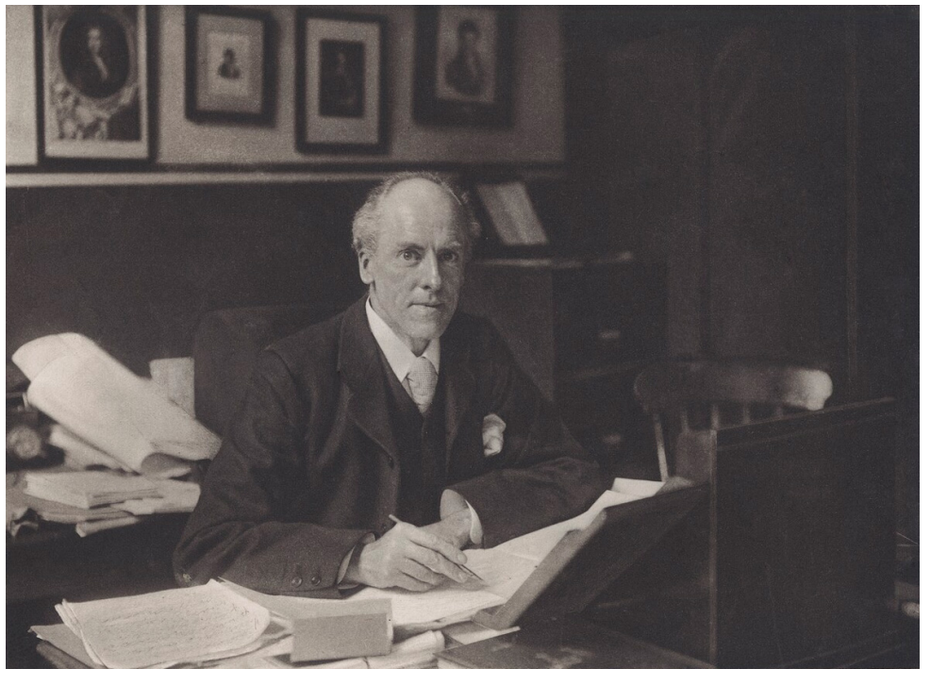
Method of Moments

STAT 37710 / CMSC 35300

Rebecca Willett and Yuxin Chen

In a nutshell



- History: Karl Pearson (1894)
- Method of moments: estimating population moment (or function of population moments) by using the corresponding sample moments (or function of sample moments)

