CS/CNS/EE/IDS 165: Foundations of Machine Learning and Statistical Inference

Maximum Likelihood Estimation

http://tensorlab.cms.caltech.edu/users/anima/cms165-2020.html

Anima Anandkumar
Computing and Mathematical Sciences
California Institute of Technology, Pasadena CA 91125
anima@caltech.edu

Outline

Topics

- The ML Estimator.
- Finite Sample Properties of MLE.
- Scoring and EM Methods.
- Asymptotic Properties of MLE.

References:

1. H.V. Poor, An Introduction to Signal Detection and Estimation, 2nd Ed., Springer-Verlag, 1994, Chapter IV-D.
2. B. Porat, Digital Processing of Random Signals: Theory and Methods, Prentice Hall, 1994.
3. S. M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice Hall, 1993.
4. P.J. Bickel and K.A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, Prentice Hall, Englewood Cliffs, NJ, 1977.
5. E.L. Lehmann, Theory of Point Estimation, Chapman & Hall, New York, 1991.

ML Estimation

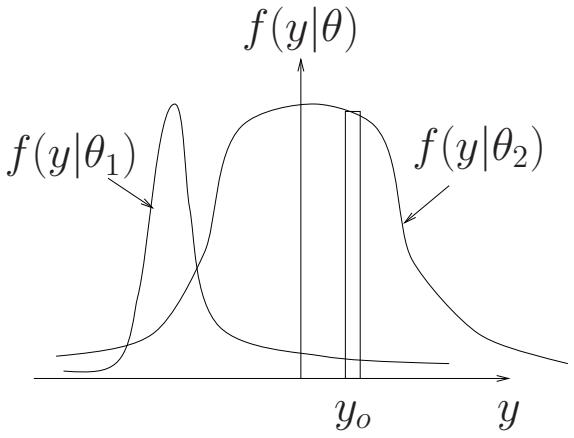
The Problem:

Estimate (vector) parameter θ from random vector $Y \sim f(y|\theta), \quad \theta \in \Lambda$.

The ML Estimation

$$\begin{aligned}\hat{\theta}_{\mathsf{ML}}(y) &= \arg\max_{\theta\in\Lambda} f(y|\theta) \\ &= \arg\max_{\theta\in\Lambda} \ln f(y|\theta).\end{aligned}$$

The intuition is that our observation y_o must come from the population that makes y_o most likely to occur.



Remarks

The density $f(y|\theta)$ as a function of θ is referred to as the likelihood function, and $\ln f(y|\theta)$ is the log-likelihood function. (In fact, functions of the form $g(y)f(y|\theta)$ can all be considered likelihood functions.)

An Example

Consider i.i.d. $Y_i \sim \mathcal{N}(\mu, \sigma^2), i = 1, \cdots N$. Estimate $\theta = [\mu, \sigma^2]^\intercal$.

1. The log-likelihood function: denote $y = [y_1, \cdots, y_n]^\intercal$

$$\ln f(y; \theta) = -\frac{1}{2\sigma^2}||y - \mu \mathbf{1}||^2 - \frac{N}{2} \ln 2\pi\sigma^2$$

2. Maximizing the log-likelihood function:

$$\frac{\partial}{\partial \mu} \ln f(y|\theta) \mid_{\theta = \begin{pmatrix} \hat{\mu}_{\mathsf{ML}} \\ \hat{\sigma}^2_{\mathsf{ML}} \end{pmatrix}} = \frac{1}{\hat{\sigma}^2_{\mathsf{ML}}}(y^\intercal \mathbf{1} - \hat{\mu}_{\mathsf{ML}} N) = 0$$

$$\frac{\partial}{\partial \sigma^2} \ln f(y|\theta) \mid_{\theta = \begin{pmatrix} \hat{\mu}_{\mathsf{ML}} \\ \hat{\sigma}^2_{\mathsf{ML}} \end{pmatrix}} = \frac{1}{2(\hat{\sigma}^2_{\mathsf{ML}})^2}||y - \hat{\mu}_{\mathsf{ML}} \mathbf{1}||^2 - \frac{N}{2\hat{\sigma}^2_{\mathsf{ML}}} = 0$$

3. The ML estimator:

$$\hat{\mu}_{\mathsf{ML}} = \frac{y^\intercal \mathbf{1}}{N} = \frac{1}{N} \sum y_i$$

$$\hat{\sigma}^2_{\mathsf{ML}} = \frac{1}{N}||\mathbf{y} - \hat{\mu}_{\mathsf{ML}} \mathbf{1}||^2 = \frac{1}{N} \sum_i (y_i - \hat{\mu}_{\mathsf{ML}})^2$$

Note: $\hat{\sigma}^2_{\mathsf{ML}}$ is biased and it is not efficient!

