Machine Learning for Signal Processing

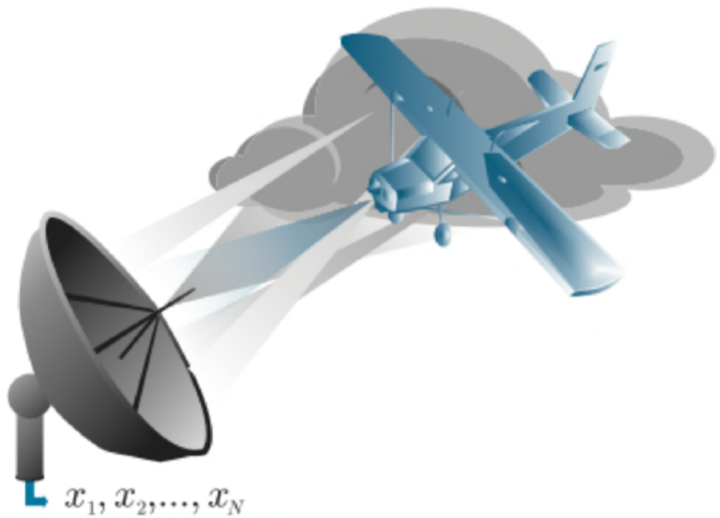
GEORGE V. MOUSTAKIDES
UNIVERSITY OF PATRAS, GREECE

Two-part presentation

- Decision making
- Parameter estimation

Decision making - Outline

- Mathematical Formulation
- Data Driven Approach
- Extensions
- Examples



Single dataset: $\{x_1, \ldots, x_n\}$
corresponding to two different scenarios (hypotheses)

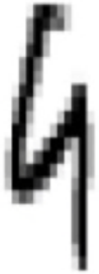
$$\mathrm{H}_0: \quad x_n \sim \text{pure noise}$$

$$\mathrm{H}_1: \quad x_n \sim \text{noise} + \underbrace{\text{reflection}}_{\text{Presence of airplane}}$$

Interested in distinguishing between handwritten numerals “4” and “9”

Single image $\Rightarrow$ Two scenarios

Distinguish between “4” and “9”

Labeled as “9” 

Labeled as “4” 

Hypothesis Testing - Decision Making - Classification

Same Mathematical Problem

Interested in Optimal Solution

Mathematical Formulation

For a random vector X we assume the following two hypotheses

$$\begin{aligned} \mathsf{H}_0 : &\quad X \sim \mathsf{f}_0(X), \ \ \mathbb{P}(\mathsf{H}_0) \\ \mathsf{H}_1 : &\quad X \sim \mathsf{f}_1(X), \ \ \mathbb{P}(\mathsf{H}_1) \end{aligned}$$

For every X need to decide if it comes from H_0 or H_1

Decide using a *Decision Function* $\mathsf{D}(X) \in \{0, 1\}$

Would like to optimize $\mathsf{D}(X)$

Plethora of applications in diverse scientific fields!!!

Bayesian Approach

Minimize decision error probability

$$\min_{\mathsf{D}} \left\{ \mathbb{P}(\mathsf{D} = 1 | \mathsf{H}_0)\mathbb{P}(\mathsf{H}_0) + \mathbb{P}(\mathsf{D} = 0 | \mathsf{H}_1)\mathbb{P}(\mathsf{H}_1) \right\}$$

$$\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)} \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \frac{\mathbb{P}(\mathsf{H}_0)}{\mathbb{P}(\mathsf{H}_1)} \equiv \frac{\mathsf{f}_1(X)\mathbb{P}(\mathsf{H}_1)}{\mathsf{f}_0(X)\mathbb{P}(\mathsf{H}_0)} \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 1$$

For $\omega(\mathsf{r})$ strictly increasing

$$\mathsf{r}(X) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 1 \equiv \omega\big(\mathsf{r}(X)\big) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1), \quad \mathsf{r}(X) = \frac{\mathsf{f}_1(X)\mathbb{P}(\mathsf{H}_1)}{\mathsf{f}_0(X)\mathbb{P}(\mathsf{H}_0)}$$

