9. Statistical Estimation

- Conditional expectation
- Mean square estimation
- Maximum likelihood estimation
- Maximum a posteriori estimation

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

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

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

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

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

Best linear unbiased estimate Now we restrict $h(y)$ to be linear:

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

In order $h(y)$ to be unbiased, we must have

$$c = \mathbf{E}[x] - K\mathbf{E}[y]$$

Define $\tilde{x} = x - \mathbf{E}[x]$ and $\tilde{y} = y - \mathbf{E}[y]$

$h(y)$ is then of the form

$$h(y) = K\tilde{y} + \mathbf{E}[x]$$

The mean square error becomes

$$\begin{aligned}\mathbf{E}\|x - h(y)\|^2 = \mathbf{E}\|\tilde{x} - K\tilde{y}\|^2 &= \mathbf{E}\,\mathbf{tr}(\tilde{x} - K\tilde{y})(\tilde{x} - K\tilde{y})^* \\ &= \mathbf{tr}(C_x - C_{xy}K^* - KC_{yx} + KC_yK^*)\end{aligned}$$

where C_x, C_y, C_{xy} are the covariance matrices

Differentiating the objective w.r.t. K gives

$$C_{xy} = KC_y$$

This equation is referred as the **Wiener-Hopf** equation

Also obtain from the condition

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

(the optimal residual is uncorrelated with the observation y)

If C_y is nonsingular, then $K = C_{xy}C_y^{-1}$

The best unbiased linear estimate is

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

It coincides with the optimal mean square estimate for Gaussian RVs

Minimizing the error covariance matrix

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

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

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

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

The objective is minimized w.r.t 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

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

Note that $f(K) \succeq 0$ because

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

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

We can verify that

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

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

The miminum covariance matrix is

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

Note that suppose $C = \begin{bmatrix} C_x & C_{xy} \\ C_{xy}^* & C_y \end{bmatrix}$

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

Maximum likelihood estimation

- $y = (y_1, \ldots, y_N)$: 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 as *fixed* 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 variables

Observe a sequence of independent random variables:

$$y_1, y_2, \ldots, y_N$$

Each y_k is multivariate 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

$$f(y_1, y_2, \ldots, y_m | \mu, \Sigma)$$

$$= \frac{1}{(2\pi)^{N/2}} \cdot \frac{1}{|\Sigma|^{N/2}} \cdot \mathbf{exp} - \frac{1}{2} \sum_{k=1}^{N} (y_k - \mu)^* \Sigma^{-1} (y_k - \mu)$$

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

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

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

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

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

We obtain the ML estimate of μ, Σ as

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} y_k, \quad \Sigma = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{\mu})(y_k - \hat{\mu})^*$$

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

Example 2 Linear measurements with IID 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 IID 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 are the columns 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 loglikelihood function is

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

Therefore the ML estimate of θ is

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

The solution of a least-squares problem

what about other distributions of v_k ?

Maximum a posteriori (MAP) estimation

Assume that θ is a *random variable*

θ and y has a joint distribution $f(y, \theta)$

In the MAP estimation, our 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
- The MAP estimate is the value that maximizes the conditional density of θ, give 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: Linear measurement with IID noise

Use the model in page 9-13 and θ has 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

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

The 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:

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

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

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

- $f(y|\theta)$ 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 equality holds is called *efficient*

Proof of the Cramér-Rao inequality

As $f(y|\theta)$ 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 of the above equations and use $\nabla_\theta \log f(y|\theta) = \frac{\nabla_\theta f(y|\theta)}{f(y|\theta)}$

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

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

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

Expand the product and this matrix is of the form

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

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

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

Use the fact that its Schur complement of the $(1,1)$ block must be nonnegative:

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

This implies the Cramér Rao inequality

Now it remains to show that

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

From the equation

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

differentiation on both sides gives

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

or

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

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

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

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

$$\nabla_\theta^2 \log f(y|\theta) = -A^* \Sigma^{-1} A$$

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

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

Linear models with additive noise

We estimate parameters in a linear model with addtive noise:

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

We explore several estimates from the following approaches

- do not use information about the noise
  - Least-squares estimate (LS)
- use information about the noise (Guassian distribution, Σ)
  - Assume θ is a fixed parameter
    - Weighted least-squares estimate (WLS)
    - Best linear unbiased estimate (BLUE)
    - Maximum likelihood estimate (ML)
  - Assume θ is random and $\theta \sim \mathcal{N}(0, \Lambda)$
    - Least mean square estimate (LMS)
    - Maximum a posteriori estimate (MAP)

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

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

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

(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 $W \succ 0$

$$\hat{\theta}_{\rm wls} = (A^*WA)^{-1}A^*Wy, \quad \text{and is unbiased}$$

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

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

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

We also have seen that in this case

$$\hat{\theta}_{\rm blue} = \hat{\theta}_{\rm wls}$$

(for Gaussian, the best linear estimate is also the best among nonlinear estimates)

Maximum likelihood

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

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

The zero gradient condition gives

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

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

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

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

Least mean square estimate Assume θ is random and independent of v

Moreover, we assume $\theta \sim \mathcal{N}(0, \Lambda)$

Hence, θ and y are jointly Gaussian with zero mean and the covariance matrix

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

$\hat{\theta}_{\rm lms}$ is essentially the conditional mean which can be computed readily for Gaussian distribution

$$\begin{aligned} \hat{\theta}_{\rm lms} &= \mathbf{E}[\theta|y] = C_{\theta y} C_{yy}^{-1} y \\ &= \Lambda A^* (A\Lambda A^* + \Sigma)^{-1} y \end{aligned}$$

Alternatively, we can claim that $\mathbf{E}[\theta|y]$ is a linear function of y (because θ, y are Gaussian)

$$\hat{\theta}_{\rm lms} = Ky$$

and K can be computed from the Wiener-Hopf equation

Maximum a posteriori Assume θ is random and independent of v and assume $\theta \sim \mathcal{N}(0, \Lambda)$

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

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

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

Nevertheless, we can write down the posteriori density function

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

$$\log f(\theta) = -\frac{1}{2} \log \det \Lambda - \frac{1}{2} \theta^* \Lambda \theta$$

(these terms are up to a constant)

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

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

which gives

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

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

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

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

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

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

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

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

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

Use the fact that the optimal residual is uncorrelated with y

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

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

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

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

(ML does not use a prior knowledge about θ)

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