12. Statistical Estimation

- conditional expectation
- mean square estimation (MSE)
- maximum likelihood estimation (MLE)
- maximum a posteriori estimation (MAP)
- Cramér-Rao inequality
- properties of MLE
- linear model with additive noise

Conditional expectation

let x, y be random variables with a joint density function $f(x, y)$

the conditional expectation of x given y is

$$\mathbf{E}[x|y] = \int x f(x|y) dx$$

where $f(x|y)$ is the conditional density: $f(x|y) = f(x, y)/f(y)$

Facts:

- $\mathbf{E}[x|y]$ is a function of y
- $\mathbf{E}[\mathbf{E}[x|y]] = \mathbf{E}[x]$
- for any scalar function $g(y)$ such that $\mathbf{E}[|g(y)|^2] < \infty$

$$\mathbf{E}\left[(x - \mathbf{E}[x|y])g(y)\right] = 0$$

Mean square estimation

suppose x, y are random with a joint distribution

problem: find an estimate $h(y)$ that minimizes the mean square error:

$$\mathbf{E}\|x - h(y)\|^2$$

result: the optimal estimate in the mean square is *the conditional mean*:

$$h(y) = \mathbf{E}[x|y]$$

Proof. use the fact that $x - \mathbf{E}[x|y]$ is uncorrelated with any function of y

$$\begin{aligned}\mathbf{E}\|x - h(y)\|^2 &= \mathbf{E}\ \|x - \mathbf{E}[x|y] + \mathbf{E}[x|y] - h(y)\|^2 \\ &= \mathbf{E}\ \|x - \mathbf{E}[x|y]\|^2 + \mathbf{E}\ \|\mathbf{E}[x|y] - h(y)\|^2\end{aligned}$$

hence, the error is minimized only when $h(y) = \mathbf{E}[x|y]$

Gaussian case: x, y are jointly Gaussian: $(x, y) \sim \mathcal{N}(\mu, \Sigma)$ where

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{bmatrix}$$

the conditional density function of x given y is also Gaussian with conditional mean

$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y),$$

and conditional covariance matrix

$$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^T$$

hence, for Gaussian distribution, the optimal mean square estimate is

$$\mathbf{E}[x|y] = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y),$$

the optimal estimate is **linear** in y

conclusions:

- $\mathbf{E}[x|y]$ is called the minimum mean square error (MMSE) estimator
- the MMSE estimator is typically nonlinear in y and is obtained from $f(x, y)$
- for Gaussian case, the MMSE estimator is **linear** in y
- the MMSE estimator must satisfy the **orthogonal principle**:

$$\mathbf{E}[(x - \hat{x}_{\mathrm{mmse}})g(y)] = 0$$

  where g is any function of y such that $\mathbf{E}[|g(y)|^2] < \infty$

- MMSE estimator can be difficult to evaluate, so one can consider a linear MMSE estimator

Linear MMSE estimator

the linear unbiased MMSE estimator takes the affine form:

$$h(y) = K\tilde{y} + \mathbf{E}[x], \quad \text{(with } \tilde{y} = y - \mathbf{E}[y])$$

important results: define $\tilde{x} = x - \mathbf{E}[x]$

- the linear MMSE estimator minimizes

$$\mathbf{E}\|x - h(y)\|^2 = \mathbf{E}\|\tilde{x} - K\tilde{y}\|^2$$

- the linear MMSE estimator is

$$h(y) = \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mathbf{E}[y]) + \mathbf{E}[x]$$

- the form of linear MMSE requires just covariance matrices of x, y
- it coincides with the optimal mean square estimate for Gaussian RVs

Wiener-Hopf equation

the optimal condition for linear MMSE estimator

$$\Sigma_{xy} = K\Sigma_{yy}$$

- obtained by differentiating MSE w.r.t. K

$$\text{MSE} = \mathbf{E}\,\mathbf{tr}(\tilde{x} - K\tilde{y})(\tilde{x} - K\tilde{y})^T = \mathbf{tr}(\Sigma_{xx} - \Sigma_{xy}K^T - K\Sigma_{yx} + K\Sigma_y K^T)$$

$$\frac{\partial \text{MSE}}{\partial K} = -\Sigma_{yx} - \Sigma_{yx} + 2\Sigma_{yy}K^T = 0$$

- also obtained from the condition

$$\mathbf{E}[(x - h(y))y^T] = 0 \quad \Rightarrow \quad \mathbf{E}[(\tilde{x} - K\tilde{y})\tilde{y}^T] = 0$$

(the optimal residual is uncorrelated with the observation y)

Minimum variance unbiased estimator (MVUE)

for any estimate $h(y)$, the covariance matrix of the corresponding error is

$$C = \mathbf{E}\left[(x - h(y))(x - h(y))^T\right]$$

- different choices of h lead to different covariances, say C_1, C_2
- we can compare two matrices in *matrix sense* by saying

$$C_1 \succeq C_2 \quad \text{if} \quad C_1 - C_2 \succeq 0 \quad \text{(the difference is positive semidefinite)}$$

