

# DATA DRIVEN TECHNIQUES

---


## Lecture 10 Data Driven Parameter Estimation

---

---

---

---



# Outline

- Data driven parameter estimation
- Data driven Bayesian estimation
  - Unknown joint density
  - Unknown parameter prior density
- Data driven non-Bayesian estimation
  - A class of parameter estimation problems
  - Density matching solution
- Examples

# Data driven parameter estimation

Given a vector  $X$  of measurements  
interested in estimating parameter vector  $\theta$

Estimator  $\hat{\theta}(X)$  : **any deterministic function of  $X$**

Statistical estimation assumes availability of  $f(X|\theta)$

In Bayesian approaches also availability of parameter prior  $h(\theta)$

Provides optimum or asymptotically optimum estimators

Goal: **Replace probability densities with data** sampled from them

Produce parameter estimates based solely on data

# Data driven Bayesian estimation

For conditional  $f(X|\theta)$ , prior  $h(\theta)$ , and cost  $C(\hat{\theta}, \theta)$ , define Average Cost

$$\mathcal{C}(\hat{\theta}) = \mathbb{E}_{X, \theta} [C(\hat{\theta}(X), \theta)] = \iint C(\hat{\theta}(X), \theta) f(X|\theta) h(\theta) dX d\theta$$

$$\min_{\hat{\theta}(X)} \mathcal{C}(\hat{\theta})$$

$$\hat{\theta}_o(X) = \arg \min_{\hat{\theta}(X)} \mathcal{C}(\hat{\theta}) = \arg \min_{\hat{\theta}(X)} \mathbb{E}_{X, \theta} [C(\hat{\theta}(X), \theta)]$$

$$\hat{\theta}_o(X) = \arg \min_U \mathbb{E}_{\theta} [C(U, \theta) | X] = \arg \min_U \mathbb{E}_{\theta} [C(U, \theta) f(X|\theta)]$$

$$G(u, x) = \int C(u, \theta) f(\theta/x) d\theta$$

$$= \int C(u, \theta) \frac{f(x|\theta) h(\theta)}{g(x)} d\theta \quad \rightarrow$$

$$\hat{\theta}_0(x) = \arg \max_u \frac{\int C(u, \theta) f(x|\theta) h(\theta) d\theta}{\int f(x|\theta) h(\theta) d\theta}$$

$$= \arg \max_u \int C(u, \theta) f(x|\theta) h(\theta) d\theta$$

$$= \arg \max_u \mathbb{E}_\theta [C(u, \theta) f(x|\theta)]$$

Unknown  $f(X|\theta)$ ,  $h(\theta)$      $\mathcal{C}(u, \theta)$  cost

Combine  $f(X|\theta)$ ,  $h(\theta)$  into equivalent  $f(X, \theta) = f(X|\theta)h(\theta)$

Data driven: Replace  $f(X, \theta)$  with data

$f(X, \theta) : (X_1, \theta_1), \dots, (X_n, \theta_n)$  sampled **simultaneously** from  $f(X, \theta)$

Replace (approximate)  $\hat{\theta}(X)$  with  $u(X, \alpha)$ ,  $\alpha$ : network parameters

$$\begin{aligned} \mathcal{C}(\alpha) = \mathbb{E}_{X, \theta} [\mathcal{C}(u(X, \alpha), \theta)] &\Rightarrow \min_{\alpha} \mathbb{E}_{X, \theta} [\mathcal{C}(u(X, \alpha), \theta)] \Rightarrow \alpha_o \\ u(X, \alpha_o) &\approx \hat{\theta}_o(X) \end{aligned}$$

Stochastic Gradient Descent

$$\begin{aligned} \alpha_t &= \alpha_{t-1} - \mu \nabla_{\alpha} \mathcal{C}(u(X_t, \alpha_{t-1}), \theta_t) \\ &= \alpha_{t-1} - \mu (\mathbb{J}_{\alpha} u(X_t, \alpha_{t-1}))^{\top} \nabla_U \mathcal{C}(u(X_t, \alpha_{t-1}), \theta_t) \end{aligned}$$

$$\hat{\mathcal{C}}(\alpha) = \frac{1}{n} \sum_{i=1}^n \mathbf{C}(\mathbf{u}(X_i, \alpha), \theta_i) \Rightarrow \min_{\alpha} \hat{\mathcal{C}}(\alpha)$$

Gradient Descent

$$\alpha_t = \alpha_{t-1} - \frac{\mu}{n} \sum_{i=1}^n (\mathbb{J}_{\alpha} \mathbf{u}(X_i, \alpha_{t-1}))^{\top} \nabla_U \mathbf{C}(\mathbf{u}(X_i, \alpha_{t-1}), \theta_i)$$

Known  $\mathbf{f}(X|\theta)$ , unknown  $\mathbf{h}(\theta)$

$$\mathbf{h}(\theta) : \theta_1, \dots, \theta_n$$

$$\hat{\theta}_o(X) = \arg \min_U \mathbb{E}_{\theta} [\mathbf{C}(U, \theta) | X] = \arg \min_U \mathbb{E}_{\theta} [\mathbf{C}(U, \theta) \mathbf{f}(X|\theta)]$$

