ECE 275A: Parameter Estimation I
Method of Moments

Florian Meyer

Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Convergence in Probability

- A sequence of random vectors $\{\boldsymbol{X}_k\}$ with $\boldsymbol{X}_k \in \mathbb{R}^n$, converges to random vector $\boldsymbol{X} \in \mathbb{R}^n$ "in probability" if

$$\lim_{k\to\infty} P(\|\boldsymbol{X}_k - \boldsymbol{X}\| \geq \epsilon) = 0, \forall \epsilon > 0$$

- This is typically denoted either as $\boldsymbol{X}_k \overset{\text{prob.}}{\to} \boldsymbol{X}$ or as $\operatorname*{p-lim}_{k\to\infty} \boldsymbol{X}_k = \boldsymbol{X}$

Carry-Over Property of Convergence in Probability

- Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a continuous vector-valued function
- For any two vectors $\boldsymbol{x}_k$ and $\boldsymbol{x}$ and $\forall \delta > 0$, there exists $\epsilon > 0$ such that we obtain
$$\|\boldsymbol{x}_k - \boldsymbol{x}\| < \epsilon \Rightarrow \|f(\boldsymbol{x}_k) - f(\boldsymbol{x})\| < \delta$$
- Note that the contrapositive statement is
$$\|f(\boldsymbol{x}_k) - f(\boldsymbol{x})\| \geq \delta \Rightarrow \|\boldsymbol{x}_k - \boldsymbol{x}\| \geq \epsilon \tag{1}$$

Carry-Over Property of Convergence in Probability

- Let $\{\boldsymbol{X}_k\}$ be a sequence that converges in probability, i.e., $\underset{k\to\infty}{\text{p-lim}}\, \boldsymbol{X}_k = \boldsymbol{X}$

- Using equation (1), we have

$$\forall \delta > 0, \exists \epsilon > 0, \text{ such that } P(\|f(\boldsymbol{X}_k) - f(\boldsymbol{X})\| \geq \delta) \leq P(\|\boldsymbol{X}_k - \boldsymbol{X}\| \geq \epsilon)$$

- Hence, from $\underset{k\to\infty}{\text{p-lim}}\, \boldsymbol{X}_k = \boldsymbol{X}$ it follows that

$$\lim_{k\to\infty} P(\|f(\boldsymbol{X}_k) - f(\boldsymbol{X})\| \geq \delta) = 0, \forall \delta > 0,$$

i.e., $\underset{k\to\infty}{\text{p-lim}}\, f(\boldsymbol{X}_k) = f(\boldsymbol{X})$

- This carry-over property of convergence in probability can also be denoted as

$$\boldsymbol{X}_k \overset{\text{prob.}}{\rightarrow} \boldsymbol{X} \Rightarrow f(\boldsymbol{X}_k) \overset{\text{prob.}}{\rightarrow} f(\boldsymbol{X})$$

Convergence in Mean Square

- A sequence $\{\boldsymbol{X}_k\}$ converges to $\boldsymbol{X}$ "in mean square" if

$$\lim_{k\to\infty} E\big(\|\boldsymbol{X}_k - \boldsymbol{X}\|^2\big) = 0$$

- This is typically denoted as $\boldsymbol{X}_k \stackrel{\text{m.s.}}{\to} \boldsymbol{X}$

- By using the Markov inequality, we obtain

$$P(\|\boldsymbol{X}_k - \boldsymbol{X}\| \geq \epsilon) = P(\|\boldsymbol{X}_k - \boldsymbol{X}\|^2 \geq \epsilon^2) \leq \frac{E(\|\boldsymbol{X}_k - \boldsymbol{X}\|^2)}{\epsilon^2}$$

- We have $\lim_{k\to\infty} P(\|\boldsymbol{X}_k - \boldsymbol{X}\| \geq \epsilon) \leq \lim_{k\to\infty} \frac{E(\|\boldsymbol{X}_k - \boldsymbol{X}\|^2)}{\epsilon^2} = 0, \forall \epsilon > 0$ and thus shown that convergence in mean square implies convergence in probability, i.e.,

$$\boldsymbol{X}_k \stackrel{\text{m.s.}}{\to} \boldsymbol{X} \Rightarrow \boldsymbol{X}_k \stackrel{\text{prob.}}{\to} \boldsymbol{X}$$

Method of Moments

- Let $\boldsymbol{\theta} \in \mathbb{R}^p$ be a parameter vector that parameterizes a (statistical family) of model distributions $p_{\boldsymbol{\theta}}(x) = p(x; \boldsymbol{\theta})$
- Assume the statistical family is **identifiable**

$$p(x; \boldsymbol{\theta}) = p(x; \boldsymbol{\theta}'), \forall x \ \Leftrightarrow \ \boldsymbol{\theta} = \boldsymbol{\theta}'$$

  and **well-specified**

$$\exists\, \boldsymbol{\theta} \in \Theta \ \text{ such that } \ p_{\text{true}}(x) = p(x; \boldsymbol{\theta})$$

  where p_{true} is the true distribution

- If the family is identifiable and well-specified, then the truth distribution is uniquely represented by a "true parameter vector" $\boldsymbol{\theta}_{\text{true}}$:

$$\exists\, \boldsymbol{\theta}_{\text{true}} \ \text{ such that } \ p_{\text{true}}(x) = p(x; \boldsymbol{\theta}_{\text{true}})$$