- if $C_1 \succeq C_2$ then $\mathbf{tr}(C_1) \geq \mathbf{tr}(C_2)$ (MSE 1 is is bigger than MSE 2)

problem: restrict $h(y)$ to the linear case:

$$h(y) = Ky + c$$

and choose $h(y)$ to yield the **minimum covariance** (instead of minimum MSE) the covariance matrix can be written as

$$(\mu_x - (K\mu_y + c))(\mu_x - (K\mu_y + c))^T + \Sigma_x - K\Sigma_{yx} - \Sigma_{xy}K^T + K\Sigma_y K^T$$

the objective is minimized with respect to c when

$$c = \mu_x - K\mu_y$$

(same as the best unbiased linear estimate of the mean square error)

the covariance matrix of the error can be expressed as a quadratic function in K

$$f(K) = \Sigma_{xx} - K\Sigma_{yx} - \Sigma_{xy}K^T + K\Sigma_y K^T = \begin{bmatrix} -I & K \end{bmatrix} \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{bmatrix} \begin{bmatrix} -I \\ K^T \end{bmatrix} \succeq 0$$

let K_0 be a solution to the Wiener-Hopf equation: $\Sigma_{xy} = K_0\Sigma_{yy}$, we can write

$$f(K) = f(K_0) + (K - K_0)\Sigma_{yy}(K - K_0)^T$$

so $f(K)$ is minimized when $K = K_0$

the miminum covariance matrix is

$$f(K_0) = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^T$$

for $\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{bmatrix}$, note that

- the minimum covariance matrix is the Schur complement of Σ_{xx} in Σ
- it is exactly a conditional covariance matrix for Gaussian variables
- in conclusion, the linear MVUE estimate is given by

$$h(y) = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$$

  (same as linear MMSE estimator and MMSE estimator in Gaussian case)
- note: in order to compute the estimate, we only need up to second-moment of x and y (no distribution is assumed)

Maximum likelihood estimation

- log-likelihood function
- maximum likelihood principle
- models with and without predictors
- dynamical models, linear regression models

Log-likelihood function

setting: let $(y_1, \ldots, y_N)$ be i.i.d. observations of random variable Y with pdf

$$f(y; \theta^\star), \quad \text{and } \theta^\star \text{ is unknown}$$

likelihood function: the joint pdf of $y = (y_1, y_2, \ldots, y_N)$

$$\ell(\theta; y) = f(y_1, y_2, \ldots, y_N; \theta) = \prod_{i=1}^{N} f(y_i; \theta)$$

- $f(y_1, y_2, \ldots, y_N; \theta)$ is a function of data and parametrized by θ
- view ℓ as function of θ, giving a likelihood of θ that fits well with data

log-likelihood function: take the logarithmic function of ℓ

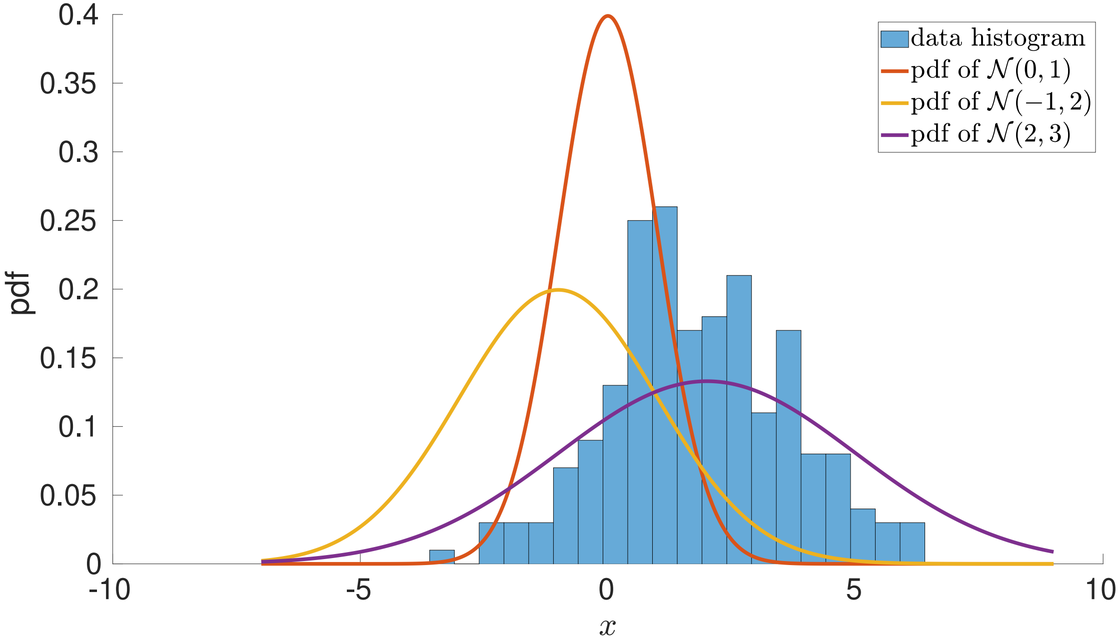
$$\mathcal{L}(\theta; y) = \sum_{i=1}^{N} \log f(y_i; \theta)$$

