LECTURE 4 NOTES

1. MoM estimators. More generally, the MoM finds the root of a system of *moment equations* that vanish in expectation. That is,

1. construct $g_{\mathbf{x}} : \Theta \to \mathbf{R}^m$ such that $\mathbf{E}_{\theta^*}[g_{\mathbf{x}}(\theta^*)] = 0$
2. find $\hat{\theta}_{\text{MoM}} \in \Theta$ such that $g_{\mathbf{x}}(\hat{\theta}_{\text{MoM}}) = 0$.

In their most simple form, the moment equations are

$$g_{\mathbf{x}}(\theta) = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^1 - \mathbf{E}_\theta\big[\mathbf{x}_1\big] \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^m - \mathbf{E}_\theta\big[\mathbf{x}_1^m\big] \end{bmatrix}.$$

Another instance of the (more general) MoM is maximum likelihood. By the usual zero gradient optimality conditions, the MLE is a root of $\nabla \ell_x$. Thus it is the MoM estimator based on the moment equation

$$g_{\mathbf{x}}(\theta) = \nabla \ell_{\mathbf{x}}(\theta).$$

EXAMPLE 1.1. *Consider estimating the natural parameters of an exponential family:*

$$f_\theta(x) = \exp(\theta^T \phi(x) - a(\theta))h(x).$$

The MLE of θ is

$$\hat{\theta}_{\text{ML}} \in \arg\max_{\theta \in \Theta} \theta^T \phi(\mathbf{x}) - a(\theta),$$

where $a(\theta) := \log \int_{\mathcal{X}} \exp(\theta^T \phi(x))h(x)dx$. By the optimality of $\hat{\theta}_{\text{ML}}$,

$$0 = \phi(\mathbf{x}) - \nabla a(\hat{\theta}_{\text{ML}}).$$

It is possible to show that $\nabla a(\theta) = \mathbf{E}_\theta\big[\phi(\mathbf{x})\big]$ for any $\theta \in \Theta$. Thus an obvious set of moment equations is

$$g_{\mathbf{x}}(\theta) = \phi(\mathbf{x}) - \nabla a(\theta).$$

Its root is exactly the MLE! Thus, when the model is an exponential family, the obvious MoM estimator is the MLE.

To ensure the uniqueness of a MoM estimator, the moment equations must be invertible. When they are not, it is often possible to add equations (e.g. matching higher order moments) to ensure the moment equations are

invertible. However, care must be taken not to add too many equations or the moment equations may become inconsistent; i.e. there is no $\theta \in \Theta$ such that $g_x(\theta) = 0$.

EXAMPLE 1.2. *Consider fitting a mixture model:*

$$\begin{aligned} \mathbf{h}_i &\overset{\text{i.i.d.}}{\sim} \text{cat}(p), \\ \mathbf{x}_{i,j} \mid \mathbf{h}_i &\overset{\text{i.i.d.}}{\sim} \text{multi}(1, \mu_{\mathbf{h}_i}), \end{aligned} \tag{1.1}$$

where $p \in \Delta^k$ is the mixture weights and $\mu_h \in \Delta^d$ are the means of the mixture components.[1] *The model (1.1) is a popular model for text documents:*

1. *$\mathbf{h}_i \in [k]$ corresponds to the topic (e.g. arts, science, technology etc.) of the i-th document.*
2. *$\mathbf{x}_{i,j} \in \{0,1\}^d$ indicates the j-th word in the i-th document: there are d possible words in the dictionary, and $\mathbf{x}_{i,j}$ is an indicator vector.*

The MLE is intractable because the log-likelihood function is non-convex. However, as long as

1. *$\{\mu_h\}_{h\in[k]}$ are linearly independent,*
2. *we observe at least 3 realization of $\mathbf{x}_{i,j}$ for each realization of $\mathbf{h}_i$,*

it is possible to estimate the parameters by a MoM approach.

The first moment of $\mathbf{x}_{i,j}$ is

$$\mathbf{E}[\mathbf{x}_{i,j}] = \sum_{h\in[k]} \mathbf{E}[\mathbf{x}_{i,j} \mid \mathbf{h}_i = h] p_k = \sum_{h\in[k]} \mu_h p_h.$$

Clearly the first moment is not enough to identify the parameters. We are led to consider a second moment:

$$\begin{aligned} \mathbf{E}\left[\mathbf{x}_{i,j_1}\mathbf{x}_{i,j_2}^T\right] &= \sum_{h\in[k]} \mathbf{E}\left[\mathbf{x}_{i,j_1}\mathbf{x}_{i,j_2}^T \mid \mathbf{h}_i = h\right] p_k \\ &= \sum_{h\in[k]} \mathbf{E}\left[\mathbf{x}_{i,j_1} \mid \mathbf{h}_i = h\right] \mathbf{E}\left[\mathbf{x}_{i,j_2} \mid \mathbf{h}_i = h\right]^T p_k \\ &= \sum_{h\in[k]} \mu_h p_h \mu_h^T, \end{aligned} \tag{1.2}$$