- In this case, learning the true distribution is equivalent to learning the true parameter $\boldsymbol{\theta}_{\text{true}}$

Method of Moments

- Let $\mu_k(\boldsymbol{\theta}) = E(X^k; \boldsymbol{\theta})$ be the k-th non-central moment $(k = 1, 2, \cdots)$ of the model distribution $p(x; \boldsymbol{\theta})$

- Define $\boldsymbol{m}(\cdot)$: $\mathbb{R}^p \to \mathbb{R}^p$ as a vector-valued function, i.e.,

$$\boldsymbol{m}(\boldsymbol{\theta}) = \begin{bmatrix} \mu_{k_1}(\boldsymbol{\theta}) \\ \mu_{k_2}(\boldsymbol{\theta}) \\ \vdots \\ \mu_{k_p}(\boldsymbol{\theta}) \end{bmatrix}$$

  where the p elements are non-central moments $m_{k_i}(\boldsymbol{\theta})$ selected such that the $p \times p$ Jacobian matrix $\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{m}(\boldsymbol{\theta})$ is nonsingular for all $\boldsymbol{\theta}$

Method of Moments

- Let $\boldsymbol{m}_{\text{true}}$ be the vector of true moments corresponding to the vector of model moments $\boldsymbol{m}(\boldsymbol{\theta})$
- Under the nonsingular Jacobian matrix assumption, the vector equation
$$\boldsymbol{m}(\boldsymbol{\theta}) = \boldsymbol{m}_{\text{true}}$$
is a system of p independent scalar equations that can be solved for $\boldsymbol{\theta}_{\text{true}}$, i.e.,
$$\boldsymbol{\theta}_{\text{true}} = \boldsymbol{m}^{-1}(\boldsymbol{m}_{\text{true}})$$
where $\boldsymbol{m}^{-1}(\cdot)$ is the inverse function of $\boldsymbol{m}(\cdot)$ which is also continuously differentiable
- However, $\boldsymbol{m}_{\text{true}}$ is unknown

Method of Moments

- Assume that we observe N i.i.d. samples $\{x_1, \cdots, x_N\}$ drawn from the true distribution $p_{\text{true}}(x)$
- Let $\hat{\mu}_k^{(N)} = \frac{1}{N}\sum_{i=1}^{N} x_i^k$ be the k-th non-central sample moment
- Define $\hat{\boldsymbol{m}}^{(N)}$ be the vector whose elements are the sample moments of the elements (which are moments) of $\boldsymbol{m}(\cdot)$

$$\hat{\boldsymbol{m}}^{(N)} = \begin{bmatrix} \hat{\mu}_{k_1}^{(N)} \\ \hat{\mu}_{k_2}^{(N)} \\ \vdots \\ \hat{\mu}_{k_p}^{(N)} \end{bmatrix}$$

- Note that $E(\hat{\mu}_k^{(N)}) = \mu_k(\boldsymbol{\theta}_{\text{true}}) = \mu_{k*}$

Method of Moments

- Assume $\text{Var}(X^k) < \infty, \forall k \in \{k_1, k_2, \cdots, k_p\}$

$$\begin{aligned} E\big((\hat{\mu}_k^{(N)} - \mu_{k*})^2\big) &= \text{Var}\big(\hat{\mu}_k^{(N)}\big) \\ &= \text{Var}\Big(\frac{1}{N}\sum_{i=1}^{N} X_i^k\Big) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\text{Var}\big(X_i^k\big) \\ &= \frac{1}{N}\text{Var}\big(X^k\big) \end{aligned}$$

- Thus, we obtain

$$\lim_{N\to\infty} E\big((\hat{\mu}_k^{(N)} - \mu_{k*})^2\big) = 0 \;\Rightarrow\; \hat{\mu}_k^{(N)} \overset{\text{m.s.}}{\to} \mu_{k*} \;\Rightarrow\; \hat{\mu}_k^{(N)} \overset{\text{prob.}}{\to} \mu_{k*}$$

and $\underset{N\to\infty}{\text{p-lim}}\, \hat{\boldsymbol{m}}^{(N)} = \boldsymbol{m}_{\text{true}}$

Method of Moments

- Now define the **Method of Moments (MOM)** estimate

$$\hat{\boldsymbol{\theta}}^{(N)} = \boldsymbol{m}^{-1}(\hat{\boldsymbol{m}}^{(N)})$$

- The *Carry-Over Property of Convergence in Probability* yields consistency, i.e.,

$$\operatorname*{p-lim}_{N\to\infty} \hat{\boldsymbol{\theta}}^{(N)} = \operatorname*{p-lim}_{N\to\infty} \boldsymbol{m}^{-1}(\hat{\boldsymbol{m}}^{(N)}) = \boldsymbol{m}^{-1}(\boldsymbol{m}_{\text{true}}) = \boldsymbol{\theta}_{\text{true}},$$

- Thus, we have

$$\hat{\boldsymbol{\theta}}^{(N)} \overset{\text{prob.}}{\to} \boldsymbol{\theta}_{\text{true}}$$