Maximum likelihood principle

the distribution of data, $f(y;\theta)$, is known but θ is to be estimated

MLE principle: choose θ that the observed data becomes *as likely as possible*

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \ \mathcal{L}(\theta; y) := \sum_{i=1}^{N} \log f(y_i; \theta)$$



MLE estimate must satisfy the zero-gradient condition:

$$\nabla_\theta \mathcal{L}(\theta; y) = 0$$

$\mathcal{N}(2,3)$ is more likely to explain data better than other Gaussians

example 1: estimate the mean and covariance matrix of Gaussian RVs

- observe a sequence of i.i.d. random variables: $y_1, y_2, \ldots, y_N$
- each y_k is an n-dimensional Gaussian: $y_k \sim \mathcal{N}(\mu, \Sigma)$, but μ, Σ are unknown
- the likelihood function of $y_1, \ldots, y_N$ for given μ, Σ is

$$\begin{aligned}\ell(\mu, \Sigma; y) &= f(y_1, y_2, \ldots, y_N | \mu, \Sigma) \\ &= \frac{1}{(2\pi)^{Nn/2}} \cdot \frac{1}{|\Sigma|^{N/2}} \cdot \textbf{exp} - \frac{1}{2} \sum_{k=1}^{N} (y_k - \mu)^T \Sigma^{-1} (y_k - \mu)\end{aligned}$$

- the **log-likelihood function** is (up to a constant)

$$\mathcal{L}(\mu, \Sigma) = \log \ell(\mu, \Sigma; y) = \frac{N}{2} \log\det \Sigma^{-1} - \frac{1}{2} \sum_{k=1}^{N} (y_k - \mu)^T \Sigma^{-1} (y_k - \mu)$$

- the log-likelihood is concave in Σ^{-1}, μ, so the ML estimate satisfies the zero gradient conditions:

$$\frac{\partial L}{\partial \Sigma^{-1}} = \frac{N\Sigma}{2} - \frac{1}{2}\sum_{k=1}^{N}(y_k - \mu)(y_k - \mu)^T = 0$$

$$\frac{\partial L}{\partial \mu} = \sum_{k=1}^{N}\Sigma^{-1}(y_k - \mu) = 0$$

- we obtain the ML estimate of μ, Σ as

$$\hat{\mu}_{\mathrm{ml}} = \frac{1}{N}\sum_{k=1}^{N} y_k, \quad \hat{\Sigma}_{\mathrm{ml}} = \frac{1}{m}\sum_{k=1}^{m}(y_k - \hat{\mu}_{\mathrm{ml}})(y_k - \hat{\mu}_{\mathrm{ml}})^T$$

  - $\hat{\mu}_{\mathrm{ml}}$ is the sample mean and $\hat{\Sigma}_{\mathrm{ml}}$ is a (biased) sample covariance matrix
  - in this example, MLE estimate is obtained in closed-form

Models with predictors

assume RVs X (predictor) and Y (response) with joint pdf

$$f_{xy}(x, y; \theta^\star), \quad \text{and } \theta^\star \text{ is unknown}$$

let $z = \{(x_i, y_i)\}_{i=1}^N$ be i.i.d. observations of (X, Y), we can write

$$\ell(y, x; \theta) = f(y, x; \theta) = f(y|x; \theta) f(x; \theta)$$

in regression, we aim to explain y when x is given

MLE problem is then to maximize the *conditional log-likelihood* of y given x

$$\mathcal{L}(\theta) = \sum_{i=1}^N \log f(y_i \mid x_i; \theta)$$

(though x is random, its values are given beforehand)

example 2: we aim to explain the number of car accidents (y) from

$x =$ number of junctions, populations, sold liquor, and incoming cars $\quad (x \in \mathbf{R}^4)$

- $\{(x_i, y_i)\}_{i=1}^N$ are i.i.d. observations collected from several cities
- y should be modeled as Poisson(λ)
- we model $\lambda = e^{x^T\theta}$ to *link* the mean of y with predictors