Data Driven Approach

$$\mathsf{H}_0: \quad X \sim \mathsf{f}_0(X), \quad \mathbb{P}(\mathsf{H}_0) \quad X_1^0 \; X_2^0 \; \ldots \; X_{n_0}^0$$

$$\mathsf{H}_1: \quad X \sim \mathsf{f}_1(X), \quad \mathbb{P}(\mathsf{H}_1) \quad X_1^1 \; X_2^1 \; \ldots \; X_{n_1}^1$$

Sampled from f_0

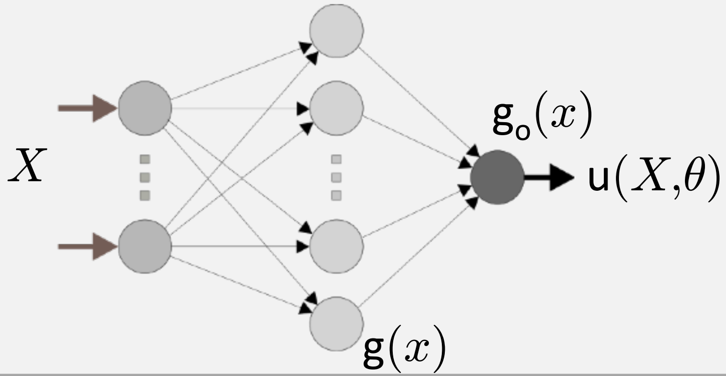
Sampled from f_1

$$\mathbb{P}(\mathsf{H}_i) \approx \frac{n_i}{n_0 + n_1}$$

Design a decision like function $\mathsf{v}(X) = \begin{cases} -1 & \text{when } X \text{ from } \mathsf{H}_0 \\ 1 & \text{when } X \text{ from } \mathsf{H}_1. \end{cases}$

Cybenko 1989 (universal approximation)

For sufficiently large neural network $\mathsf{u}(X,\theta)$ we can find suitable parameters θ such that we can approximate arbitrarily close any function $\mathsf{v}(X)$



$$|\mathsf{v}(X) - \mathsf{u}(X,\theta)| \leq \epsilon$$

Use neural network $\mathsf{u}(X,\theta)$ and optimize θ solving

$$\mathsf{J}(\theta) = \frac{1}{n_0 + n_1} \left\{ \sum_{i=1}^{n_0} \Big(-1 - \mathsf{u}(X_i^0, \theta) \Big)^2 + \sum_{j=1}^{n_1} \Big(1 - \mathsf{u}(X_j^1, \theta) \Big)^2 \right\}$$

$$\min_{\theta} \mathsf{J}(\theta) \;\Rightarrow\; \theta_o \;\Rightarrow\; \mathsf{u}(X, \theta_o)$$

For every X to test decide as follows: $\mathsf{u}(X, \theta_o) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 0$

Works “well”!! Why??

Understanding using Asymptotic Analysis

$$n_0, n_1 \to \infty, \qquad \mathsf{u}(X,\theta) \to \mathsf{v}(X)$$

$$\mathsf{J}(\theta) = \frac{n_0}{n_0+n_1} \frac{1}{n_0} \sum_{i=1}^{n_0} \Big(1 + \mathsf{u}(X_i^0,\theta)\Big)^2 + \frac{n_1}{n_0+n_1} \frac{1}{n_1} \sum_{j=1}^{n_1} \Big(1 - \mathsf{u}(X_j^1,\theta)\Big)^2$$

$$\mathsf{J}(\mathsf{v}) = \mathbb{P}(\mathsf{H}_0)\mathbb{E}_0\left[\Big(1+\mathsf{v}(X)\Big)^2\right] + \mathbb{P}(\mathsf{H}_1)\mathbb{E}_1\left[\Big(1-\mathsf{v}(X)\Big)^2\right]$$

$$\min_\theta \mathsf{J}(\theta) \to \min_{\mathsf{v}} \mathsf{J}(\mathsf{v})$$

$$\theta_o \Rightarrow \mathsf{u}(X,\theta_o) \approx \mathsf{v}_o(X)$$

$$\mathbb{E}_1\left[\Big(1-\mathsf{v}(X)\Big)^2\right]=\mathbb{E}_0\left[\Big(1-\mathsf{v}(X)\Big)^2\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}\right]$$