[1] Δ^k is the probability simplex in $\mathbf{R}^k$: $\Delta^k := \{p \in (0,1)^k : \sum_{i\in[k]} p_i = 1\}$

From the second moment, it is possible to identify $\operatorname{span}(\{\mu_h\}_{h\in[k]})$*, but not the means. We are led to consider (a "slice" of) a third moment:*

$$\begin{aligned}\mathbf{E}\left[\mathbf{x}_{i,j_1}(a^T\mathbf{x}_{i,j_2})\mathbf{x}_{i,j_3}^T\right] &= \sum_{h\in[k]} \mathbf{E}\left[\mathbf{x}_{i,j_1}(a^T\mathbf{x}_{i,j_2})\mathbf{x}_{i,j_3}^T \mid \mathbf{h}_i = h\right] p_k \\ &= \sum_{h\in[k]} \mu_h(a^T\mu_h)p_h\mu_h^T\end{aligned} \tag{1.3}$$

As it turns out, for a generic $a \in \mathbf{R}^d$*, the third moment, together with the first and second moments, is enough to identify the parameters.*

Let $M := \begin{bmatrix}\mu_1 & \dots & \mu_k\end{bmatrix} \in \mathbf{R}^{d\times k}$*. The moment equations are*

$$g_{\mathbf{x}}(M,p) = \begin{bmatrix}\hat{u}_1 - Mp \\ \widehat{U}_2 - M\,\mathbf{diag}(p)M^T \\ \widehat{U}_3 - M\,\mathbf{diag}(M^Ta)\,\mathbf{diag}(p)M^T\end{bmatrix},$$

where $\hat{u}_1 \in \mathbf{R}^d$*,* $\widehat{U}_2, \widehat{U}_3 \in \mathbf{R}^{d\times d}$ *are estimates of the respective moments. Let* $\widehat{Q} \in \mathbf{R}^{d\times k}$ *be a subunitary matrix that spans* $\mathcal{R}(\widehat{U}_2)$*. It is possible to show that (as long as* $\{a^T\mu_h\}_{h\in[k]}$ *are distinct) the root of* g_x *is*

$$\widehat{M}_{MoM} = \widehat{Q}\widehat{V},\ \hat{p}_{\mathrm{MoM}} = \widehat{M}_{MoM}^{\dagger}\hat{u}_1$$

where $\widehat{V}\widehat{D}\widehat{V}^{-1}$ *is the eigen-decomposition of* $\widehat{Q}^T\widehat{U}_2\widehat{Q}(\widehat{Q}^T\widehat{U}_3\widehat{Q})^{-1}$*. To ensure* $\{a^T\mu_h\}_{h\in[k]}$ *are distinct, it suffices to choose* a *randomly.*

A potential advantage of MoM estimators over the MLE is they tend to be (more) robust to model misspecification. In Example 1.2, the normality of $\mathbf{x}_{i,j} \mid \mathbf{h}_i$, while crucial to the MLE (the likelihood function is totally dependent on the density of $\mathbf{x}_{i,j} \mid \mathbf{h}_i$), is inconsequential to the MoM estimator. If, unbeknown to the investigator, $\mathbf{x}_{i,j} \mid \mathbf{h}_i$ is non-normal, the MLE will perform poorly, but the MoM will return a reasonable estimator.

EXAMPLE 1.3. *Consider fitting a linear model to observation pairs*

$$(\mathbf{x}_i, \mathbf{y}_i) \overset{\text{i.i.d.}}{\sim} P.$$

The target is the best linear predictor

$$\begin{aligned}\beta^* &\in \operatorname*{arg\,min}_{\beta\in\mathbf{R}^p} \mathbf{E}_P\left[\tfrac{1}{2}(\mathbf{y}_1 - \mathbf{x}_1^T\beta)_2^2\right] \\ &= \operatorname*{arg\,min}_{\beta\in\mathbf{R}^p} \tfrac{1}{2}\beta^T\,\mathbf{E}_P\left[\mathbf{x}_1\mathbf{x}_1^T\right]\beta - \mathbf{E}_P\left[\mathbf{x}_1\mathbf{y}_1\right]^T\beta + \tfrac{1}{2}\,\mathbf{E}_P\left[\mathbf{y}_1^2\right].\end{aligned}$$