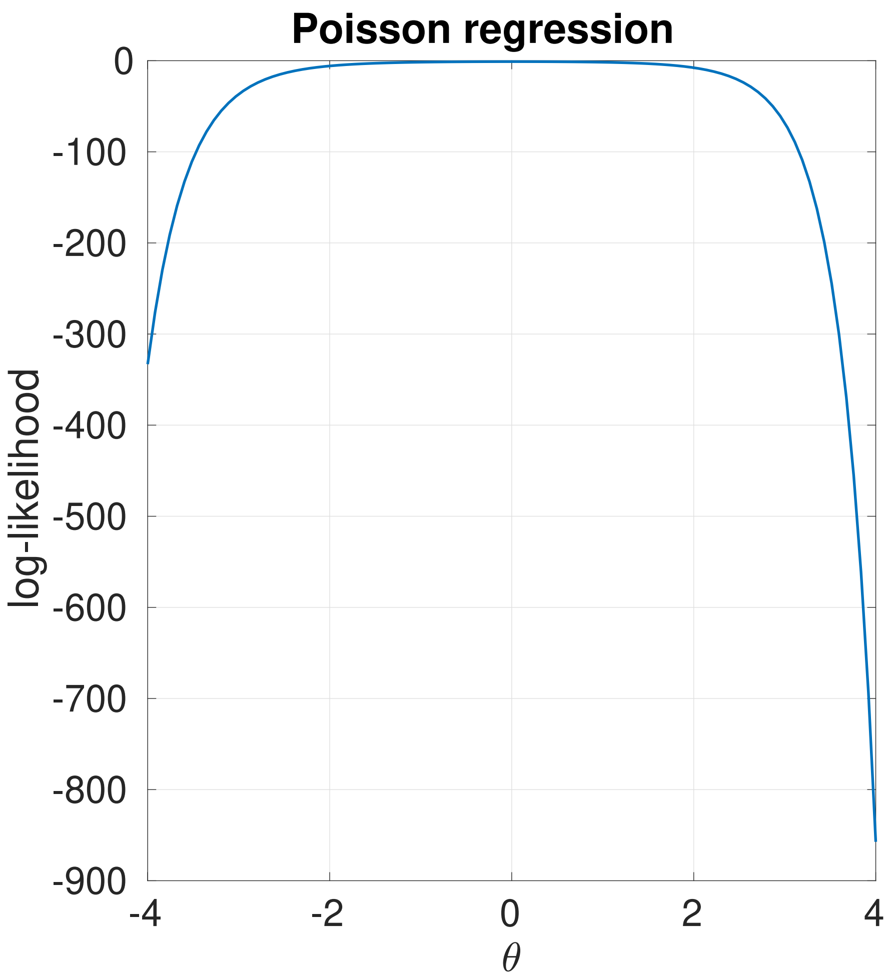
from the assumed model, the likelihood function of ith sample is

$$f(y_i|x_i;\theta) = e^{-\exp(x_i^T\theta)}\exp(x_i^T\theta)^{y_i}/y_i!$$

the conditional log-likelihood function of y given x is

$$\mathcal{L}(\theta) = \sum_{i=1}^N -e^{x_i^T\theta} + y_i x_i^T\theta - \log(y_i!)$$

the zero gradient condition is $\nabla_\theta \mathcal{L}(\theta) = \sum_{i=1}^{N} \left(-x_i e^{x_i^T \theta} + y_i x_i \right) = 0$



the log-likelihood function is concave (use a numerical method to fine $\hat{\theta}$)

Linear model with additive noise

when y and x have a linear relationship with i.i.d. corrupted noise, $e_i \sim f_e(e)$

$$y_i = x_i^T \beta + e_i, \quad i = 1, 2, \ldots, N$$

settings: i.i.d. observations $\{(x_i, y_i)\}_{i=1}^N$ are given and β is to be estimated

unlike the least-squares apporach, we can use *statistical info* of noise in estimation

- when x_i is given, the variable $y_i | x_i$ is an affine transformation of e_i

$$f(y_i | x_i; \beta) = f_e(y_i - x_i^T \beta)$$

- since data are i.i.d., and e_i's are all distributed by f_e

$$\mathcal{L}(\beta; y | x) = \sum_{i=1}^N \log f_e(y_i - x_i^T \beta)$$

MLE as minimizing MSE

estimate β in a linear model with **Gaussian noise** $f_e(u) = \frac{e^{-u^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$

$$
\begin{aligned}
\mathcal{L}(\beta, \sigma^2; y|x) &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 \\
&\triangleq -(N/2)\log(2\pi\sigma^2) - (1/2\sigma^2)\|y - X\beta\|_2^2
\end{aligned}
$$

- maximizing $\mathcal{L}(\beta, \sigma^2; y|x)$ over β is equivalent to minimizing $\|y - X\beta\|_2$
- from the zero gradient condition, the estimate of noise variance is

$$
\hat{\sigma}^2_{\mathrm{mle}} = (1/N)\|y - X\hat{\beta}\|_2^2
$$

- ML estimation of linear model with additive **Gaussian** noise is equivalent to a least-squares problem

