Empirical Bayes methods & the EM algorithm

(I) Setup & Notation

(II) Introductory Example

(III) Basic Ideas

(IV) EM Algorithm

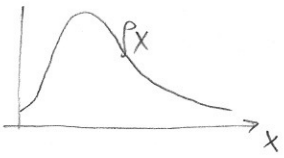
Efron – Large-scale inference : empirical Bayes methods for estimation, testing and prediction (2012)

McLachlan, Krishnan – The EM algorithm and extensions (2007)

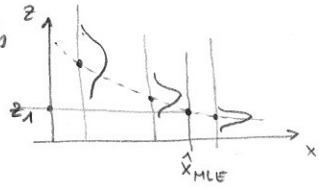
(I) Setup & Notation

$X \in \mathbb{R}^d$ unknown parameter, $Z \in \mathbb{R}^n$ measurement

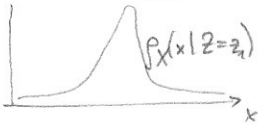
Last time



Input:
- prior ρ_X of X
- likelihood model $\{\rho_Z(\cdot \mid X = x) \mid x \in \mathbb{R}^d\}$
- $Z_1, \dots, Z_M \sim Z_{True}$, $Z_{True} \sim \rho_Z(\cdot \mid X = x_{True})$ measurements for the same person / individual with true parameter value x_{True}



Output:
- Estimator $\hat{x}_M$ of $X = x_{True}$ & its uncertainty



Today

Input:
- ~~prior ρ_X~~
- likelihood model $\{\rho_Z(\cdot \mid X = x) \mid x \in \mathbb{R}^d\}$
- $\vec{z} = (z_1, \dots, z_M)$ measurements of several (M) individuals with M parametrizations X_j, i.e.

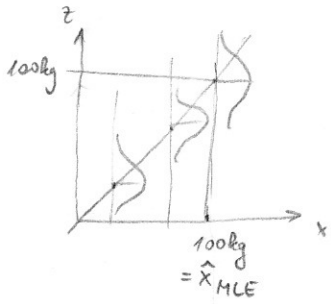
$$Z_m \sim \rho_Z(\cdot \mid X = X_m) \text{ indep.}, \quad X_m \overset{iid}{\sim} \underbrace{\rho_X}_{\text{unknown}}$$

$\vec{X} := (X_1, \dots, X_M)$

Output:
- Estimation of ρ_X
- Estimators $\hat{x}_1, \dots, \hat{x}_M$ of $X_1, \dots, X_M$ (& their uncertainty)

(II) Introductory Example

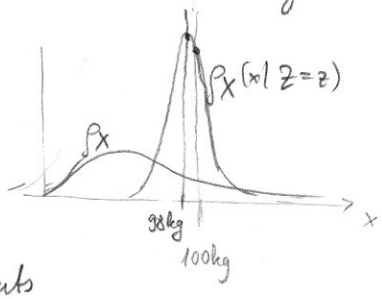
(a) Estimating the weight X of a person after measuring it with inaccurate weighing scales:



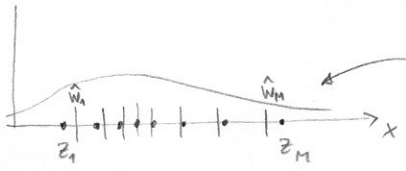
Assume that we measure $z = 100\,kg$.

Is $\hat{x}_{MLE} = 100\,kg$ a good estimator? Yes, unless we have additional knowledge:

(b) If we know the weight distribution ρ_X within the population, Bayes' rule gives us better criteria for estimating $X|Z=z$; e.g. $\hat{x}_{MAP} = \max_x \rho_X(x|Z=z) \overset{\text{maybe}}{=} 98\,kg$.



