

18.650
Statistics for Applications

Chapter 4: The Method of Moments

Weierstrass Approximation Theorem (WAT)

Theorem

Let f be a continuous function on the interval $[a, b]$, then, for any $\varepsilon > 0$, there exists $a_0, a_1, \dots, a_d \in \mathbb{R}$ such that

$$\max_{x \in [a, b]} \left| f(x) - \sum_{k=0}^d a_k x^k \right| < \varepsilon.$$

In word: “continuous functions can be arbitrarily well approximated by polynomials”

Statistical application of the WAT (1)

- ▶ Let X_1, \dots, X_n be an i.i.d. sample associated with a (identified) statistical model $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$. Write θ^* for the true parameter.
- ▶ Assume that for all θ , the distribution \mathbb{P}_θ has a density f_θ .
- ▶ If we find θ such that

$$\int h(x) f_{\theta^*}(x) dx = \int h(x) f_\theta(x) dx$$

for all (bounded continuous) functions h , then $\theta = \theta^*$.

- ▶ Replace expectations by averages: find estimator $\hat{\theta}$ such that

$$\frac{1}{n} \sum_{i=1}^n h(X_i) = \int h(x) f_{\hat{\theta}}(x) dx$$

for all (bounded continuous) functions h . There is an **infinity** of such functions: not doable!

Statistical application of the WAT (2)

- ▶ By the WAT, it is enough to consider polynomials:

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=0}^d a_k X_i^k = \int \sum_{k=0}^d a_k x^k f_{\hat{\theta}}(x) dx, \quad \forall a_0, \dots, a_d \in \mathbb{R}$$

Still an infinity of equations!

- ▶ In turn, enough to consider

$$\frac{1}{n} \sum_{i=1}^n X_i^k = \int x^k f_{\hat{\theta}}(x) dx, \quad \forall k = 1, \dots, d$$

(only $d + 1$ equations)

- ▶ The quantity $m_k(\theta) := \int x^k f_{\theta}(x) dx$ is the k th moment of \mathbb{P}_{θ} . Can also be written as

$$m_k(\theta) = \mathbb{E}_{\theta}[X^k].$$

Gaussian quadrature (1)

- ▶ The Weierstrass approximation theorem has limitations:
 1. works only for continuous functions (not really a problem!)
 2. works only on intervals $[a, b]$
 3. Does not tell us what d (# of moments) should be
- ▶ What if E is discrete: no PDF but PMF $p(\cdot)$?
- ▶ Assume that $E = \{x_1, x_2, \dots, x_r\}$ is finite with r possible values. The PMF has $r - 1$ parameters:

$$p(x_1), \dots, p(x_{r-1})$$

because the last one: $p(x_r) = 1 - \sum_{j=1}^{r-1} p(x_j)$ is given by the first $r - 1$.

- ▶ Hopefully, we do not need much more than $d = r - 1$ moments to recover the PMF $p(\cdot)$.

Gaussian quadrature (2)

- ▶ Note that for any $k = 1, \dots, r_1$,

$$m_k = \mathbb{E}[X^k] = \sum_{j=1}^r p(x_j) x_j^k$$

and

$$\sum_{j=1}^r p(x_j) = 1$$

This is a *system of linear equations* with unknowns $p(x_1), \dots, p(x_r)$.

- ▶ We can write it in a compact form:

$$\begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_r^1 \\ x_1^2 & x_2^2 & \cdots & x_r^2 \\ \vdots & & \ddots & \vdots \\ x_1^{r-1} & x_2^{r-1} & \cdots & x_r^{r-1} \\ 1 & 1 & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} p(x_1) \\ p(x_2) \\ \vdots \\ p(x_{r-1}) \\ p(x_r) \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_{r-1} \\ 1 \end{pmatrix}$$

Gaussian quadrature (2)

- ▶ Check if matrix is invertible: **Vandermonde determinant**

$$\det \begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_r^1 \\ x_1^2 & x_2^2 & \cdots & x_r^2 \\ \vdots & & \ddots & \vdots \\ x_1^{r-1} & x_2^{r-1} & \cdots & x_r^{r-1} \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \prod_{1 < j < k < r} (x_j - x_k) \neq 0$$

- ▶ So given m_1, \dots, m_{r-1} , there is a **unique** PMF that has these moments. It is given by

$$\begin{pmatrix} p(x_1) \\ p(x_2) \\ \vdots \\ p(x_{r-1}) \\ p(x_r) \end{pmatrix} = \begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_r^1 \\ x_1^2 & x_2^2 & \cdots & x_r^2 \\ \vdots & & \ddots & \vdots \\ x_1^{r-1} & x_2^{r-1} & \cdots & x_r^{r-1} \\ 1 & 1 & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_{r-1} \\ 1 \end{pmatrix}$$

Conclusion from WAT and Gaussian quadrature

- ▶ Moments contain important information to recover the PDF or the PMF
- ▶ If we can estimate these moments accurately, we may be able to recover the distribution
- ▶ In a parametric setting, where knowing the distribution \mathbb{P}_θ amounts to knowing θ , it is often the case that even less moments are needed to recover θ . This is on a case-by-case basis.
- ▶ Rule of thumb if $\theta \in \Theta \subset \mathbb{R}^d$, we need d moments.

