U(C_t, C_{t+1}) = u(C_t) + \frac{1}{\beta} u(C_{t+1})$$

- Consumer's budget

$$C_t + \frac{C_{t+1}}{R_{t+1}} \leq W_t$$

- Consumer chooses C_t to maximize expected utility

$$\mathbb{E}\left[u(C_t) + \frac{1}{\beta} u((W_t - C_t)R_{t+1})\right]$$

- FOC is

$$0 = u'(C_t) - \mathbb{E}\left[\frac{R_{t+1}}{\beta} u'(C_{t+1})\right]$$

- Assuming $u(c) = \frac{c^{1-\alpha}}{1-\alpha}$, the Euler equation is

$$\mathbb{E}\left[R_{t+1}\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha} - \beta\right] = 0$$

- Suppose β is known and we are interested in estimating α
- Then α satisfies $\mathbb{E}\left[m(R_{t+1}, C_{t+1}, C_t, \alpha)\right] = 0$, where

$$m(R_{t+1}, C_{t+1}, C_t, \alpha) = R_{t+1}\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha} - \beta$$

- The MME for α solves

$$\frac{1}{n}\sum_{t=1}^{n}\left[m(R_{t+1}, C_{t+1}, C_t, \hat{\alpha})\right] = 0$$

Asymptotic property of MME

- If there is a unique β_0 that solves

$$\mathbb{E}[m(X,\beta)] = 0$$

  and further technical conditions hold so that

$$\frac{1}{n}\sum_{i=1}^{n}[m(X_i,\beta)] \xrightarrow{p} \mathbb{E}[m(X,\beta)]$$

  uniformly for all β in some set B, then MME $\hat{\beta} \xrightarrow{p} \beta_0$

- With more technical conditions, we can also show

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} (0, V)$$

  where $V = (Q')^{-1}\Omega Q^{-1}$, $\Omega = \text{var}(m(X,\beta_0))$,
  $Q' = \mathbb{E}\left[\frac{\partial}{\partial\beta'}m(X,\beta_0)\right]$

Efficiency of MME Estimator

- We know sample mean $\hat{\mu}$ is BLUE for population mean μ, which might justify use of MME
- Restriction to linear models is not convincing
- In fact, we can show $\hat{\mu}$ has the lowest variance among **all unbiased estimators**
- **Theorem**: Let X be a random vector and $\mathcal{F}$ be a set of distributions such such that $\mathbb{E}\left\|X\right\|^2 < \infty$. If $\tilde{\mu}$ is an unbiased estimator for $\mu = \mathbb{E}X$ for all distributions in $\mathcal{F}$, then

  $$\text{var}(\tilde{\mu}) \geq \frac{1}{n}\Sigma$$

  where $\Sigma = \text{var}(X)$
- Since sample mean $\hat{\mu}$ is unbiased and $\text{var}(\hat{\mu}) = \frac{1}{n}\Sigma$, we conclude $\hat{\mu}$ has the lowest variance among all unbiased estimators

Proof (non-examinable)

- Basic Idea
  - If X has a parametric pdf $f(x|\theta)$, we can apply Cramér-Rao theory to find lower bound
  - However, the distribution of X is left unspecified (the space of possible distributions is too big)
  - Construct a smaller class of correctly specified parametric distributions $f(x|\alpha)$ so that when $\alpha = 0$, $f(x|0) = f(x)$
  - Since $\tilde{\mu}$ is unbiased for all distributions, it is also unbiased for $f(x|\alpha)$
  - The variance lower bound among all distributions must at least as large as the Cramér-Rao bound for the subclass of distributions $f(x|\alpha)$

- Focus on the case when X continuous with $f(x)$. Wlog, assume $\mu = 0$ and X is bounded so that $\|X\| \leq C$ for some $0 < C < \infty$
- Extending to cases with $\mu \neq 0$ and unbounded X only involves some more technicality
- Now let $\mathcal{F}$ be the set of distributions such that $\mathbb{E}X = 0$ and $\|X\| \leq C$ with probability 1
- Note $\|X\| \leq C$ with probability 1 implies $\mathbb{E}\,\|X\|^2 < \infty$ is automatically satisfied

- Step 1: construct a parametric subclass of distributions

$$f(x|\alpha) = f(x)\left\{1 + \alpha'\Sigma^{-1}x\right\}$$

where $\alpha \in \left\{\alpha : \left\|\Sigma^{-1}\alpha\right\| \le \frac{1}{C}\right\}$,

$$\Sigma = \mathrm{var}(X) = \mathbb{E}[XX']$$

Note $\mathbb{E}X = 0$, $|x| \le C$

- Let $\mathbb{E}_\alpha[\cdot]$ denote expectation under $f(x|\alpha)$
- Step 2: verify that $f(x|\alpha) \in \mathcal{F}$
  - $f(x|\alpha)$ is a valid pdf sharing same support with $f(x)$

$$f(x|\alpha) \geq 0 \text{ since } |\alpha'\Sigma^{-1}x| \leq \left\|\Sigma^{-1}\alpha\right\| \|x\| \leq 1 \tag{3}$$

$$\begin{aligned}\int f(x|\alpha)dx &= \int f(x)dx + \int f(x)\alpha'\Sigma^{-1}xdx \\ &= 1 + \alpha'\Sigma^{-1}\mathbb{E}X = 1\end{aligned}$$

  - $f(x|\alpha)$ is correctly specified: when $\alpha = \mathbf{0}$, $f(x|\alpha) = f(x)$
  - Variance of X under $f(x|\alpha)$ is finite:
    (3) implies $f(x|\alpha) \leq 2f(x)$. Thus $\mathbb{E}_\alpha \|X\|^2 \leq 2\mathbb{E}\|X\| < \infty$
  - Expectation of X under $f(x|\alpha)$ is

$$\begin{aligned}\int xf(x|\alpha)dx &= \int f(x)xdx + \left(\int xx'f(x)dx\right)\Sigma^{-1}\alpha \\ &= 0 + \Sigma^{-1}\Sigma^{-1}\alpha = \alpha\end{aligned}$$

- Step 3: apply Cramér-Rao Theorem for model $f(x|\alpha)$
  - Unbiasedness of $\tilde{\mu}$ means it is unbiased for all $f(x) \in \mathcal{F}$. Since $f(x|\alpha) \in \mathcal{F}$, it must hold that $\tilde{\mu}$ is unbiased for model $f(x|\alpha)$
  - By Cramér-Rao Theorem,

$$\text{var}(\tilde{\mu}) \geq n^{-1}\mathscr{F}_\alpha$$

    where

$$\mathscr{F}_\alpha = \mathbb{E}\left[\frac{\partial}{\partial\alpha}\log f(X|0)\frac{\partial}{\partial\alpha'}\log f(X|0)\right]$$

  - Note

$$\frac{\partial}{\partial\alpha}\log f(X|\alpha) = \frac{\Sigma^{-1}X}{\{1+\alpha'\Sigma^{-1}X\}}$$

  - Hence $\mathscr{F}_\alpha = \Sigma^{-1}\mathbb{E}\left[XX'\right]\Sigma^{-1} = \Sigma^{-1}$ as desired