$$\mathsf{J}(\mathsf{v})=\mathbb{P}(\mathsf{H}_0)\mathbb{E}_0\left[\Big(1+\mathsf{v}(X)\Big)^2+\mathsf{r}(X)\Big(1-\mathsf{v}(X)\Big)^2\right] \qquad \mathsf{r}(X)=\frac{\mathsf{f}_1(X)\mathbb{P}(\mathsf{H}_1)}{\mathsf{f}_0(X)\mathbb{P}(\mathsf{H}_0)}$$

minimize for each X

$$\mathsf{v_o}(X)=\frac{\mathsf{r}(X)-1}{\mathsf{r}(X)+1}=\omega\big(\mathsf{r}(X)\big),$$ where $\omega(\mathsf{r})=\frac{\mathsf{r}-1}{\mathsf{r}+1}$ strictly increasing

Test equivalent to Bayes: $\mathsf{v_o}(X)=\omega\big(\mathsf{r}(X)\big)\underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}}\omega(1)=0 \quad\Rightarrow\quad \mathsf{u}(X,\theta_\mathsf{o})\underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}}0$

Equivalence in the limit

Develop data driven methods for estimation of $\omega\big(\mathsf{r}(X)\big)$ for other $\omega(\mathsf{r})$

Extensions to other functions

For strictly increasing function $\omega(\mathrm{r})$ can we define cost

$$\mathrm{J}(\mathrm{v}) = \mathbb{P}(\mathrm{H}_0)\mathbb{E}_0\left[\phi\big(\mathrm{v}(X)\big)\right] + \mathbb{P}(\mathrm{H}_1)\mathbb{E}_1\left[\psi\big(\mathrm{v}(X)\big)\right]$$

so that $\min_{\mathrm{v}} \mathrm{J}(\mathrm{v}) \Rightarrow \mathrm{v}_o(X) = \omega\big(\mathrm{r}(X)\big)$?

THEOREM: <u>Select</u> strictly increasing function $\omega(\mathrm{r})$ and strictly negative function $\rho(z)$. Define

$$\psi'(z) = \rho(z), \quad \phi'(z) = -\omega^{-1}(z)\rho(z)$$

then $\mathrm{v}_o(X) = \arg\min_{\mathrm{v}} \mathrm{J}(\mathrm{v}) = \omega\big(\mathrm{r}(X)\big)$

Examples of functions

A: $\omega(\mathsf{r}) = \mathsf{r} \in \mathbb{R}_+$ (likelihood ratio)

$$\rho(z) = -1, z \geq 0 \;\Rightarrow\; \phi(z) = \frac{z^2}{2},\;\; \psi(z) = -z$$

Mean Square

B: $\omega(\mathsf{r}) = \log(\mathsf{r}) \in \mathbb{R}$ (log-likelihood ratio)

$$\rho(z) = -e^{-0.5z} \;\Rightarrow\; \phi(z) = 2e^{0.5z},\;\; \psi(z) = 2e^{-0.5z}$$

Exponential

C: $\omega(\mathsf{r}) = \frac{\mathsf{r}}{\mathsf{r}+1} \in [0,1]$ (posterior probability)

$$\rho(z) = -\frac{1}{z}, z \in [0,1] \;\Rightarrow\; \phi(z) = -\log(1-z),\;\; \psi(z) = -\log(z)$$

Cross Entropy

Data Driven Implementation

$$\mathsf{J}(\mathsf{v}) = \mathbb{P}(\mathsf{H}_0)\mathbb{E}_0\left[\phi\big(\mathsf{v}(X)\big)\right] + \mathbb{P}(\mathsf{H}_1)\mathbb{E}_1\left[\psi\big(\mathsf{v}(X)\big)\right]$$

$$\mathsf{v}_\mathsf{o}(X) = \arg\min_{\mathsf{v}} \mathsf{J}(\mathsf{v}) = \omega\big(\mathsf{r}(X)\big)$$