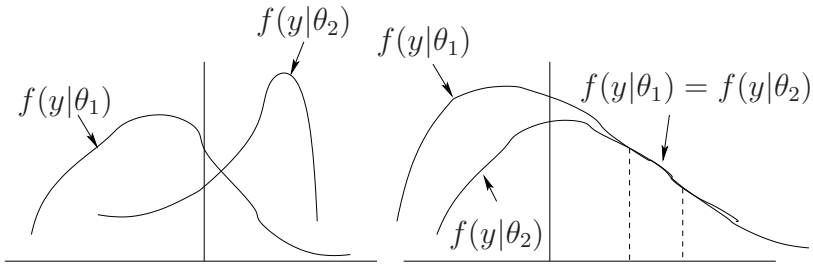
Identifiability

Definition Two parameter values θ_1 and θ_2 are observational equivalent

$$f(y|\theta_1) = f(y|\theta_2) \quad a.e.$$

A parameter θ is (globally) identifiable if there is no other $\theta' \in \Lambda$ observationally equivalent to θ, *i.e.,*

$$f(y|\theta) = f(y|\theta') \ a.e. \Rightarrow \theta = \theta'$$



Identifiable Unidentifiable

Example: For the linear model, $y = H\theta + w$, $w \sim \mathcal{N}(0, C)$, θ is identifiable iff H has full column rank.

Proof: "Only if": If H does not have full column rank, then there exists $\theta_\perp \neq 0$ such that $\mathbf{H}\theta_\perp = 0$. Hence

$$y = H\theta + w = H(\theta + \theta_\perp) + w, \quad \forall \theta_\perp \in \text{Ker}(H)$$

"if": Let θ and $\bar{\theta}$ be such that $f(y|\theta) = f(y|\bar{\theta})$. We then have

$$(y - H\theta)'C^{-1}(y - H\theta) = (y - H\bar{\theta})'C^{-1}(\mathbf{y} - H\bar{\theta}) \ \ a.e.$$

$$2(\theta - \bar{\theta})^\intercal H^\intercal C^{-1} y = (H\theta)^\intercal C^{-1} H\theta - (H\bar{\theta})^\intercal C{-}1 H\bar{\theta} \quad a.e.$$

$$\theta = \bar{\theta}$$

ML for Linear Models

Theorem If y is from the linear model

$$y = H\theta + w,$$

where H is known with full column rank, and $w \sim \mathcal{N}(0, \Sigma)$.

1. The ML estimate of θ is given by a linear estimator
$$\hat{\theta}_{\mathsf{ML}} = (H^{\intercal}\Sigma^{-1}H)^{-1}H^{\intercal}\Sigma^{-1}y.$$
2. $\hat{\theta}_{\mathsf{ML}}$ is efficient (hence UMVU).
3. $\hat{\theta}_{\mathsf{ML}} \sim \mathcal{N}(\theta, (H^{\intercal}\Sigma^{-1}H)^{-1})$.

$$\begin{aligned}\nabla_\theta \ln f(y|\theta) &= \nabla_\theta\{-\frac{1}{2}(y-H\theta)^{\intercal}\Sigma^{-1}(y-H\theta)\} \\ &= H^{\intercal}\Sigma^{-1}(y-H\theta) \\ &= (H^{\intercal}\Sigma^{-1}H)(\hat{\theta}_{\mathsf{ML}}-\theta)\end{aligned}$$

BLUE

The best linear unbiased estimator (BLUE) is defined by $\hat{\theta}_{\mathsf{BLUE}} = A_{\mathsf{BLUE}}y$ where

$$A_{\mathsf{BLUE}} = \arg\min \mathbb{E}(||\theta - AY||^2) \quad \text{subject to} \quad \mathbb{E}(AY) = \theta$$

Corollary

For the linear model with zero mean but otherwise arbitrarily distributed $\mathbf{w}$,

$$\hat{\theta}_{\mathsf{BLUE}} = \hat{\theta}_{\mathsf{ML}}$$

ML for the Exponential Family

Theorem: Consider the K-parameter exponential family

$$f(y|\theta) = exp\{\sum_{i=1}^{K} c_i(\theta)t_i(y) + d(\theta) + s(y)\}.$$

Let $\mathcal{C}$ be the interior of the range of $[c_1(\theta), \cdots, c_K(\theta)]^\intercal$ (assuming the existence). If

$$\mathbb{E}(t_i(Y)) = t_i(y), \quad i = 1, \cdots, K$$

have a solution $\hat{\theta}(y)$ for which $[c_1(\hat{\theta}), \cdots, c_K(\hat{\theta})]^\intercal \in \mathcal{C}$, then $\hat{\theta}$ is the unique ML estimator of θ.