Karl Pearson

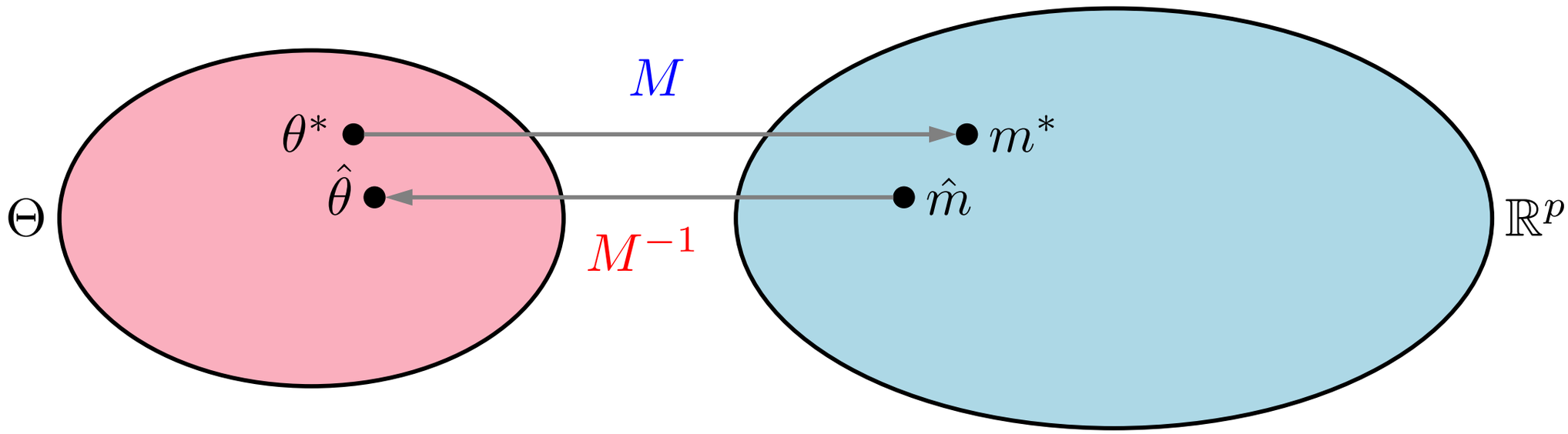
FRS

Pearson in 1912

Born	Carl Pearson 27 March 1857 Islington, London, England
Died	27 April 1936 (aged 79) Coldharbour, Surrey, England
Nationality	British
Alma mater	King's College, Cambridge University of Heidelberg
Known for	Principal component analysis Pearson distribution Pearson's chi-squared test Pearson's r Phi coefficient *The Grammar of Science*

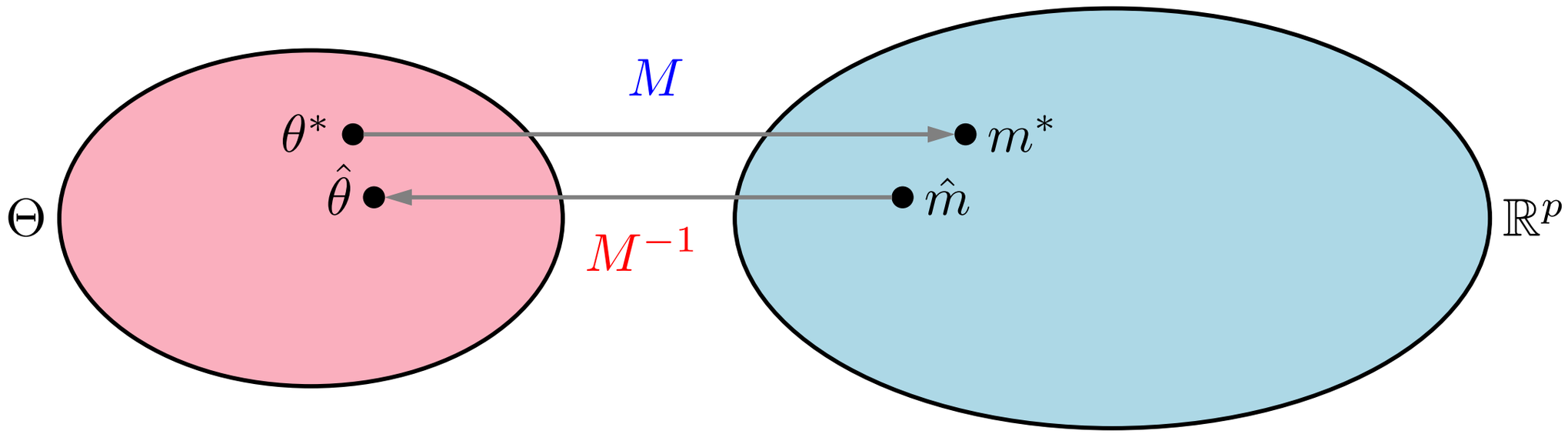
A general framework for the method of moment

1. define a *moment mapping* M relating the model parameters θ to moments m of the data distribution specified by θ.



Method of moment: A general framework

1. define a *moment mapping* M relating the model parameters θ to moments m of the data distribution specified by θ.
2. *plug in* the empirical moments $\hat{m}$ and invert the mapping to get parameter estimates θ.



Moment mapping

Let $\phi(z) \in \mathbb{R}^p$ be an observation function which only depends on the observed variables z.