$X_1^0 \; X_2^0 \; \ldots \; X_{n_0}^0$ Sampled from f_0 $\quad X_1^1 \; X_2^1 \; \ldots \; X_{n_1}^1$ Sampled from f_1

$$\mathsf{J}(\theta) = \frac{1}{n_0 + n_1}\left\{\sum_{i=1}^{n_0}\phi\big(\mathsf{u}(X_i^0,\theta)\big) + \sum_{j=1}^{n_1}\psi\big(\mathsf{u}(X_j^1,\theta)\big)\right\}$$

$$\min_\theta \mathsf{J}(\theta) \;\Rightarrow\; \theta_\mathsf{o} \;\Rightarrow\; \mathsf{u}(X,\theta_\mathsf{o}) \qquad \mathsf{u}(X,\theta_\mathsf{o}) \approx \omega\left(\frac{\mathsf{f}_1(X)\mathbb{P}(\mathsf{H}_1)}{\mathsf{f}_0(X)\mathbb{P}(\mathsf{H}_0)}\right)$$

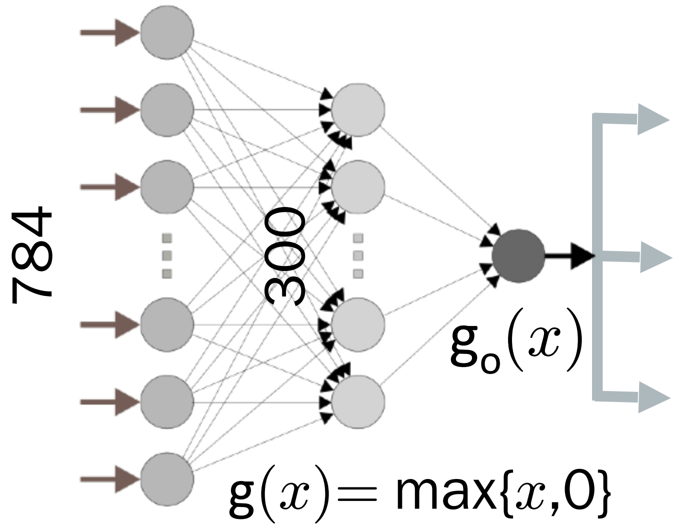
Close to optimum Bayes test: $\mathsf{u}(X,\theta_\mathsf{o}) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1)$

Example: Classification Problem

From dataset MNIST isolate handwritten numerals 4 and 9

Gray scale images 28 X 28 = 784 pixels. Design classifier using training data. Examine performance using testing data.

