11. Statistical Estimation

- conditional expectation
- mean square estimation (MSE)
- maximum likelihood estimation (ML)
- maximum a posteriori estimation (MAP)

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 $\mathbf{E}[|g(y)|] < \infty$,

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

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: $(x, y) \sim \mathcal{N}(\mu, C)$ where

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \quad C = \begin{bmatrix} C_x & C_{xy} \\ C_{xy}^T & C_y \end{bmatrix}$$

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

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

and conditional covariance matrix

$$C_{x|y} = C_x - C_{xy} C_y^{-1} C_{xy}^T$$

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

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

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**:

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

  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

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

- 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 is

$$C_{xy} = KC_y$$

and is called the **Wiener-Hopf** equation

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

$$\text{MSE} = \mathbf{E}\,\mathbf{tr}(\tilde{x} - K\tilde{y})(\tilde{x} - K\tilde{y})^T = \mathbf{tr}(C_x - C_{xy}K^T - KC_{yx} + KC_yK^T)$$

- 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)

Minimizing the error covariance matrix

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

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

the problem is to choose $h(y)$ to yield the minimum covariance matrix
(instead of minimizing the mean square norm)

we compare two matrices by

$$M \preceq N \quad \text{if} \quad M - N \preceq 0$$

or $M - N$ is nonpositive definite

now restrict to the linear case:

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

the covariance matrix can be written as

$$(\mu_x - (K\mu_y + c))(\mu_x - (K\mu_y + c))^T + C_x - KC_{yx} - C_{xy}K^T + KC_yK^T$$

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 is reduced to

$$f(K) = C_x - KC_{yx} - C_{xy}K^T + KC_yK^T$$

note that $f(K) \succeq 0$ because we can write $f(K)$ as

$$f(K) = \begin{bmatrix} -I & K \end{bmatrix} \begin{bmatrix} C_x & C_{xy} \\ C_{xy}^T & C_y \end{bmatrix} \begin{bmatrix} -I \\ K^T \end{bmatrix}$$

let K_0 be a solution to the Wiener-Hopf equation: $C_{xy} = K_0 C_y$

we can verify that

$$f(K) = f(K_0) + (K - K_0) C_y (K - K_0)^T$$

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

the miminum covariance matrix is

$$f(K_0) = C_x - C_{xy} C_y^{-1} C_{xy}^T$$

for $C = \begin{bmatrix} C_x & C_{xy} \\ C_{xy}^T & C_y \end{bmatrix}$, note that

- the minimum covariance matrix is the Schur complement of C_x in C
- it is exactly a conditional covariance matrix for Gaussian variables

Maximum likelihood estimation

- $y = (y_1, \ldots, y_m)$: the observations of random variables
- θ: unknown parameters to be estimated
- $f(y|\theta)$: the probability density function of y for a fixed θ

in ML estimation, we assume θ are **fixed** (and deterministic) parameters

to estimate θ from y, we maximize the density function for a given θ:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; f(y|\theta)$$

- $f(y|\theta)$ is called the **likelihood function**
- θ is chosen so that the observed y becomes "as likely as possible"

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

- observe a sequence of *independent* random variables: $y_1, y_2, \ldots, y_m$
- 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

$$f(y_1, y_2, \ldots, y_m | \mu, \Sigma) = \frac{1}{(2\pi)^{mn/2}} \cdot \frac{1}{|\Sigma|^{m/2}} \cdot \mathbf{exp} - \frac{1}{2} \sum_{k=1}^{m} (y_k - \mu)^T \Sigma^{-1} (y_k - \mu)$$

- to maximize f, it is convenient to consider the **log-likelihood function**: (up to a constant)

$$L(\mu, \Sigma) = \log f = \frac{m}{2} \log \det \Sigma^{-1} - \frac{1}{2} \sum_{k=1}^{m} (y_k - \mu)^T \Sigma^{-1} (y_k - \mu)$$

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

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

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

- we obtain the ML estimate of μ, Σ as

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

  - $\hat{\mu}_{\mathrm{ml}}$ is the sample mean
  - $\hat{\Sigma}_{\mathrm{ml}}$ is a (biased) sample covariance matrix