For given  $X$

$$U_t = U_{t-1} - \mu \nabla_U \mathbf{C}(U_{t-1}, \theta_t) \mathbf{f}(X|\theta_t) \quad \text{SGD}$$

$$U_t = U_{t-1} - \frac{\mu}{n} \sum_{i=1}^n \nabla_U \mathbf{C}(U_{t-1}, \theta_i) \mathbf{f}(X|\theta_i) \quad \text{GD}$$

## Examples

$$\min_{\theta} \sum_{i=1}^n \|U - \theta_i\|^2 f(X|\theta_i) \rightarrow \sum_{i=1}^n (U - \theta_i) f(X|\theta_i) = 0 \rightarrow$$

$$C(U, \theta) = \|U - \theta\|^2 \quad \text{MMSE}$$

$$\hat{\theta}_{\text{MMSE}}(X) = \frac{\sum_{i=1}^n \theta_i f(X|\theta_i)}{\sum_{i=1}^n f(X|\theta_i)}$$

$$C(U, \theta) = \|U - \theta\|_{L_1} \quad \text{MMAE}$$

$$\theta_{[1]} \leq \theta_{[2]} \leq \dots \leq \theta_{[n]}$$

$$\hat{\theta}_{\text{MMAE}}(X) = \theta_{[i_o]}$$

$$i_o = \arg \begin{cases} f(X|\theta_{[1]}) + \dots + f(X|\theta_{[i_o]}) \leq f(X|\theta_{[i_o+1]}) + \dots + f(X|\theta_{[n]}) \\ f(X|\theta_{[1]}) + \dots + f(X|\theta_{[i_o+1]}) > f(X|\theta_{[i_o+2]}) + \dots + f(X|\theta_{[n]}) \end{cases}$$

**None** of the previous methods **is** applicable in the case of MAP



# Data driven non-Bayesian estimation

Non-Bayesian estimation monopolized by MLE

For parametric density  $f(X|\theta)$  we are given  $X_1, \dots, X_n$  generated by same  $\theta$

$$\hat{\theta}_{\text{MLE}}(X) = \arg \max_{\theta} \sum_{i=1}^n \log f(X_i|\theta)$$

Asymptotically optimum: Approaches CRLB as  $n \rightarrow \infty$

Estimate obtained by combining data **and** conditional density!

Cannot replace density with data

For a data-driven version we propose an indirect definition of  $f(X|\theta)$

Start with  $Z \sim h(Z)$

Consider deterministic parametric transformation  $T(Z, \theta)$

Apply transformation on  $Z$  to generate  $X = T(Z, \theta)$  then  $X \sim f(X|\theta)$

$T(Z, \theta)$  : Known functional form, unknown parameters  $\theta$

$h(Z)$  : Unknown, instead  $Z_1, \dots, Z_m$

$f(X|\theta)$  : Unknown, instead  $X_1, \dots, X_n$  for the same  $\theta$

**Goal:** Estimate transformation parameters  $\theta$  from available data

We **do not have** correspondence  $X_i = T(Z_i, \theta)$

The two datasets  $\{Z_1, \dots, Z_m\}$ ,  $\{X_1, \dots, X_n\}$  are sampled **independently**

$$T(Z, \theta) = Z + \theta$$

$$T(Z, \Theta) = \Theta Z$$

$T(Z, \theta)$  can be nonlinear

$T(Z)$  can be completely unknown. In this case we approximate with neural network  $T(Z) \approx T(Z, \theta)$

**Problem:** Transform set  $\{Z_1, \dots, Z_m\}$  into  $\{Y_1, \dots, Y_m\}$  with  $Y_i = T(Z_i, \theta)$ . Compute parameters  $\theta$  so that  $\{Y_1, \dots, Y_m\}$  exhibits **same statistical behavior** as  $\{X_1, \dots, X_n\}$

## Moment Matching

$$\frac{1}{m} \sum_{i=1}^m (T(Z_i, \theta))^s \approx \frac{1}{n} \sum_{j=1}^n (X_j)^s, \quad s = s_1, s_2, \dots$$

Notoriously non-robust

# Density Matching

**Problem:** Compute parameters  $\theta$  so that  $\{Y_1, \dots, Y_m\}$  with  $Y_i = T(Z_i, \theta)$  have the same density as  $\{X_1, \dots, X_n\}$

## Adversarial

Similar to the design of Generative Adversarial Networks (GANs)

For  $Z \sim h(Z)$  design generator  $G(Z)$  such that  $Y = G(Z)$  follows density  $f(\cdot)$

$$\min_{G(Z)} \max_{D(X)} \left\{ \mathbb{E}_f[\log(1 - D(X))] + \mathbb{E}_h[\log(D(G(Z)))] \right\} \Rightarrow Y = G(Z) \sim f(\cdot)$$

$D(X) \in (0, 1)$  is known as the “Discriminator”

$$\min_G \max_D \left\{ \mathbb{E}_f[\phi(D(X))] + \mathbb{E}_h[\phi(D(G(Z)))] \right\} \leftarrow \text{more general version}$$