Definition: Moment Mapping

$$M(\theta) := \mathbb{E}_{z \sim p_\theta}[\phi(z)]$$

Moment mapping

Let $\phi(z) \in \mathbb{R}^p$ be an observation function which only depends on the observed variables z.

Definition: Moment Mapping

$$M(\theta) := \mathbb{E}_{z \sim p_\theta}[\phi(z)]$$

Moment mapping maps each parameter vector $\theta \in \mathbb{R}^d$ to the expected value of $\phi(z)$ with respect to $p_\theta(z)$.
As an example, $\phi(z) = (z, z^2)$

Moment mapping

Let $\phi(z) \in \mathbb{R}^p$ be an observation function which only depends on the observed variables z.

Definition: Moment Mapping

$$M(\theta) := \mathbb{E}_{z \sim p_\theta}[\phi(z)]$$

Moment mapping maps each parameter vector $\theta \in \mathbb{R}^d$ to the expected value of $\phi(z)$ with respect to $p_\theta(z)$.
As an example, $\phi(z) = (z, z^2)$

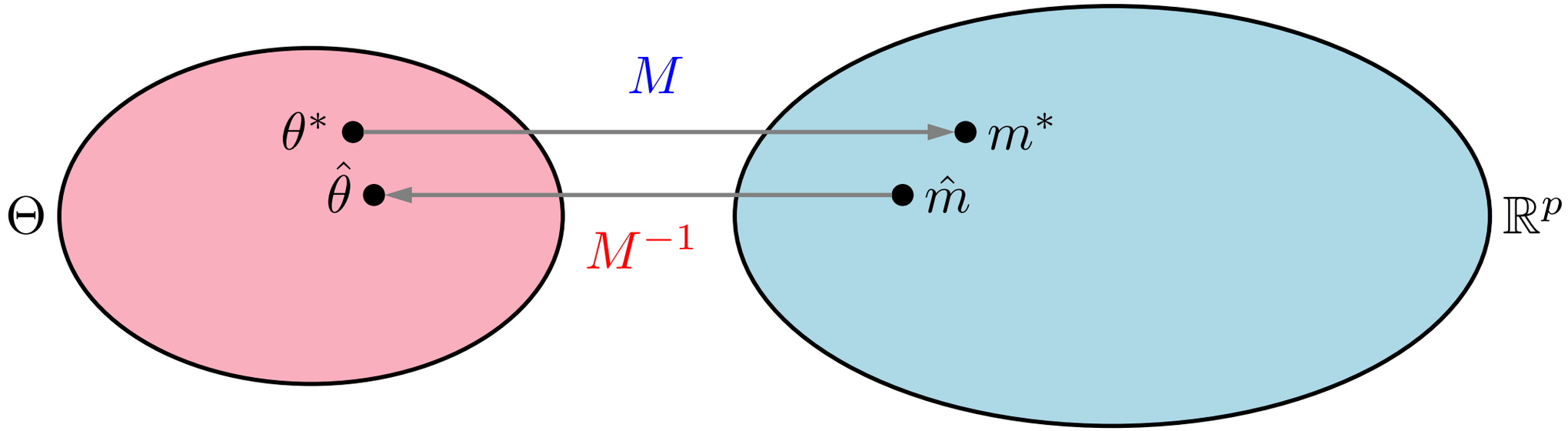
A link between moments (i.e. simple functions of the observations) with the parameters (i.e. quantities that we want to estimate).