Example 2: linear measurements with i.i.d. noise

consider a linear measurement model

$$y = A\theta + v$$

$\theta \in \mathbf{R}^n$	is parameter to be estimated
$y \in \mathbf{R}^m$	is the measurement
$v \in \mathbf{R}^m$	is i.i.d. noise

(v_i are independent, identically distributed) with density f_v

the density function of $y - A\theta$ is therefore the same as v:

$$f(y|\theta) = \prod_{k=1}^{m} f_v(y_k - a_k^T\theta)$$

where a_k^T are the row vectors of A

the ML estimate of θ depends on the noise distribution f_v

suppose v_k is Gaussian with zero mean and variance σ

- the log-likelihood function is

$$L(\theta) = \log f = -(m/2)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{m}(y_k - a_k^T\theta)^2$$

  (a_k^T are row vectors of A)

- therefore the ML estimate of θ is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ \|y - A\theta\|_2^2$$

- the solution of a least-squares problem

what about other distributions of v_k ?

Maximum a posteriori (MAP) estimation

assumptions:

- assume that θ is a *random variable*
- θ and y has a joint distribution $f(y, \theta)$

the MAP estimate of θ is given by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, f_{\theta|y}(\theta|y)$$

- $f_{\theta|y}$ is called the **posterior** density of θ
- $f_{\theta|y}$ represents our knowledge of θ after we observe y
- MAP estimate is the value that maximizes the conditional density of θ, given the observed y

from Bayes rule, the MAP estimate is also obtained by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \quad f_{y|\theta}(y|\theta) f_{\theta}(\theta)$$

taking logarithms, we can express $\hat{\theta}$ as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \quad \log f_{y|\theta}(y|\theta) + \log f_{\theta}(\theta)$$

- the only difference between ML and MAP estimate is the term $f_{\theta}(\theta)$
- f_{θ} is called the **prior** density, representing prior knowledge about θ
- $\log f_{\theta}(\theta)$ penalizes choices of θ that are unlikely to happen

under what condition on f_{θ} is the MAP estimate identical to the ML estimate ?

Example 3: linear measurement with IID noise

use the model in page 11-14 and assume θ has a prior density f_θ on $\mathbf{R}^n$

the MAP estimate can be found by solving

$$\text{maximize} \quad \log f_\theta(\theta) + \sum_{k=1}^{m} \log f_v(y_k - a_k^T \theta)$$

suppose $\theta \sim \mathcal{N}(0, \beta I)$ and $v_k \sim \mathcal{N}(0, \sigma)$, the MAP estimation is

$$\text{maximize} \quad -\frac{1}{\beta}\|\theta\|_2^2 - \frac{1}{\sigma^2}\|A\theta - y\|_2^2$$

conclusion: MAP estimate with a *Guassian prior* is the solution to a least-squares problem with ℓ_2 regularization

what if θ has a Laplacian distribution ?

Cramér-Rao inequality

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

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

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

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

- $f(y|\theta)$ is the density function of observations y for a given θ
- the RHS is called the **Cramér-Rao** lower bound
- provide the minimal covariance matrix over all possible estimators $\hat{\theta}$
- $J \triangleq \mathbf{E}\nabla_\theta^2 \log f(y|\theta)$ is called the **Fisher information matrix**
- an estimator for which the C-R equality holds is called **efficient**

Proof of the Cramér-Rao inequality

since $f(y|\theta)$ is a density function and $\hat{\theta}$ is unbiased, we have

$$1 = \int f(y|\theta)dy, \quad \theta = \int \hat{\theta}(y) f(y|\theta)dy$$

differentiate the eqs w.r.t. θ and use $\nabla_\theta \log f(y|\theta) = \dfrac{\nabla_\theta f(y|\theta)}{f(y|\theta)}$

$$0 = \int \nabla_\theta \log f(y|\theta) f(y|\theta)dy, \quad I = \int \hat{\theta}(y) \nabla_\theta \log f(y|\theta) f(y|\theta)dy$$

these two identities can be expressed as

$$\mathbf{E}\left[(\hat{\theta}(y) - \theta)\nabla_\theta \log f(y|\theta)\right] = I$$