Given any distribution P on $\mathcal{X} \times \mathcal{Y}$ (with finite second moments), the objective function is well-defined. As long as $\mathbf{E}_P\left[\mathbf{x}_1\mathbf{x}_1^T\right] \succ 0$, the objective function is strongly convex. Thus $\beta^* \in \mathbf{R}^p$ is identifiable. By the optimality of β^*,

$$\mathbf{E}_P\left[\mathbf{x}_1\mathbf{x}_1^T\right]\beta^* - \mathbf{E}_P\left[\mathbf{x}_1\mathbf{y}_1\right] = 0,$$

which suggests the moment equations

$$g_{(\mathbf{X},\mathbf{y})}(\beta) := \left(\tfrac{1}{n}\textstyle\sum_{i\in[n]} \mathbf{x}_i\mathbf{x}_i^T\right)\beta - \tfrac{1}{n}\textstyle\sum_{i\in[n]} \mathbf{x}_i\mathbf{y}_i. \tag{1.4}$$

To estimate β^*, we solve

$$g_{(\mathbf{X},\mathbf{y})}(\beta) = \left(\tfrac{1}{n}\textstyle\sum_{i\in[n]} \mathbf{x}_i\mathbf{x}_i^T\right)\beta - \tfrac{1}{n}\textstyle\sum_{i\in[n]} \mathbf{x}_i\mathbf{y}_i = 0.$$

The solution is

$$\hat{\beta} = \left(\tfrac{1}{n}\textstyle\sum_{i\in[n]} \mathbf{x}_i\mathbf{x}_i^T\right)^{-1} \tfrac{1}{n}\textstyle\sum_{i\in[n]} \mathbf{x}_i\mathbf{y}_i,$$

which is the ordinary least squares (OLS) estimator.

We remark that the moment equations (1.4) arose from the optimality of β^*, not from a parametric assumption on P. In fact, the only assumptions thus far are

1. *P has finite second moments: $\mathbf{E}_P\left[\mathbf{x}_1\mathbf{x}_1^T\right]$, $\mathbf{E}_P\left[\mathbf{x}_1\mathbf{y}_1\right]$, $\mathbf{E}_P\left[\mathbf{y}_1^2\right] < \infty$.*
2. *$\mathbf{E}_P\left[\mathbf{x}_1\mathbf{x}_1^T\right] \succ 0$.*

Thus OLS, as a MoM estimator of the best linear predictor, is justified under very weak assumptions, which partly explains its popularity.

The idea behind the MoM is also a technique for obtaining approximations to distributions. The technique, often called *moment matching* in this context, gives an approximate distribution that has the same (usually first few) moments as the original distribution.

EXAMPLE 1.4. *When n is large, evaluating the binomial density or CDF exactly is tedious. A common trick is to approximate the $\mathrm{bin}(n,p)$ distribution by the $\mathcal{N}(np, np(1-p))$ or $\mathrm{Poi}(np)$ distributions. We observe that the parameters of the normal and Poisson approximations are chosen to match the first two moments of the Binomial distribution being approximated.*

The MoM has a few drawbacks. The main drawback is the general hardness of root finding. In practice, MoM estimators are usually evaluated numerically. However, as we just saw, there are settings in which the MLE is

hard to evaluate (because the likelihood function is hard to optimize), but the MoM leads to a tractable estimator.

Another drawback is MoM estimators are often *inefficient* compared to the MLE. That is, the asymptotic variance of MoM estimators are often larger that of the MLE. The inefficiency of the MoM is the flip-side of its robustness to model misspecification. The MoM is robust to model misspecification because it does not incorporate unneccesary distributional information. The MLE, by incorporating the additional information, is more efficiency when the extra information is valid.

2. Generalized method of moments (GMM). The drawbacks of the MoM include

1. it is inefficient compared to the MLE
2. it is not clear how to proceed when the moment equations are inconsistent. We could, discard some of the equations to make the moment equations consistent, but that seems profligate.

The generalized method of moments (GMM) addresses the preceding drawbacks by solving a least-squares problem instead of root-finding:

$$\hat{\theta}_{\text{GMM}} \in \arg\min_{\theta\in\Theta} \frac{1}{2}\|g_{\mathbf{x}}(\theta)\|^2_{\widehat{W}_n} := \frac{1}{2} g_{\mathbf{x}}(\theta)^T \widehat{W}_n g_{\mathbf{x}}(\theta)$$

for some (possibly data dependent) matrix of weights $\widehat{W}_n \succ 0$. As we shall see, it is possible to boost the efficiency of the GMM estimator by carefully choosing $\widehat{W}_n$.

Yuekai Sun
Berkeley, California
December 6, 2015