Here  $G(Z) \leftarrow T(Z, \theta)$  and  $D(X) \leftarrow D(X, \vartheta)$

$$\min_{\theta} \max_{\vartheta} \left\{ \frac{1}{n} \sum_{i=1}^n \log(1 - D(X_i, \vartheta)) + \frac{1}{m} \sum_{j=1}^m \log(D(T(Z_j, \theta), \vartheta)) \right\} \Rightarrow \hat{\theta}_0$$

$$\min_{\theta} \max_{\vartheta} \left\{ \frac{1}{n} \sum_{i=1}^n \phi(D(X_i, \vartheta)) + \frac{1}{m} \sum_{j=1}^m \phi(D(T(Z_j, \theta), \vartheta)) \right\} \Rightarrow \hat{\theta}_0$$

## Maximal Correlation

If  $K(X, Y)$  positive definite kernel then

$$\max_{G(Z)} \frac{\left( \mathbb{E}_{f,h} [K(X, G(Z))] \right)^2}{\mathbb{E}_{h,h} [K(G(Z^1), G(Z^2))]} \Rightarrow Y = G(Z) \sim f(\cdot)$$

where  $Z^1, Z^2$  independent following both  $h(Z)$

Here  $G(Z) \leftarrow T(Z, \theta)$

$$\max_{\theta} \frac{\left( \sum_{i=1}^n \sum_{j=1}^m K(X_i, T(Z_j, \theta)) \right)^2}{\sum_{j=1}^m \sum_{\substack{j'=1 \\ j \neq j'}}^m K(T(Z_j, \theta), T(Z_{j'}, \theta))}$$

# Examples

Let  $h_0(z)$  zero mean. Define  $h(z) = h_0(z - \mu)$ ,  $f(x|\theta) = h(x - \theta)$   
 $h_0(z)$ ,  $\mu$ ,  $\theta$  unknown. We are given  $\{z_1, \dots, z_m\} \sim h(z)$  and  
 $\{x_1, \dots, x_n\} \sim f(x|\theta)$ . Estimate  $\theta$

Moment matching:  $\frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{m} \sum_{j=1}^m z_j$

$$\varphi(w) = \begin{cases} w^2, & |w| \leq c \\ 2c|w| - c^2, & |w| \geq c \end{cases}$$

Huber estimator:  $\arg \min_v \sum_{i=1}^n \varphi(x_i - v) - \arg \min_{\mu} \sum_{i=1}^n \varphi(x_i - \mu)$

Maximal correlation:  $K(x, y) = e^{-\frac{1}{h}|x-y|}$

MLE:  $\arg \max_v \sum_{i=1}^n \log h_0(x_i - v) - \arg \max_{\mu} \sum_{j=1}^m \log h_0(z_j - \mu)$

Estimation error power for  $n = m = 100$  and  $\theta = \mu = 1$

	Gaussian	Laplace	Cauchy	
CRLB	0.020	0.020	0.040	
MLE	0.020	0.023	0.041	
Moment Matching	0.020	0.040	$\infty$	Data-driven
Huber Estimator	0.021	0.029	0.073	Data-driven
Maximal Correlation	0.022	0.025	0.045	Data-driven

95% of Gaussian

$h = 2\text{median}\{|x_i|\}$

Let  $h_0(z)$  zero mean. Define  $h(z) = \mu h_0(\mu z)$ ,  $f(x|\theta) = \theta h(\theta x)$ . Estimate  $\theta$   
 Estimation error power for  $n = m = 100$  and  $\theta = \mu = 1$

	Gaussian	Laplace	Cauchy	
CRLB	0.010	0.020	0.040	
MLE	0.010	0.021	0.043	
Moment Matching	0.012/0.010	0.021/0.025	$\infty/\infty$	Data-driven
Robust	0.028	0.045	0.053	Data-driven
Maximal Correlation	0.014	0.027	0.055	Data-driven

## Extensions

Start with  $Z, W \sim h(Z, W)$  and generate  $X = T(Z, W, \theta)$  then  $X \sim f(X|\theta)$

Estimate  $\theta$  from  $\{(Z_1, W_1), \dots, (Z_m, W_m)\}, \{X_1, \dots, X_n\}$

If  $Z, W$  independent we can treat the “noisy” data case:  $X = T(Z, \theta) + W$

We need:  $\{Z_1, \dots, Z_m\}, \{W_1, \dots, W_l\}, \{X_1, \dots, X_n\}$

$$\max_{\theta} \frac{\left( \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^l K(X_i, T(Z_j, \theta) + W_k) \right)^2}{\sum_{j=1}^m \sum_{k=1}^l \sum_{\substack{j'=1 \\ j \neq j'}}^m \sum_{k'=1}^l K(T(Z_j, \theta) + W_k, T(Z_{j'}, \theta) + W_{k'})} \Rightarrow \hat{\theta}_0$$

We can even have noise  $W$  with parameters that can be estimated in parallel with  $\theta$ .