($\mathbf{E}$ is taken w.r.t y, and θ is fixed)

consider a positive semidefinite matrix

$$\mathbf{E}\begin{bmatrix}\hat{\theta}(y)-\theta\\(\nabla_\theta \log f(y|\theta))^T\end{bmatrix}\begin{bmatrix}\hat{\theta}(y)-\theta\\(\nabla_\theta \log f(y|\theta))^T\end{bmatrix}^T \succeq 0$$

expand the product into the form

$$\begin{bmatrix}A & I\\ I & D\end{bmatrix}$$

where $A = \mathbf{E}(\hat{\theta}(y)-\theta)(\hat{\theta}(y)-\theta)^T$ and

$$D = \mathbf{E}(\nabla_\theta \log f(y|\theta))^*(\nabla_\theta \log f(y|\theta))$$

the Schur complement of the $(1,1)$ block must be nonnegative:

$$A - ID^{-1}I \succeq 0$$

which implies the Cramér Rao inequality

now it remains to show that

$$\mathbf{E}(\nabla_\theta \log f(y|\theta))^T(\nabla_\theta \log f(y|\theta)) = -\mathbf{E}\nabla_\theta^2 \log f(y|\theta)$$

from the equation

$$0 = \int \nabla_\theta \log f(y|\theta) f(y|\theta) dy,$$

differentiating on both sides gives

$$0 = \int \nabla_\theta^2 \log f(y|\theta) f(y|\theta) dy + \int \nabla_\theta \log f(y|\theta)^T \nabla_\theta \log f(y|\theta) f(y|\theta) dy$$

or

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

Example of computing the Cramér Rao bound

revisit a linear model with correlated Gaussian noise:

$$y = A\theta + v, \quad v \sim \mathcal{N}(0, \Sigma), \quad \Sigma \text{ is known}$$

the density function $f(y|\theta)$ is given by $f_v(y - A\theta)$ which is Gaussian

$$\begin{aligned} \log f(y|\theta) &= -\frac{1}{2}(y - A\theta)^T \Sigma^{-1}(y - A\theta) - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma \\ \nabla_\theta \log f(y|\theta) &= A^T\Sigma^{-1}(y - A\theta) \\ \nabla^2_\theta \log f(y|\theta) &= -A^T\Sigma^{-1}A \end{aligned}$$

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

$$\mathbf{cov}(\hat{\theta}) \succeq (A^T\Sigma^{-1}A)^{-1}$$

Linear models with additive noise

estimate parameters in a linear model with addtive noise:

$$y = A\theta + v, \quad v \sim \mathcal{N}(0, \Sigma), \quad \Sigma \text{ is known}$$

and we explore several estimates from the following approaches

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

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

Least-squares: $\hat{\theta}_{\rm ls} = (A^T A)^{-1} A^T y$ and is unbiased

$$\mathbf{cov}(\hat{\theta}_{\rm ls}) = \mathbf{cov}((A^T A)^{-1} A^T v) = (A^T A)^{-1} A^T \Sigma A (A^T A)^{-1}$$

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

(the error covariance matrix is bigger than the CR bound)

however the bound is tight when the noise covariance is diagonal:

$$\Sigma = \sigma^2 I$$

(the noise v_k are uncorrelated)

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

$$\hat{\theta}_{\rm wls} = (A^T W A)^{-1} A^T W y, \quad \text{and is unbiased}$$

$$\begin{aligned} \mathbf{cov}(\hat{\theta}_{\rm wls}) &= \mathbf{cov}((A^T W A)^{-1} A^T W v) \\ &= (A^T W A)^{-1} A^T W \Sigma W A (A^T W A)^{-1} \end{aligned}$$