Neural network 784 X 300 X 1



Mean square $g_o(x) = \max\{x, 0\}$, $u_1(X, \theta_1)$

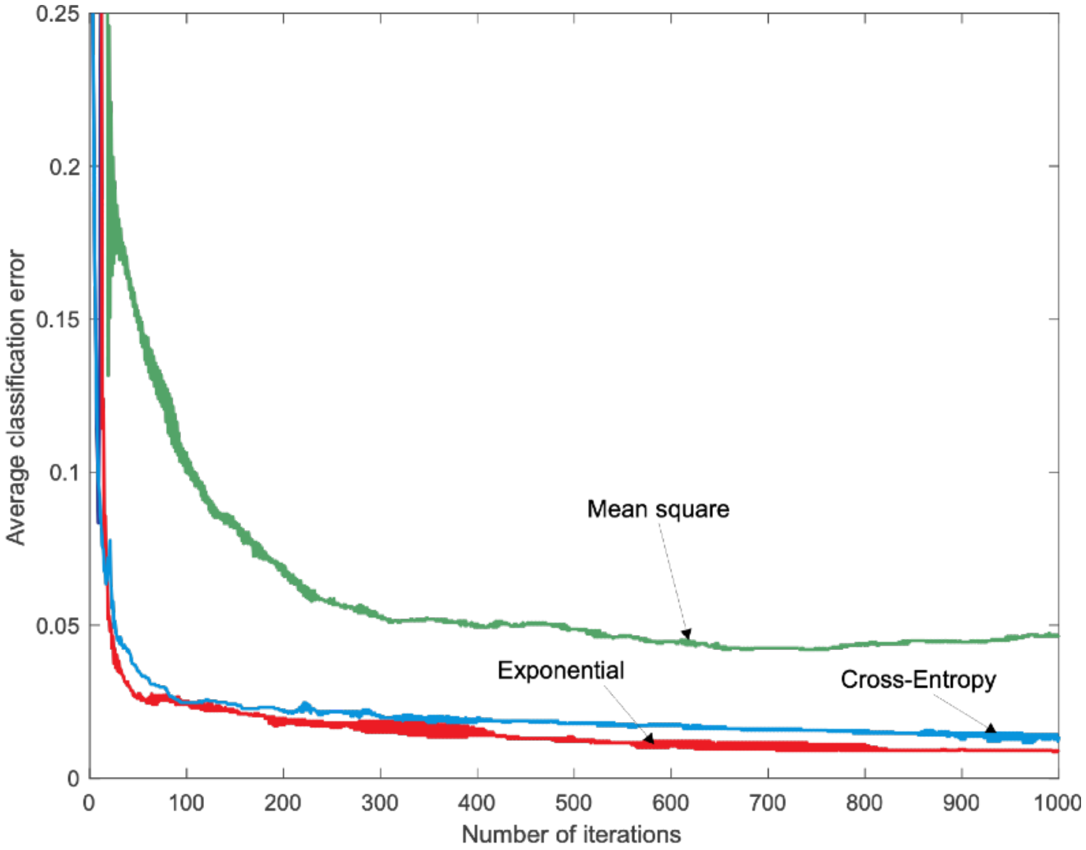
Exponential $g_o(x) = x$, $u_2(X, \theta_2)$

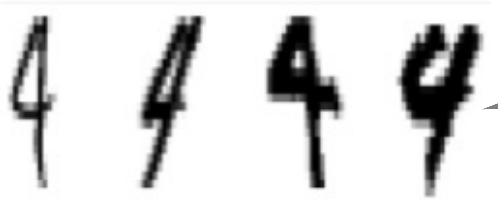
Cross entropy $g_o(x) =$ sigmoid, $u_3(X, \theta_3)$

# of parameters
235801

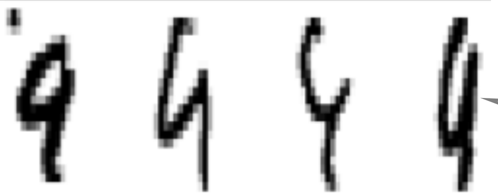
Training set: 5500 “4” and 5500 “9”. Testing set: 982 “4” and 1009 “9”

$$u_1(X,\theta_1) \underset{H_0}{\overset{H_1}{\gtreqless}} 1, \quad u_2(X,\theta_2) \underset{H_0}{\overset{H_1}{\gtreqless}} 0, \quad u_3(X,\theta_3) \underset{H_0}{\overset{H_1}{\gtreqless}} \frac{1}{2}$$