Moment mapping

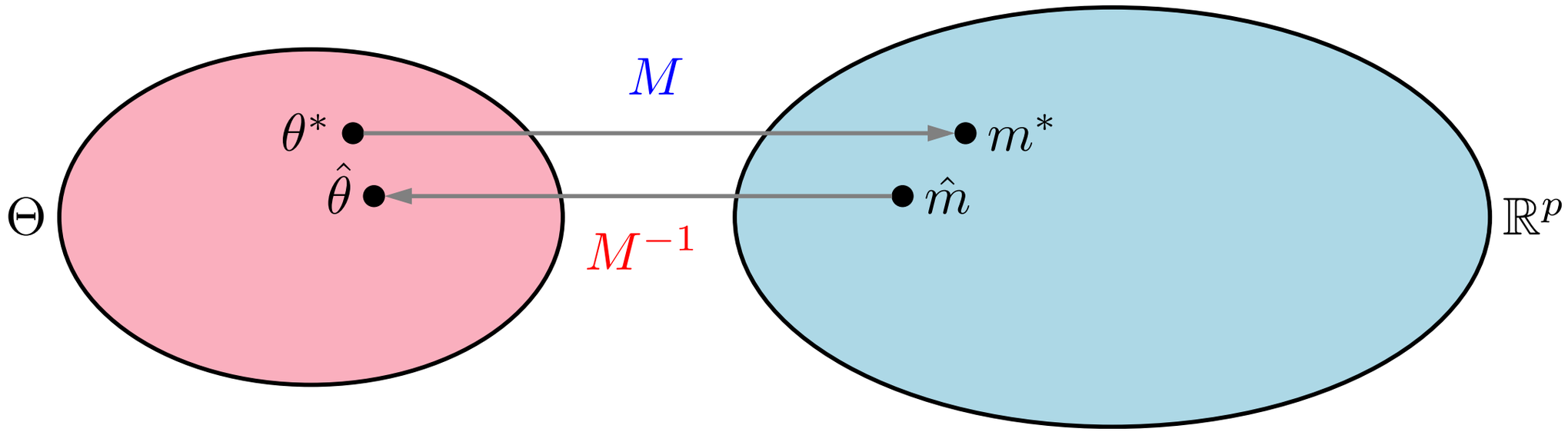
Example: Univariate Gaussion $\mathcal{N}(\mu, \sigma^2)$

Consider the observation function $\phi(z) = (z, z^2)$.
For M defined above and $\theta = (\mu, \sigma^2)$, the moment equations are:

$$M(\theta) = M((\mu, \sigma^2)) = \mathbb{E}_{z \sim \mathcal{N}(\mu, \sigma^2)}\left[(z, z^2)\right] = (\mu, \sigma^2 + \mu^2)$$



If someone told us some moments m (where $m^* = M(\theta^*)$).
Assuming M were invertible, we could solve for $\theta^* = M^{-1}(m^*)$.



Example: univariate Gaussion $\mathcal{N}(\mu, \sigma^2)$ cont.

Recall the moment equations

$$M(\theta) = M((\mu, \sigma^2)) = \mathbb{E}_{z \sim \mathcal{N}(\mu, \sigma^2)}\left[(z, z^2)\right] = (\mu, \sigma^2 + \mu^2)$$

If someone told us the moments $m^* = M(\theta^*) = (m_1^*, m_2^*)$, we can recover the parameters $\theta^* = M^{-1}(m^*)$ as follows:

- $\mu^* = m_1^*$
- $(\sigma^*)^2 = m_2^* - (m_1^*)^2$

Plug in

In practice, we don't have access to the true moments m^*, but we can estimate it extremely easily using the emprical moment (i.e. a sample average the over the data points):

$$\hat{m} := \frac{1}{n} \sum_{i=1}^{n} \phi(z^{(i)})$$

We can then plug in $\hat{m}$ for m^* to yield the method of moment estimator (MOME):

$$\widehat{\theta}_{\mathsf{MOME}} := M^{-1}(\hat{m})$$

Example: Possion(λ)

Consider the observation function $\phi(z) = z$. The moment equations are

$$M(\theta) = \mathbb{E}_{z \sim \text{Possion}(\lambda)}[z] = \lambda$$

which corresponds to the first moment. Thus, the method of moments estimator for $\phi(z) = z$ is $\widehat{\theta}_{\text{MOME}} = \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} z^{(i)}$.

For the data

```
> poissrnd(1, [1,10])
2 3 0 1 2 1 3 1 2 1
```

the method of moments estimator based on the first moment is 1.6

Example: Poisson(λ) (cont.)