MLE as minimizing MAE

estimate β in a linear model with **Laplacian noise** $f_e(u) = (1/2\lambda)e^{-|u|/\lambda}$

$$\begin{aligned}\mathcal{L}(\beta, \lambda; y|x) &= -N\log(2\lambda) - \frac{1}{\lambda}\sum_{i=1}^{N} |y_i - x_i^T\beta| \\ &\triangleq -N\log(2\lambda) - (1/\lambda)\|y - X\beta\|_1\end{aligned}$$

- maximizing $\mathcal{L}(\beta, \lambda; y|x)$ over β is equivalent to minimizing $\|y - X\beta\|_1$
- from the zero gradient condition, the ML estimate of noise variance is

$$\hat{\lambda}_{\text{mle}} = (1/N)\|y - X\hat{\beta}\|_1$$

- ML estimation of linear model with additive **Laplacian** noise is equivalent to an ℓ_1-norm estimation

Maximum a posteriori (MAP) estimation

assumption: θ is a *random variable* and jointly distributed with $f(y,\theta)$

the MAP estimate of θ is to maximize the **posterior** density (after observing y)

$$\hat{\theta} = \operatorname*{argmax}_{\theta} f_{\theta|y}(\theta|y)$$

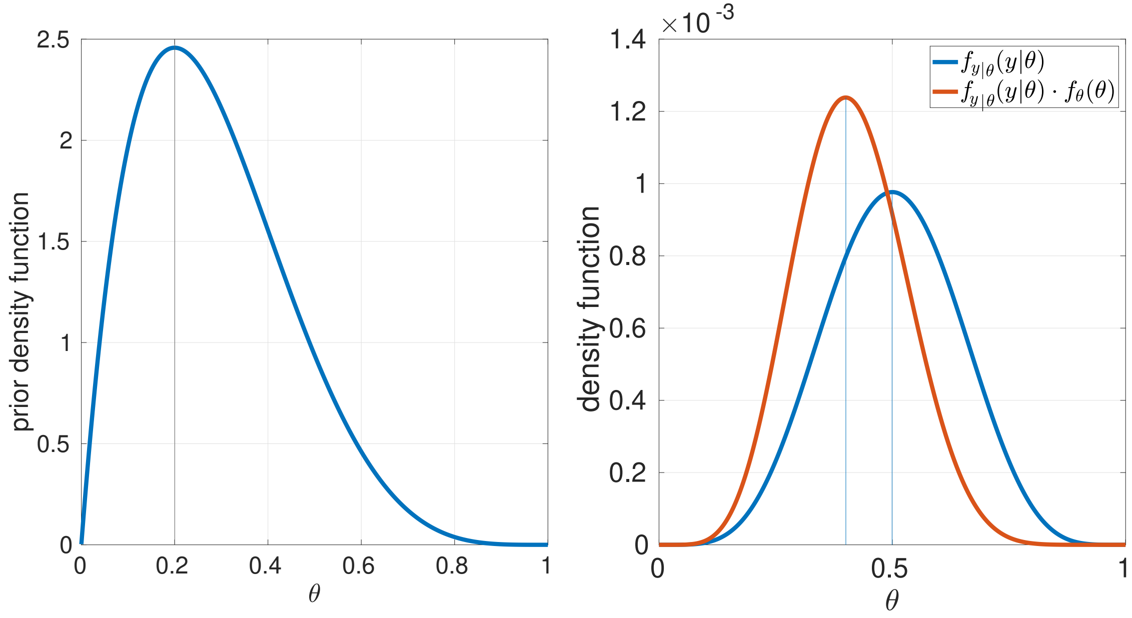
from Bayes' rule

$$f_{\theta|y}(\theta|y) = \frac{f_{y,\theta}(y,\theta)}{f_y(y)} = \frac{f_{y|\theta}(y|\theta)f_\theta(\theta)}{f_y(y)}$$

since $f_y(y)$ is not a function of θ, MAP estimation is equivalent to

$$\hat{\theta}_{\mathrm{map}} = \operatorname*{argmax}_{\theta} \;\; f_{y|\theta}(y|\theta)\cdot f_\theta(\theta) = \operatorname*{argmax}_{\theta} \;\; \log f_{y|\theta}(y|\theta) + \log f_\theta(\theta)$$

we give a varying weight of $f_{y|\theta}(y|\theta)$ for each θ given by the **prior** density of θ, $f_\theta(\theta)$



- the only difference between ML and MAP estimate is the term $f_\theta(\theta)$
- f_θ provides a prior knowledge about θ; hence, $\log f_\theta(\theta)$ penalizes choices of θ that are unlikely to happen
- (left.) from prior density, $\theta = 2$ is most likely to occur
- (right.) MLE gives $\hat{\theta} = 0.5$ but MAP estimate is < 0.5 as $f_\theta(0.5)$ is very small

under what condition on f_θ is the MAP estimate identical to the ML estimate ?