(c) Usually, ρ_X is unknown, but assume we have measurements $Z_1 = z_1, \dots, Z_M = z_M$ for several persons (with unknown weights $X_1, \dots, X_M \sim \rho_X$)



better estimators than $\hat{x}_{m,MLE} = z_m$ for $M \geq 3$ (in the mean squared error sense), see also "James Stein Estimator"

(III) Basic ideas

Roughly speaking,

(1) MLE treats every person separately

(2) Bayesian approach ——— " ———

(3) Empirical Bayes methods (i) first gather all measurements to estimate the prior f_X
(ii) then use this information to infer better individual estimators (usually by Bayes' rule)

[(i) and (ii) are sometimes hidden within one formula for the estimators]

We will concentrate on (i), approximations of f_X will be denoted by f.

Main observation: Viewing the prior as a parameter ("hyper-parameter") $\mathcal{F} = f$ (with $f_{true} = f_X$), the measurements $Z_1, \dots, Z_M$ are iid realisations of the same experiment (last time's case) with likelihood model

$$f_Z(z \mid \mathcal{F}=f) = \int f_{Z|X}(z|x) \underbrace{f_X(x \mid \mathcal{F}=f)}_{(=f(x) \text{ in non-parametric case})} dx \quad \text{"(marginal) likelihood"}$$

and total likelihood

$$f_{\vec{Z}}(\vec{z} \mid \mathcal{F}=f) = \int f_{\vec{Z}|\vec{X}}(\vec{z} \mid \vec{X}=\vec{x}) f_{\vec{X}}(\vec{x} \mid \mathcal{F}=f) \, d\vec{x} = \prod_{m=1}^{M} f_Z(z_m \mid \mathcal{F}=f)$$

discrete case: $$\mathbb{P}(Z=z \mid \mathcal{F}=f) = \sum_x \mathbb{P}(Z=z \wedge X=x \mid \mathcal{F}=f) = \sum_x \mathbb{P}(Z=z \mid X=x \wedge \mathcal{F}=f)\,\mathbb{P}(X=x \mid \mathcal{F}=f)$$
$$= \sum_x \mathbb{P}(Z=z \mid X=x)\,\mathbb{P}(X=x \mid \mathcal{F}=f)$$

So, now we can perform MLE estimation (and even Bayesian inference, after specifying a "hyper-prior", an a priori distribution of all possible priors) for the hyper-parameter $\mathcal{F}$ using the new likelihood model $\{ f_Z(\cdot \mid \mathcal{F}=f) \mid f \in \mathfrak{F} \}$, where $\mathfrak{F}$ = set of considered priors

Two cases:

- We know / assume that the prior f_X has a certain parametric form, e.g. $f_X = \mathcal{N}(\mu, \Sigma)$ (μ, Σ unknown)
  $\rightsquigarrow$ hyper-parameters $= (\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}$ spd.), finite-dim. problem
- f_X is completely unknown $\rightsquigarrow$ hyper-parameter $= \mathcal{F} \in \mathcal{M}_{\mathbb{R}^d}$, ∞-dim. problem
  ($\mathcal{M}_{\mathbb{R}^d}$ = set of all prob. distr. on $\mathbb{R}^d$)

In both cases, the EM algorithm is the standard tool to maximize the (marginal) likelihood (PMLE & NPMLE).

(IV) Expectation - maximization (EM) algorithm

Iterative algorithm that (locally) maximizes the (marginal) log-likelihood

$$\mathcal{L}(f) = \log \rho_{\vec{Z}}(\vec{z} \mid \mathcal{F}=f) = \log \prod_{m=1}^{M} \rho_Z(z_m \mid \mathcal{F}=f) = \sum_{m=1}^{M} \log \int \rho_Z(z \mid X=x) \underbrace{\rho_X(x \mid \mathcal{F}=f)}_{\text{often } f(x)} \, dx$$

Starting with f_0 iterate:

E-step: Formulate the "complete data ($\vec{X}=\vec{x}$, $Z=\vec{z}$) likelihood fct."