We could also consider a different observation function $\phi(z) = z^2$ (which leads to the second moment). The moment equations are

$$M(\theta) = \mathbb{E}_{z \sim \text{Possion}(\lambda)}\left[z^2\right] = \lambda + \lambda^2$$

The method of moments estimator based on the second moment solves $\frac{1}{n}\sum_{i=1}^{n}\left(z^2\right)^{(i)} = \hat{\lambda} + \hat{\lambda}^2$ which gives

$$\widehat{\theta}_{\text{MOME}} = \hat{\lambda} = -\frac{1}{2} + \left[\frac{1}{4} + \frac{1}{n}\sum_{i=1}^{n}\left(z^2\right)^{(i)}\right]^{\frac{1}{2}}$$

The method of moments estimator for the above sequence based on the second moment is 1.4105

Asymptotic analysis

Conceptually, $\hat{m}$ is close m^*, and thus $\widehat{\theta}$ is close to θ^*.

To make it more concrete, we notice that $\hat{m}$ is an average of i.i.d. variables. Apply the central limit theorem

$$\sqrt{n}(\hat{m} - m^*) \to \mathcal{N}(0, \mathsf{Cov}_{z \sim p^*}[\phi(z)])$$

We want to analyze the asympototic behavior of $\theta = M^{-1}(m)$.

The delta method

Let $\hat{m}$ be a consistent estimator that converges in probability to its true value m^* with asymptotic normality:

$$\sqrt{n}(\hat{m} - m^*) \to \mathcal{N}(0, \Sigma)),$$

where n is the number of observations and Σ is a covariance matrix. Given a scalar-valued function h of $\hat{m}$, the delta method implies that

$$\sqrt{n}(h(\hat{m}) - h(m^*)) \to \mathcal{N}(0, \nabla h(m^*)^\top \Sigma \nabla h(m^*)))$$

Sketch proof: the delta method

Keeping only the first two terms of the Taylor series, and using vector notation for the gradient, we can estimate $h(\hat{m})$ as

$$h(\hat{m}) \approx h(m^*) + \nabla h(m^*)^T \cdot (\hat{m} - m^*)$$

Sketch proof: the delta method

Keeping only the first two terms of the Taylor series, and using vector notation for the gradient, we can estimate $h(\hat{m})$ as

$$h(\hat{m}) \approx h(m^*) + \nabla h(m^*)^T \cdot (\hat{m} - m^*)$$

which implies the variance of $h(\hat{m})$ is approximately

$$\begin{aligned}
\mathrm{Var}\left(h(\hat{m})\right) &\approx \mathrm{Var}\left(h(m^*) + \nabla h(m^*)^T \cdot (\hat{m} - m^*)\right) \\
&= \mathrm{Var}\left(h(m^*) + \nabla h(m^*)^T \cdot \hat{m} - \nabla h(m^*)^T \cdot m^*\right) \\
&= \mathrm{Var}\left(\nabla h(m^*)^T \cdot \hat{m}\right) \\
&= \nabla h(m^*)^T \cdot \mathrm{Cov}(\hat{m}) \cdot \nabla h(m^*) \\
&= \nabla h(m^*)^T \cdot (\Sigma) \cdot \nabla h(m^*)
\end{aligned}$$

Asymptotic analysis (cont.)

Conceptually, $\hat{m}$ is close m^*, and thus $\widehat{\theta}$ is close to θ^*.

Assuming that M^{-1} is continuous around m^*, by the Delta method we can argue that

$$\sqrt{n}(\widehat{\theta} - \theta^*) \to \mathcal{N}(0, \nabla M^{-1}(m^*) \mathsf{Cov}_{z \sim p^*}[\phi(z)] \nabla M^{-1}(m^*)^\top)$$

where $\nabla M^{-1}(m^*) \in \mathbb{R}^{d \times p}$ is the Jacobian matrix for M^{-1}.
Note that $\nabla M^{-1}(m^*) = \nabla M(\theta^*)^\dagger$ ($\dagger$ denotes pseudoinverse).
Therefore, for the method of moments to work well, the first d singular values of $\nabla M(\theta)$ should be far from 0. Intuitively, if we perturb θ by a little, we want m to move a lot.