Proof for the scalar 1-parameter case:

$$f(y|\theta) = \exp(\theta t(y) + d(\theta) + s(y)).$$

From the moment generating function of $t(y)$

$$\phi(\omega) := \mathbb{E}[e^{\omega t(y)}] = \int exp((\theta+\omega)t(y) + d(\theta) + s(y))dy = exp(d(\theta) - d(\theta+\omega))$$

The first two moments of $t(y)$ are given by

$$\begin{aligned} \mathbb{E}(t(Y)) &= -d'(\theta), \\ \mathbb{E}(t^2(Y)) &= -d''(\theta) + \mathbb{E}^2(t(Y)) \rightarrow d''(\theta) = -\mathsf{Var}(t(Y)) \leq 0 \end{aligned}$$

The ML estimator is the solution of the ML equation,

$$\frac{\partial}{\partial\theta} \ln f(y|\theta)|_{\theta=\hat{\theta}_{ML}} = t(y) + d'(\theta_{ML}) = 0.$$

The uniqueness comes from

$$\frac{\partial^2}{\partial\theta^2} \ln f(y|\theta)|_{\theta=\hat{\theta}_{ML}} = d''(\theta_{ML}) \leq 0$$

Efficiency and Invariance

Efficiency

If the efficient estimator exists, it is the ML estimator.

Proof: If the CRB is achieved by an unbiased estimator $\hat{\theta}(y)$,

$$\frac{\partial}{\partial \theta} \ln f(y|\theta) = \mathbf{I}(\theta)(\hat{\theta}(y) - \theta) \quad \text{a.s.} \rightarrow \hat{\theta}_{\mathsf{ML}} = \hat{\theta}(y)$$

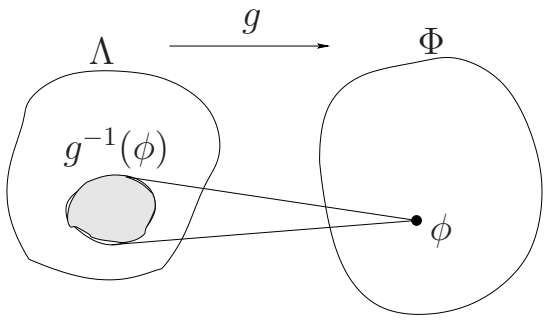
Invariance

Let $g(\theta)$ be a function from Λ onto Φ. For any $\phi \in \Phi$, let $g^{-1}(\phi) \in \Lambda$ be the inverse image of ϕ. Define the induced likelihood function $l(y;\phi)$ by

$$l(y;\phi) \stackrel{\Delta}{=} \sup_{\theta \in g^{-1}(\phi)} f(y|\theta)$$

If $\hat{\theta}_{\mathsf{ML}}$ is the ML estimate of θ,

$$\hat{\phi}_{\mathsf{ML}} \stackrel{\Delta}{=} \arg\sup_{\phi \in \Phi} l(y;\phi) = g(\hat{\theta}_{\mathsf{ML}})$$



Proof:

$$f(y|\hat{\theta}_{\mathsf{ML}}) \leq \sup_{\theta \in g^{-1}(g(\hat{\theta}_{\mathsf{ML}}))} f(y|\theta) = l(y; g(\hat{\theta}_{\mathsf{ML}})) \leq \sup_{\phi \in \Phi} l(\mathbf{y};\phi) = \sup_{\theta \in \Theta} f(y;\theta) = f(y|\hat{\theta}_{\mathsf{ML}})$$

ML and Kullback-Leibler Distance

The Kullback-Leibler Distance

To measure the "distance" between $f(y|\theta_0)$ and $f(y|\theta_1)$, the Kullback Leibler Distance is defined by

$$D(P_0||P_1) \triangleq \mathbb{E}_{\theta_0}(\ln \frac{f(Y|\theta_0)}{f(Y|\theta_1)}) = \int f(y|\theta_0) \ln \frac{f(y|\theta_0)}{f(y|\theta_1)} d\mathbf{y}$$

We will also use the notation $D(\theta_0||\theta_1)$. Note that $D(P_0||P_1)$ is not a true distance measure $(D(P_0||P_1) \neq D(P_1||P_0))$.

Property: $D(P_0||P_1) \geq 0$ with equality iff $f(Y|\theta_0) = f(Y|\theta_1)$, a.e.. If θ_0 is identifiable, then $D(P_0||P_1) = 0 \Leftrightarrow \theta_1 = \theta_0$

Proof: By the Jensen's inequality, since $-\ln x$ is strictly convex, we have

$$\mathbb{E}_{\theta_0}(-\ln \frac{f(y|\theta_0)}{f(y|\theta)}) \leq \ln(\mathbb{E}_{\theta_0}\{\frac{f(y|\theta)}{f(y|\theta_0)}\}) = 0$$

with equality iff $\frac{f(y|\theta_0)}{f(y|\theta)}$ is a constant.