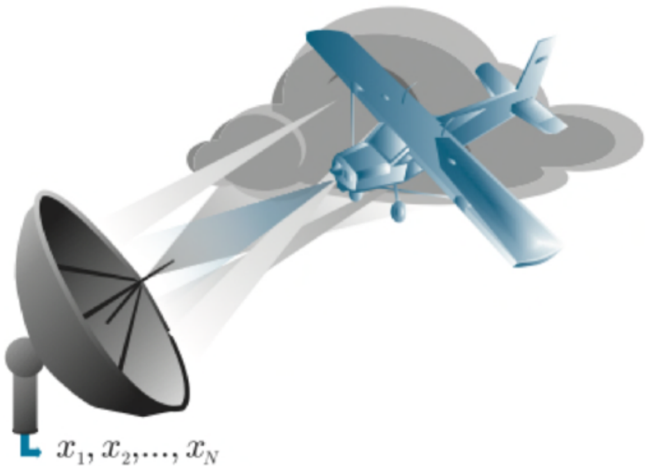
4 mistaken for 9



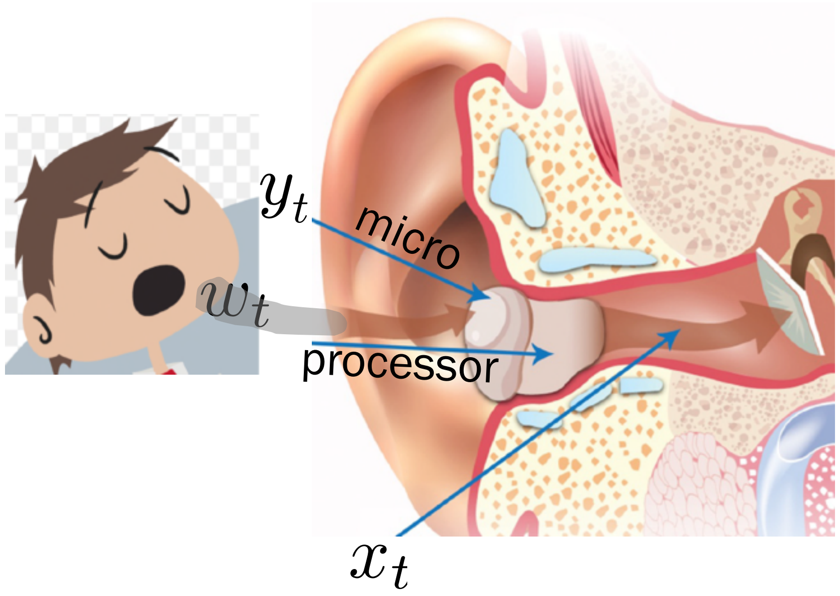
9 mistaken for 4

Parameter Estimation - Outline

- Data driven non-Bayesian estimation
- A class of parameter estimation problems
- Density matching
- Example



We would like to estimate speed and position



$$y_t = w_t + h_1 x_{t-\tau_1} + \cdots + h_k x_{t-\tau_k}$$

Echo, must be removed

Data driven non-Bayesian estimation

Non-Bayesian estimation monopolized by Maximum Likelihood Estimator (MLE)

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

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

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

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 $Z \sim \mathrm{h}(Z)$

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

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

$\mathrm{T}(Z,\theta)$: Known functional form, unknown parameters θ

$\mathrm{h}(Z)$: Unknown, instead $Z_1, \ldots, Z_m$

$\mathrm{f}(X|\theta)$: Unknown, instead $X_1, \ldots, X_n$ for the same θ

Goal: Estimate transformation parameters θ from available data

We do not have correspondence $X_i = \mathrm{T}(Z_i,\theta)$

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

$\mathsf{T}(Z, \theta) = Z + \theta$

$\mathsf{T}(Z, \Theta) = \Theta Z$

$\mathsf{T}(Z, \theta)$ can be nonlinear

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

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

Moment Matching

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

Notoriously non-robust

Density Matching

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

Maximal Correlation   If $\mathsf{K}(X,Y)$ positive definite kernel then

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

where Z^1, Z^2 independent following both $\mathsf{h}(Z)$

Here $\mathsf{G}(Z) \leftarrow \mathsf{T}(Z,\theta)$

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

Example

Let $\mathsf{h}_0(z)$ zero mean. Define $\mathsf{h}(z) = \mathsf{h}_0(z-\mu)$, $\mathsf{f}(x|\theta) = \mathsf{h}(x-\theta)$
$\mathsf{h}_0(z), \mu, \theta$ unknown. We are given $\{z_1, \dots, z_m\} \sim \mathsf{h}(z)$ and
$\{x_1, \dots, x_n\} \sim \mathsf{f}(x|\theta)$. Estimate θ

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

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

Huber estimator: $\arg\min\limits_{v} \sum_{i=1}^{n} \varphi(x_i - v) - \arg\min\limits_{\mu} \sum_{j=1}^{m} \varphi(z_j - \mu)$

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

MLE: $\arg\max\limits_{v} \sum_{i=1}^{n} \log \mathsf{h}_0(x_i - v) - \arg\max\limits_{\mu} \sum_{j=1}^{m} \log \mathsf{h}_0(z_j - \mu)$

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	∞	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

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