When is MOME useful?

Method of moments is only useful if we can find an appropriate observation function ϕ such that

- The moment mapping M is invertible, and hopefully has singular values that are bounded below (i.e. M provides enough information about the parameters θ).
- The inverse moment mapping M^{-1} is computationally tractable.

Summary

- Fix observation function $\phi(z) \in \mathbb{R}^p$
- Define moment mapping $M(\theta) := \mathbb{E}_{z \sim p_\theta}[\phi(z)]$
- Estimate the empirical moments $\hat{m} := \frac{1}{n} \sum_{i=1}^{n} \phi(z^{(i)})$
- Plug in $\hat{m}$ for m^* to yield the method of moments estimator $\widehat{\theta}_{\mathsf{MOME}} := M^{-1}(\hat{m})$

MOME: a classical view

- Define $\mu_j = \mathbb{E}_{z \sim p_\theta}\left[z^j\right]$ for $j = 1, \ldots, k$. Consider

$$\phi(z) = (z - \mu_1, \ldots, z^k - \mu_k)$$

MOME: a classical view

- Define $\mu_j = \mathbb{E}_{z\sim p_\theta}\left[z^j\right]$ for $j = 1, \ldots, k$. Consider

$$\phi(z) = (z - \mu_1, \ldots, z^k - \mu_k)$$

- Apply moment mapping to get

$$\begin{aligned} M(\theta) &= \mathbb{E}_{z\sim p_\theta}[\phi(z)] \\ &= \mathbb{E}_{z\sim p_\theta}\left[(z - \mu_1, \ldots, z^k - \mu_k)\right] \\ &= 0 \end{aligned} \tag{1}$$

MOME: a classical view

- Define $\mu_j = \mathbb{E}_{z\sim p_\theta}\left[z^j\right]$ for $j = 1, \ldots, k$. Consider

$$\phi(z) = (z - \mu_1, \ldots, z^k - \mu_k)$$

- Apply moment mapping to get

$$\begin{aligned} M(\theta) &= \mathbb{E}_{z\sim p_\theta}[\phi(z)] \\ &= \mathbb{E}_{z\sim p_\theta}\left[(z - \mu_1, \ldots, z^k - \mu_k)\right] \\ &= 0 \end{aligned} \tag{1}$$

- The sample moment (or emprical moment) of $\phi(z)$ is

$$\begin{aligned} \hat{m} &= \frac{1}{n}\sum_{i=1}^{n}\phi(z^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}(z - \mu_1, \ldots, z^k - \mu_k)^{(i)} \\ &= \frac{1}{n}\sum_{i=1}^{n}(z, \ldots, z^k)^{(i)} - (\mu_1, \ldots, \mu_k) \end{aligned} \tag{2}$$

MOME: a classical view (cont.)

Equate the RHS of (1) with the RHS of (2), we get

$$\frac{1}{n}\sum_{i=1}^{n}(z,\ldots,z^k)^{(i)} - (\mu_1,\ldots,\mu_k) = 0$$

MOME: a classical view (cont.)

Equate the RHS of (1) with the RHS of (2), we get

$$\frac{1}{n}\sum_{i=1}^{n}(z,\ldots,z^k)^{(i)} - (\mu_1,\ldots,\mu_k) = 0$$

Hence it reduces to the classical view of moment-based estimator:

$$\frac{1}{n}\sum_{i=1}^{n} z^{(i)} \qquad \mu_1 = \mathbb{E}_{z\sim p_\theta}[z]$$

$$\vdots \qquad\qquad \vdots$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(z^k\right)^{(i)} \qquad \mu_k = \mathbb{E}_{z\sim p_\theta}\left[z^k\right]$$

MOME: a classical view (cont.)