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

$$\hat{\theta}_{\rm wls} = (A^T \Sigma^{-1} A)^{-1} A^T \Sigma^{-1} y$$

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 $f(y|\theta) = f_v(y - A\theta)$,

$$\log f(y|\theta) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma - \frac{1}{2}(y - A\theta)^T\Sigma^{-1}(y - A\theta)$$

the zero gradient condition gives

$$\nabla_\theta \log f(y|\theta) = A^T\Sigma^{-1}(y - A\theta) = 0$$

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

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

moreover, we can verify that

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

minimum mean square estimate:

- θ is random and independent of v
- $\theta \sim \mathcal{N}(0, \Lambda)$

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

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

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

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

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

$$\hat{\theta}_{\text{mmse}} = \hat{\theta}_{\text{lms}} = Ky$$

and K can be computed from the Wiener-Hopf equation

Maximum a posteriori:

- θ is random and independent of v
- $\theta \sim \mathcal{N}(0, \Lambda)$

the MAP estimate can be found by solving

$$\hat{\theta}_{\text{map}} = \underset{\theta}{\text{argmax}} \;\; \log f(\theta|y) = \underset{\theta}{\text{argmax}} \;\; \log f(y|\theta) + \log f(\theta)$$

without having to solve this problem, it is immediate that

$$\hat{\theta}_{\text{map}} = \hat{\theta}_{\text{mmse}}$$

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

nevertheless, we can write down the posteriori density function

$$\log f(y|\theta) = -\frac{1}{2}\log\det\Sigma - \frac{1}{2}(y - A\theta)^T\Sigma^{-1}(y - A\theta)$$
$$\log f(\theta) = -\frac{1}{2}\log\det\Lambda - \frac{1}{2}\theta^T\Lambda^{-1}\theta$$

(these terms are up to a constant)

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

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

which gives

$$\hat{\theta}_{\rm map} = (A^T\Sigma^{-1}A + \Lambda^{-1})^{-1}A^T\Sigma^{-1}y$$

- $\hat{\theta}_{\rm map}$ is clearly similar to $\hat{\theta}_{\rm ml}$ except the extra term Λ^{-1}
- when $\Lambda = \infty$ or *maximum ignorance*, it reduces to ML estimate

- from $\hat{\theta}_{\mathrm{mmse}} = \hat{\theta}_{\mathrm{map}}$, it is interesting to verify

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

(see the proof next page - it can be skipped)

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

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

we start with the expression of $\hat{\theta}_{\text{lms}}$

$$\begin{aligned}
\hat{\theta}_{\text{mmse}} &= \Lambda A^T (A\Lambda A^T + \Sigma)^{-1} y = \Lambda A^T H \\
A\hat{\theta}_{\text{mmse}} &= A\Lambda A^T H = y - \Sigma H \\
\Lambda A^T \Sigma^{-1} A\theta_{\text{mmse}} &= \Lambda A^T \Sigma^{-1} y - \Lambda A^T H \\
&= \Lambda A^T \Sigma^{-1} y - \hat{\theta}_{\text{mmse}} \\
(I + \Lambda A^T \Sigma^{-1} A)\hat{\theta}_{\text{mmse}} &= \Lambda A^T \Sigma^{-1} y \\
(\Lambda^{-1} + A^T \Sigma^{-1} A)\hat{\theta}_{\text{mmse}} &= A^T \Sigma^{-1} y \\
\hat{\theta}_{\text{mmse}} &= (\Lambda^{-1} + A^T \Sigma^{-1} A)^{-1} A^T \Sigma^{-1} y \triangleq \hat{\theta}_{\text{map}}
\end{aligned}$$

to compute the covariance matrix of the error, we use $\hat{\theta}_{\mathrm{map}} = \mathbf{E}[\theta|y]$

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

use the fact that the optimal residual is uncorrelated with y

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

next $\hat{\theta}_{\mathrm{map}} = \mathbf{E}[\theta|y]$ is a linear function in y

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

$\hat{\theta}_{\mathrm{map}}$ yields a smaller covariance matrix than $\hat{\theta}_{\mathrm{ml}}$ as it should be

(ML does not use a prior knowledge about θ)

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