Remarks

- θ_0 is the global minimum of $D(\theta_0||\theta)$, and

$$\min_{\theta} D(\theta_0||\theta) \leftrightarrow \max \mathbb{E}_{\theta_0} \ln f(y|\theta)$$

- If we have i.i.d $Y_i \sim f(y|\theta_0)$, then

$$\hat{\theta}_{\mathsf{ML}} = \arg\max_{\theta}\{\frac{1}{N}\sum_i \ln f(y_i|\theta)\} \rightarrow \arg\max_{\theta} \mathbb{E}_{\theta_0}\{\ln f(Y|\theta)\}$$

Therefore, ML is the same as minimizing wrt empirical distribution.

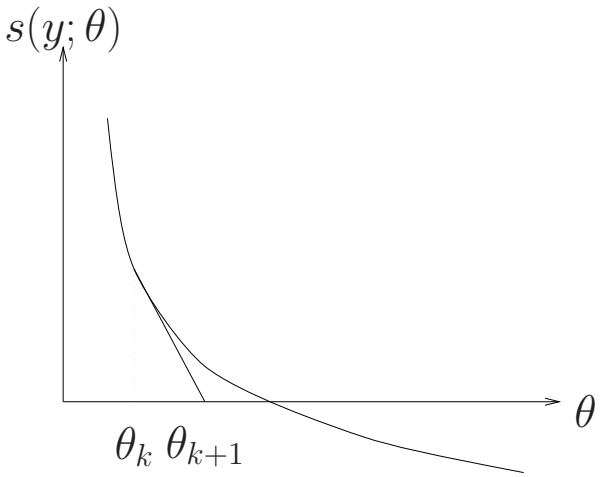
Numerical Methods

To solve the ML equation

$$s(\mathbf{y};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{y};\boldsymbol{\theta}) = \mathbf{0}$$

The Newton Raphson Iteration

$$\theta_{k+1} = \theta_k - \mathbf{J}^{-1}(\theta_k; y) s(y;\theta_k), \quad \mathbf{J}(\theta; y) = \frac{\partial^2 \ln f(y|\theta)}{\partial\theta\partial\theta^\intercal}$$



The Scoring Method

$$\theta_{k+1} = \theta_k + \mathbf{I}^{-1}(\theta_k) s(y;\theta_k), \quad \mathbf{I}(\theta) = -\mathbb{E}\{\frac{\partial^2 \ln f(y|\theta)}{\partial\theta\partial\theta^\intercal}\}$$

Remark:

Neither iterative algorithm guarantees convergence.

The EM Algorithm

Estimation with Missing Data

Let the complete data $z = [s^\intercal \quad y^\intercal]^\intercal \sim f(z|\theta)$. Suppose that only $y \sim f(y|\theta)$ is observed. The MLE is given by

$$\hat{\theta}_{\mathsf{ML}} = \arg\max \ln f(y|\theta)$$

The EM Approach:

- Consider the joint log-likelihood function

$$\ln f(y|\theta) = \ln f(z;\theta) - \ln f(z|y;\theta)$$

- Taking the conditional expectation (on $Y = y$) under parameter θ'.

$$\ln f(y|\theta) = \underbrace{\mathbb{E}_{\theta'}[\ln f(Z|\theta) \mid Y = y]}_{Q(\theta;\theta')} - \underbrace{\mathbb{E}_{\theta'}[\ln f(Z|y;\theta) \mid Y = y]}_{P(\theta;\theta')}$$

- Around θ'

$$\begin{aligned} \ln f(y|\theta) \;\; - \;\; & \ln f(y|\theta') = Q(\theta;\theta') - Q(\theta';\theta') \\ + \;\; & \underbrace{\{P(\theta';\theta') - P(\theta;\theta')\}}_{\text{KL distance } \geq 0} \end{aligned}$$

- The Key Step:

$$Q(\theta'';\theta') > Q(\theta';\theta') \Rightarrow \ln f(y|\theta'') > \ln f(y|\theta')$$

Hence, improving $Q(\theta;\theta')$ also improves likelihood.

The EM Algorithm

The EM Algorithm

- Pick an initial θ_0
- At the kth iteration:
  - The E-step: Form the conditional expectation

$$Q(\theta;\theta_k) \stackrel{\Delta}{=} \mathbb{E}_{\theta_k}[\ln f(Z|\theta) \mid Y = y]$$

  - The M-step: Maximization

$$\theta_{k+1} = \arg\max_{\theta} Q(\theta;\theta_k)$$

- $k \uparrow$

Remarks

- Guaranteed convergence, and potentially simpler implementation.
- Global convergence is not guaranteed.
- Usually the optimization at the M-step is global.
- The joint density function is exploited.
- The trick is the selection of complete data $\mathbf{z}$ and the computation of the conditional mean.