Equate the RHS of (1) with the RHS of (2), we get

$$\frac{1}{n}\sum_{i=1}^{n}(z,\ldots,z^k)^{(i)} - (\mu_1,\ldots,\mu_k) = 0$$

Hence it reduces to the classical view of moment-based estimator:

sample moments of z	$\approx$	population moments of z
$\frac{1}{n}\sum_{i=1}^{n} z^{(i)}$		$\mu_1 = \mathbb{E}_{z\sim p_\theta}[z]$
$\vdots$		$\vdots$
$\frac{1}{n}\sum_{i=1}^{n}\left(z^k\right)^{(i)}$		$\mu_k = \mathbb{E}_{z\sim p_\theta}\left[z^k\right]$

Connection to other point estimators

- Now let us consider a different observation function

$$\phi(z) = \frac{\partial}{\partial \theta} \log p_\theta(z)$$

Connection to other point estimators

- Now let us consider a different observation function

$$\phi(z) = \frac{\partial}{\partial \theta} \log p_\theta(z)$$

- Apply moment mapping to get

$$\begin{aligned}
M(\theta) &= \mathbb{E}_{z\sim p_\theta}[\phi(z)] \\
&= \mathbb{E}_{z\sim p_\theta}\left[\frac{\partial}{\partial \theta} \log p_\theta(z)\right] \\
&= \int \frac{\partial \log p(z|\theta)}{\partial \theta} p(z|\theta) dz \\
&= \int \frac{1}{p(z|\theta)} \frac{\partial p(z|\theta)}{\partial \theta}\, p(z|\theta) dz = \int \frac{\partial p(z|\theta)}{\partial \theta} dz \\
&= \frac{\partial}{\partial \theta}\left[\int p(z|\theta) dz\right] = 0 \qquad (3)
\end{aligned}$$

Connection to other point estimators (cont.)

- We assume the order of integral and differential can be changed, this is justified under some reasonable conditions. We also have

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n}\phi(z^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p_\theta(z) \tag{4}$$

Connection to other point estimators (cont.)

- We assume the order of integral and differential can be changed, this is justified under some reasonable conditions. We also have

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n}\phi(z^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p_\theta(z) \tag{4}$$

- Equate the RHS of (3) with the RHS of (4), we get

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p_\theta(z) = 0$$

Connection to other point estimators (cont.)

- We assume the order of integral and differential can be changed, this is justified under some reasonable conditions. We also have

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n}\phi(z^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p_\theta(z) \tag{4}$$

- Equate the RHS of (3) with the RHS of (4), we get

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p_\theta(z) = 0$$

This is exactly the maximal likelihood estimator[a]!

[a] (cf MLE lecture notes for details)

Remarks

- The method of moments dates back to Pearson from 1894, which predates the full development of maximum likelihood by Fisher in the 1920s. Since then, maximum likelihood has been the dominant paradigm for parameter estimation, mainly due to its statistical efficiency and naturalness (see next lecture).

Remarks

- The method of moments dates back to Pearson from 1894, which predates the full development of maximum likelihood by Fisher in the 1920s. Since then, maximum likelihood has been the dominant paradigm for parameter estimation, mainly due to its statistical efficiency and naturalness (see next lecture).
- There is more ad-hocness in the method of moments, which allows us to get computationally efficient algorithms at the price of reduced statistical efficiency.

Remarks

- The method of moments dates back to Pearson from 1894, which predates the full development of maximum likelihood by Fisher in the 1920s. Since then, maximum likelihood has been the dominant paradigm for parameter estimation, mainly due to its statistical efficiency and naturalness (see next lecture).
- There is more ad-hocness in the method of moments, which allows us to get computationally efficient algorithms at the price of reduced statistical efficiency.
- The method of moment has been adopted for parameter estimation in latent-variable models, which include examples such as Gaussian mixture models (GMMs), Hidden Markov Models (HMMs), etc. We will explore some of these models in latter part of this course.