MAP estimation of linear model

consider a linear model with i.i.d. additive Gaussian noise: $y_i = x_i^T \beta + e_i$

when assuming β is random with a prior density $f_\beta(\beta)$, MAP estimation is

$$\underset{\beta}{\text{maximize}} \quad -\frac{1}{2\sigma^2}\sum_{k=1}^{N}(y_i - x_i^T\beta)^2 + \log f_\beta(\beta)$$

- **Gaussian prior:** $\beta \sim \mathcal{N}(0, \alpha I)$ (ℓ_2-regularized least-squares)

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{\sigma^2}\|y - X\beta\|_2^2 + \frac{1}{\alpha}\|\beta\|_2^2$$

- **Laplacian prior:** $f_\beta(\beta) = (1/2\alpha)e^{-\|\beta\|_1/\alpha}$ (ℓ_1-regularized least-squares)

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{\sigma^2}\|y - X\beta\|_2^2 + \frac{1}{\alpha}\|\beta\|_1$$

ML regularity conditions

let $\{(x_i, y_i)\}_{i=1}^N$ be i.i.d. samples according to MLE problem

$$s_i(\theta) = \frac{\partial \log f(y_i|x_i,\theta)}{\partial \theta} = \frac{1}{f(y_i|x_i,\theta)} \nabla_\theta f(y_i|x_i,\theta)$$

is called **score of the loglikelihood**

ML regularity conditions are

1. expected score is zero

$$\mathbf{E}_{y|x}\left[\nabla_\theta \log f(y|x;\theta)\right] = \int \nabla_\theta \log f(y|x;\theta) f(y|x;\theta) dy = 0$$

2. expected outer product of score is the negative expected Hessian of score

$$-\mathbf{E}_{y|x}\left[\nabla_\theta^2 \log f(y|x;\theta)\right] = \mathbf{E}_{y|x}\left[(\nabla_\theta \log f(y|x;\theta))(\nabla_\theta \log f(y|x;\theta))^T\right]$$

Cramér-Rao inequality

for any **unbiased** estimator $\hat{\theta}$ with the error covariance

$$\mathbf{cov}(\hat{\theta}) = \mathbf{E}(\theta - \hat{\theta})(\theta - \hat{\theta})^T$$

we always have a lower bound on $\mathbf{cov}(\hat{\theta})$:

$$\mathbf{cov}(\hat{\theta}) \succeq \left[\mathbf{E}(\nabla_\theta \log f(y|x;\theta))^T(\nabla_\theta \log f(y|x;\theta))\right]^{-1} = -\left(\mathbf{E}\left[\nabla_\theta^2 \log f(y|x;\theta)\right]\right)^{-1}$$

- the RHS is called the **Cramér-Rao** lower bound where two equal terms obtained by ML regularity condition
- provide the minimal covariance matrix over all possible estimators $\hat{\theta}$
- $\mathcal{I}(\theta) \triangleq -\mathbf{E}[\nabla_\theta^2 \log f(y|x;\theta)]$ is called the **Fisher information matrix** (note: $\log f(y|x;\theta) := \log f(y_1,\ldots,y_N|x_1,\ldots,x_N;\theta)$)
- an estimator for which the C-R equality holds is called **efficient**

Cramér Rao bound for linear model estimate

revisit a linear model with correlated Gaussian noise:

$$y = X\beta + e, \quad X \in \mathbf{R}^{N \times n}, \quad e \sim \mathcal{N}(0, \Sigma)$$

the density function $f(y|X;\beta)$ is given by $f_e(y - X\beta)$ which is Gaussian

$$\begin{aligned}
\log f(y|X;\beta) &= -\frac{1}{2}(y - X\beta)^T \Sigma^{-1} (y - X\beta) - \frac{N}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma \\
\nabla_\theta \log f(y|X;\beta) &= X^T \Sigma^{-1}(y - X\beta) \\
\nabla_\theta^2 \log f(y|X;\beta) &= -X^T \Sigma^{-1} X
\end{aligned}$$

hence, for any unbiased estimate $\hat{\beta}$,

$$\mathbf{cov}(\hat{\beta}) \succeq (X^T \Sigma^{-1} X)^{-1}$$

compare this LB with covariance of estimators you have seen ?