Method of moments (1)

Let X_1, \dots, X_n be an i.i.d. sample associated with a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$. Assume that $\Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$.

► *Population moments*: Let $m_k(\theta) = \mathbb{E}_\theta[X_1^k]$, $1 \leq k \leq d$.

► *Empirical moments*: Let $\hat{m}_k = \overline{X_n^k} = \frac{1}{n} \sum_{i=1}^n X_i^k$, $1 \leq k \leq d$.

► Let

$$\begin{aligned} \psi &: \Theta \subset \mathbb{R}^d &\rightarrow &\mathbb{R}^d \\ &\theta &\mapsto &(m_1(\theta), \dots, m_d(\theta)). \end{aligned}$$

Method of moments (2)

Assume ψ is one to one:

$$\theta = \psi^{-1}(m_1(\theta), \dots, m_d(\theta)).$$

Definition

Moments estimator of θ :

$$\hat{\theta}_n^{MM} = \psi^{-1}(\hat{m}_1, \dots, \hat{m}_d),$$

provided it exists.

Method of moments (3)

Analysis of $\hat{\theta}_n^{MM}$

- ▶ Let $M(\theta) = (m_1(\theta), \dots, m_d(\theta))$;
- ▶ Let $\hat{M} = (\hat{m}_1, \dots, \hat{m}_d)$.
- ▶ Let $\Sigma(\theta) = \mathbb{V}_\theta(X, X^2, \dots, X^d)$ be the covariance matrix of the random vector (X, X^2, \dots, X^d) , where $X \sim \mathbb{P}_\theta$.
- ▶ Assume ψ^{-1} is continuously differentiable at $M(\theta)$. Write $\nabla\psi^{-1}|_{M(\theta)}$ for the $d \times d$ gradient matrix at this point.

Method of moments (4)

- ▶ LLN: $\hat{\theta}_n^{MM}$ is weakly/strongly consistent.
- ▶ CLT:

$$\sqrt{n} \left(\hat{M} - M(\theta) \right) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \Sigma(\theta)) \quad (\text{w.r.t. } \mathbb{P}_\theta).$$

Hence, by the Delta method (see next slide):

Theorem

$$\sqrt{n} \left(\hat{\theta}_n^{MM} - \theta \right) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \Gamma(\theta)) \quad (\text{w.r.t. } \mathbb{P}_\theta),$$

where $\Gamma(\theta) = [\nabla \psi^{-1}|_{M(\theta)}]^\top \Sigma(\theta) [\nabla \psi^{-1}|_{M(\theta)}]$.

Multivariate Delta method

Let $(T_n)_{n \geq 1}$ sequence of random vectors in \mathbb{R}^p ($p \geq 1$) that satisfies

$$\sqrt{n}(T_n - \theta) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \Sigma),$$

for some $\theta \in \mathbb{R}^p$ and some symmetric positive semidefinite matrix $\Sigma \in \mathbb{R}^{p \times p}$.

Let $g : \mathbb{R}^p \rightarrow \mathbb{R}^k$ ($k \geq 1$) be continuously differentiable at θ .
Then,

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, \nabla g(\theta)^\top \Sigma \nabla g(\theta)),$$

where $\nabla g(\theta) = \left(\frac{\partial g_j}{\partial \theta_i} \right)_{1 \leq i \leq p, 1 \leq j \leq k} \in \mathbb{R}^{k \times p}$.

MLE vs. Moment estimator

- ▶ Comparison of the quadratic risks: In general, the MLE is more accurate.
- ▶ Computational issues: Sometimes, the MLE is intractable.
- ▶ If likelihood is concave, we can use optimization algorithms (Interior point method, gradient descent, etc.)
- ▶ If likelihood is not concave: only heuristics. Local maxima. (Expectation-Maximization, etc.)

MIT OpenCourseWare
<https://ocw.mit.edu>

18.650 / 18.6501 Statistics for Applications
Fall 2016

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.