$$L_c(\vec{x}, \vec{z} \mid \mathcal{F}=f) = \prod_{m=1}^{M} \rho_Z(z_m \mid X=x_m) \underbrace{\rho_X(x_m \mid \mathcal{F}=f)}_{\text{often } f(x)}$$

and the expectation value in $\vec{X}$ given $\mathcal{F}=f_n$, $\vec{Z}=\vec{z}$

$$Q(f \mid f_n) := \mathbb{E}\left[\log L_c\left((\vec{X} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n), \vec{z} \mid \mathcal{F}=f \right) \right]$$

$$\left(= \mathbb{E}\left[\log L_c(\vec{X}, \vec{z} \mid \mathcal{F}=f) \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n \right] \right)$$

$$= \int \log L_c(\vec{x}, \vec{z} \mid \mathcal{F}=f) \underbrace{\rho_{\vec{X}}(\vec{x} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n)}_{\substack{\text{density of } (\vec{X} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n) \\ = \text{posterior, if } f_n \text{ was the prior}}} d\vec{x}$$

M-step: maximize $Q(\cdot \mid f_n)$: $f_{n+1} = \operatorname{argmax}_f Q(f \mid f_n)$

(IV.1) Some theory:

1) $\mathcal{L}(f) = \log \rho_{\vec{Z}}(\vec{z} \mid \mathcal{F}=f) = \log \int \rho_{\vec{Z}}(\vec{z} \mid \vec{X}=\vec{x}) \rho_{\vec{X}}(\vec{x} \mid \mathcal{F}=f) \cdot \frac{\rho_{\vec{X}}(\vec{x} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n)}{\rho_{\vec{X}}(\vec{x} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n)} d\vec{x}$ (density in $\vec{x}$)

$$\overset{\text{Jensen}}{\geq} \int \rho_{\vec{X}}(\vec{x} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n) \log\left(\frac{\rho_{\vec{Z}}(\vec{z} \mid \vec{X}=\vec{x}) \rho_{\vec{X}}(\vec{x} \mid \mathcal{F}=f)}{\rho_{\vec{X}}(\vec{x} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n)} \right) d\vec{x}$$

$$= \mathcal{L}(f_n) + \underbrace{\int \rho_{\vec{X}}(\vec{x} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n) \log\left(\frac{\rho_{\vec{Z}}(\vec{z} \mid \vec{X}=\vec{x}) \rho_{\vec{X}}(\vec{x} \mid \mathcal{F}=f)}{\rho_{\vec{X}}(\vec{x} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n) \rho_{\vec{Z}}(\vec{z} \mid \mathcal{F}=f_n)} \right) d\vec{x}}_{=: \ell(f \mid f_n)}$$

2) $\ell(f \mid f_n) = 0$ for $f = f_n$, because then denominator $\overset{\text{Bayes}}{=} \frac{\rho_{\vec{X}}(\vec{x} \mid \mathcal{F}=f_n) \rho_{\vec{Z}}(\vec{z} \mid \vec{X}=\vec{x}, \mathcal{F}=f_n)}{\rho_{\vec{Z}}(\vec{z} \mid \mathcal{F}=f_n)} \rho_{\vec{Z}}(\vec{z} \mid \mathcal{F}=f_n) =$ numerator

3) $\operatorname{argmax}_f \ell(f \mid f_n) = \operatorname{arg\,max}_f \int \rho_{\vec{X}}(\vec{x} \mid \vec{Z}=\vec{z}, \mathcal{F}=f_n) \log \underbrace{\left(\rho_{\vec{Z}}(\vec{z} \mid \vec{X}=\vec{x}) \rho_{\vec{X}}(\vec{x} \mid \mathcal{F}=f) \right)}_{L_c(\vec{x}, \vec{z} \mid \mathcal{F}=f)} d\vec{x} = \operatorname{arg\,max}_f Q(f \mid f_n)$

Therefore:

$$f_{n+1} = \operatorname{argmax}_f Q(f \mid f_n) \overset{(3)}{=} \operatorname{argmax}_f \ell(f \mid f_n) \overset{(2)}{\Rightarrow} \ell(f_{n+1} \mid f_n) \geq 0 \overset{(1)}{\Rightarrow} \boxed{\mathcal{L}(f_{n+1}) \geq \mathcal{L}(f_n)}$$

Picture?