Linear models with additive noise

estimate parameters in a linear model with additive noise:

$$y = X\beta + e, \quad e \sim \mathcal{N}(0, \Sigma), \quad \Sigma \text{ is known}$$

and we explore several estimates from the following approaches

- no use of noise information
  - least-squares estimate (LS)
- use information about the noise (e.g., Gaussian distribution, Σ)

assume β is a fixed parameter	assume $\beta \sim \mathcal{N}(0, \Lambda)$
weighted least-squares (WLS)	minimum mean square (MMSE)
best linear unbiased (BLUE)	maximum a posteriori (MAP)
maximum likelihood (ML)	

least-squares: $\hat{\beta}_{\rm ls} = (X^TX)^{-1}X^Ty$ and is unbiased

$$\mathbf{cov}(\hat{\beta}_{\rm ls}) = \mathbf{cov}((X^TX)^{-1}X^Te) = (X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}$$

we can verifty that $\mathbf{cov}(\hat{\beta}_{\rm ls}) \succeq (X^T\Sigma^{-1}X)^{-1}$

it is bigger than the CR bound but the inequality is tight when $\Sigma = \sigma^2 I$ (the noise e_i's are uncorrelated)

generalized LS estimate (or BLUE): $\hat{\beta}_{\sf blue} = (X^T\Sigma X)^{-1}X^T\Sigma^{-1}y$
(obtained from the normalized model by $\Sigma^{-1/2}$)

$$\mathbf{cov}(\hat{\beta}_{\sf blue}) = (X^T\Sigma^{-1}X)^{-1}$$

(the covariance matrix achieves the CR bound)

weighted least-squares: for a given weight matrix $W \succ 0$

$$\hat{\beta}_{\text{wls}} = (X^T W X)^{-1} X^T W y \quad \text{and is unbiased}$$

it follows that the covariance of estimator is

$$\begin{aligned}
\mathbf{cov}(\hat{\beta}_{\text{wls}}) &= \mathbf{cov}((X^T W X)^{-1} X^T W e) \\
&= (X^T W X)^{-1} X^T W \Sigma W X (X^T W X)^{-1}
\end{aligned}$$

$\mathbf{cov}(\hat{\beta}_{\text{wls}})$ attains the minimum (the CR bound) when $W = \Sigma^{-1}$

$$\hat{\beta}_{\text{wls}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y$$

interpretation:

- large Σ_{ii} means the ith measurement is highly uncertain
- should put less weight on the corresponding ith entry of the residual

maximum likelihood: from $f(y|X;\beta) = f_e(y - X\beta)$,

$$\log f(y|X;\beta) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma - \frac{1}{2}(y - X\beta)^T\Sigma^{-1}(y - X\beta)$$

the zero gradient condition gives

$$\nabla_\beta \log f(y|X;\beta) = X^T\Sigma^{-1}(y - X\beta) = 0$$

$$\hat{\beta}_{\mathrm{ml}} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$$

$\hat{\beta}_{\mathrm{ml}}$ is also efficient (achieves the minimum covariance matrix)

as this point, three types of estimators (for linear model) are identical

$$\hat{\beta}_{\mathrm{ml}} = \hat{\beta}_{\mathrm{wls}} = \hat{\beta}_{\mathrm{blue}}$$

minimum mean square estimate:

- β is random and independent of e
- $\beta \sim \mathcal{N}(0, \Lambda)$

hence, y and β are jointly Gaussian with zero mean and the covariance:

$$C = \begin{bmatrix} C_\beta & C_{\beta y} \\ C_{\beta y}^T & C_{yy} \end{bmatrix} = \begin{bmatrix} \Lambda & \Lambda X^T \\ X\Lambda & X\Lambda X^T + \Sigma \end{bmatrix}$$

$\hat{\beta}_{\rm mmse}$ is essentially the conditional mean (readily computed for Gaussian)

$$\hat{\beta}_{\rm mmse} = \mathbf{E}[\beta|y] = C_{\beta y} C_{yy}^{-1} y = \Lambda X^T (X\Lambda X^T + \Sigma)^{-1} y$$

alternatively, we claim that $\mathbf{E}[\beta|y]$ is linear in y (because β, y are Gaussian)

$$\hat{\beta}_{\rm mmse} = \hat{\beta}_{\rm lms} = Ky$$

and K can be computed from the Wiener-Hopf equation

Maximum a posteriori:

- β is random and independent of e
- $\beta \sim \mathcal{N}(0, \Lambda)$

the MAP estimate can be found by solving

$$\hat{\theta}_{\mathrm{map}} = \underset{\beta}{\mathrm{argmax}} \quad \log f(\beta | y) = \underset{\beta}{\mathrm{argmax}} \quad \log f(y|\beta) + \log f(\beta)$$

without having to solve this problem, it is immediate that

$$\hat{\beta}_{\mathrm{map}} = \hat{\beta}_{\mathrm{mmse}}$$

since for Gaussian density function, $\mathbf{E}[\beta|y]$ maximizes $f(\beta|y)$

nevertheless, we can write down the posteriori density function (up to a constant)