(IV.2) Further applications of EM

The EM algorithm has further applications in statistics and data analysis, e.g. in incomplete data analysis
grouped / censored / truncated

Example: You model the age of death (AoD) of patients undergoing some treatment by an exponential distribution $X_m = \text{AoD}_m \sim \text{Exp}(\mu)$ with unknown mean $\mathcal{M} = \mu$

$$f_X(x \mid \mathcal{M}=\mu) = \mu^{-1} \exp(-x/\mu) \cdot \mathbb{1}_{(0,\infty)}$$

However, when evaluating the data some patients (w.l.o.g. $m = 1, \dots, \tau$) are still alive, so your data is incomplete:

$$z_m = (a_m, d_m)$$

(a_m: age reached; d_m: $d_m = \begin{cases} 1 & \text{if uncensored (person died at age } a_m) \\ 0 & \text{if censored} \quad (\text{AoD} \geq a_m) \end{cases}$)

Since $\text{Exp}(\mu)$ is memoryless we have for $m = 1, \dots, \tau$

$$(X_m \mid Z_m = z_m) = a_m + \tilde{X}_m \quad , \text{ where } \tilde{X}_m \sim \text{Exp}(\mu)$$

E-step:

$$L_c(\vec{x}, \vec{z} \mid \mathcal{M}=\mu) = \prod_{m=1}^{M} \mu^{-1} \exp(-x_m/\mu) = \prod_{m=1}^{M} \mu^{-1} \exp(-a_m/\mu) \cdot \prod_{m=1}^{\tau} \exp(-\tilde{x}_m/\mu)$$

$$\log L_c(\vec{x}, \vec{z} \mid \mathcal{M}=\mu) = -M \log\mu - \sum_{m=1}^{M} a_m/\mu - \sum_{m=1}^{\tau} \tilde{x}_m/\mu$$

$$Q(\mu \mid \mu_n) = \mathbb{E}\left[\log L_c\left((\vec{X} \mid \vec{z}=\vec{z}, \mathcal{M}=\mu_n), \vec{z} \mid \mathcal{M}=\mu \right) \right] = -M\log\mu - \sum_{m=1}^{M} a_m/\mu - \sum_{m=1}^{\tau} \frac{\mu_n}{\mu}$$

M-step:

$$Q(\mu \mid \mu_n) = -M \log\mu - \mu^{-1} \left(\tau \cdot \mu_n + \sum_{m=1}^{M} a_m \right)$$

$$0 \overset{!}{=} \partial_\mu Q(\mu \mid \mu_n) = -\frac{M}{\mu} + \frac{1}{\mu^2} \left(\tau \cdot \mu_n + \sum_{m=1}^{M} a_m \right)$$

$$\Rightarrow \mu_{n+1} = \frac{1}{M} \left(\tau \cdot \mu_n + \sum_{m=1}^{M} a_m \right)$$

$\mu_n \longrightarrow \mu_*$ with $\mu_* = \frac{1}{M} \left(\tau \mu_* + \sum_{m=1}^{M} a_m \right)$, i.e. $\mu_* = \frac{1}{M-\tau} \sum_{m=1}^{M} a_m$

μ_* finely maximizes the likelihood

$$f_{\vec{Z}}(\vec{z} \mid \mathcal{M}=\mu) = \prod_{m=1}^{\tau} \underbrace{\mathbb{P}(X \geq a_m \mid \mathcal{M}=\mu)}_{\exp(-a_m/\mu)} \cdot \prod_{m=\tau+1}^{M} \underbrace{f_X(a_m \mid \mathcal{M}=\mu)}_{\mu^{-1}\exp(-a_m/\mu)}$$

as can be seen by differentiating its logarithm ∎