$$\log f(y|\beta) = -(1/2)\log\det\Sigma - (1/2)(y - X\beta)^T\Sigma^{-1}(y - X\beta)$$
$$\log f(\beta) = -(1/2)\log\det\Lambda - (1/2)\beta^T\Lambda^{-1}\beta$$

the MAP estimate satisfies the zero gradient (w.r.t. β) condition:

$$-X^T\Sigma^{-1}(y - X\beta) + \Lambda^{-1}\beta = 0$$

that gives the form similar to MLE except the extra term Λ^{-1}

$$\hat{\beta}_{\mathrm{map}} = (X^T\Sigma^{-1}X + \Lambda^{-1})^{-1}X^T\Sigma^{-1}y$$

when $\Lambda = \infty$ or *maximum ignorance*, it reduces to ML estimate

it is a fact that $\hat{\beta}_{\mathrm{mmse}} = \hat{\beta}_{\mathrm{map}}$, so it is interesting to verify

$$\Lambda X^T(X\Lambda X^T + \Sigma)^{-1}y = (X^T\Sigma^{-1}X + \Lambda^{-1})^{-1}X^T\Sigma^{-1}y$$

(the two terms are equivalent – proved by some algebraic operations)

proof: $\hat{\beta}_{\rm mmse} = \hat{\beta}_{\rm map}$

define $H = (X\Lambda X^T + \Sigma)^{-1}y$ and we have

$$X\Lambda X^T H + \Sigma H = y$$

we start with the expression of $\hat{\beta}_{\rm mmse}$

$$\begin{aligned}
\hat{\beta}_{\rm mmse} &= \Lambda X^T(X\Lambda X^T + \Sigma)^{-1}y = \Lambda X^T H \\
X\hat{\beta}_{\rm mmse} &= X\Lambda X^T H = y - \Sigma H \\
\Lambda X^T\Sigma^{-1}X\beta_{\rm mmse} &= \Lambda X^T\Sigma^{-1}y - \Lambda X^T H \\
&= \Lambda X^T\Sigma^{-1}y - \hat{\beta}_{\rm mmse} \\
(I + \Lambda X^T\Sigma^{-1}X)\hat{\beta}_{\rm mmse} &= \Lambda X^T\Sigma^{-1}y \\
(\Lambda^{-1} + X^T\Sigma^{-1}X)\hat{\beta}_{\rm mmse} &= X^T\Sigma^{-1}y \\
\hat{\beta}_{\rm mmse} &= (\Lambda^{-1} + X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y \triangleq \hat{\beta}_{\rm map}
\end{aligned}$$

covariance of MAP estimate: we use $\hat{\beta}_{\rm map} = \mathbf{E}[\beta|y]$

$$\mathbf{cov}(\hat{\beta}_{\rm map}) = \mathbf{E}\left[(\beta - \mathbf{E}[\beta|y])(\beta - \mathbf{E}[\beta|y])^T\right]$$

use the fact that the optimal residual is uncorrelated with y

$$\mathbf{cov}(\hat{\beta}_{\rm map}) = \mathbf{E}\left[(\beta - \mathbf{E}[\beta|y])\beta^T\right]$$

next, use the fact that $\hat{\beta}_{\rm map} = \mathbf{E}[\beta|y]$ is a linear function in y

$$\begin{aligned}\mathbf{cov}(\hat{\beta}_{\rm map}) &= C_\beta - KC_{y\beta} = \Lambda - (X^T\Sigma^{-1}X + \Lambda^{-1})^{-1}X^T\Sigma^{-1}X\Lambda \\ &= (X^T\Sigma^{-1}X + \Lambda^{-1})^{-1}\left[(X^T\Sigma^{-1}X + \Lambda^{-1})\Lambda - X^T\Sigma^{-1}X\Lambda\right] \\ &= (X^T\Sigma^{-1}X + \Lambda^{-1})^{-1} \preceq (X^T\Sigma^{-1}X)^{-1}\end{aligned}$$

$\hat{\beta}_{\rm map}$ yields a smaller covariance matrix than that of $\hat{\beta}_{\rm ml}$
(because ML does not use a prior knowledge about β)

Summary

- estimate methods in this section require statistical properties of random entities in the model
- minimum-mean-square estimate is the conditional mean and typically a nonlinear function in the measurement data
- a maximum-likelihood estimation is a nonlinear optimization problem; it can reduce to have a closed-form solution in some special case of noise distribution (e.g. Gaussian)
- a maximum a posteriori estimation takes model parameters as random variables; it requires a prior distribution of these parameters

References

Appendix B in
T. Söderström and P. Stoica, *System Identification*, Prentice Hall, 1989

Chapter 2-3 in
T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall, 2000

Chapter 9 in
A. V. Balakrishnan, *Introduction to Random Processes in Engineering*, John Wiley & Sons, Inc., 1995

Chapter 7 in
S. Boyd, and L. Vandenberghe, *Convex Optimization*, Cambridge press, 2004