Solutions To Problems of Chapter 3

3.1. Let $\hat{\theta}_i$, $i = 1, 2, \ldots, m$, be unbiased estimators of a parameter vector $\boldsymbol{\theta}$, i.e., $\mathbb{E}[\hat{\theta}_i] = \boldsymbol{\theta}$, $i = 1, \ldots, m$. Moreover, assume that the respective estimators are uncorrelated to each other and that all have the same (total) variance, $\sigma^2 = \mathbb{E}[(\theta_i - \boldsymbol{\theta})^T(\theta_i - \boldsymbol{\theta})]$. Show that by averaging the estimates, e.g.,

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i,$$

the new estimator has total variance $\sigma_c^2 := \mathbb{E}[(\hat{\theta} - \boldsymbol{\theta})^T(\hat{\theta} - \boldsymbol{\theta})] = \frac{1}{m}\sigma^2$.

Solution: First, it is easily checked out that the new estimator is also unbiased. By the definition of the total variance (which is the trace of the respective covariance matrix), we have

$$\begin{aligned}
\sigma_c^2 &= \mathbb{E}[(\hat{\theta} - \boldsymbol{\theta})^T(\hat{\theta} - \boldsymbol{\theta})] \\
&= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\theta}_i - \boldsymbol{\theta}\right)\right)^T \left(\frac{1}{m}\sum_{j=1}^{m}\left(\hat{\theta}_j - \boldsymbol{\theta}\right)\right)\right] \\
&= \frac{1}{m^2}\sum_{i,j=1}^{m} \mathbb{E}\left[\left(\hat{\theta}_i - \boldsymbol{\theta}\right)^T \left(\hat{\theta}_j - \boldsymbol{\theta}\right)\right] = \frac{1}{m}\sigma^2,
\end{aligned}$$

since $\mathbb{E}[(\hat{\theta}_i - \boldsymbol{\theta})^T(\hat{\theta}_j - \boldsymbol{\theta})] = \delta_{ij}\sigma^2$.

3.2. Let a random variable x being described by a uniform pdf in the interval $[0, \frac{1}{\theta}]$, $\theta > 0$. Assume a function[1] g, which defines an estimator $\hat{\theta} := g(\mathrm{x})$ of θ. Then, for such an estimator to be unbiased, the following must hold:

$$\int_0^{\frac{1}{\theta}} g(x)dx = 1.$$

However, such a function g does not exist.

Solution: Necessarily, the pdf of x must be

$$p(x) = \begin{cases} \theta, & x \in [0, \frac{1}{\theta}], \\ 0, & \text{otherwise.} \end{cases}$$

For the estimator to be unbiased,

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \int_{-\infty}^{\infty} g(x)p(x)dx \\
&= \int_0^{\frac{1}{\theta}} g(x)\theta dx = \theta, \quad \forall\theta > 0.
\end{aligned}$$

[1] To avoid any confusion, let g be Lebesgue integrable on intervals of $\mathbb{R}$.

Hence,

$$G(\theta) := \int_0^{\frac{1}{\theta}} g(x)dx = 1, \quad \forall\theta > 0. \tag{1}$$

However, such a function g cannot exist. Indeed, one can easily verify by the basic integral theory that $\lim_{\theta\to\infty} G(\theta) = 0$, and $\lim_{\theta\to 0} G(\theta) = 1$, by (1). Clearly, these results contradict each other.

3.3. A family $\{p(\mathcal{D};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{A}\}$ is called *complete* if, for any vector function $\boldsymbol{h}(\mathcal{D})$ such that $\mathbb{E}_{\mathcal{D}}[\boldsymbol{h}(\mathcal{D})] = \mathbf{0}$, $\forall\boldsymbol{\theta}$, then $\boldsymbol{h} = \mathbf{0}$.

Show that if $\{p(\mathcal{D};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{A}\}$ is complete, and there exists an MVU estimator, then this estimator is unique.

Solution: Assume two MVU estimators $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. Then, $\mathbb{E}_{\mathcal{D}}[\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2] = \boldsymbol{\theta} - \boldsymbol{\theta} = \mathbf{0}$. Hence, by the definition of completeness, $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

3.4. Let $\hat{\theta}_u$ be an unbiased estimator, i.e., $\mathbb{E}[\hat{\theta}_u] = \theta_0$. Define a biased one by $\hat{\theta}_b = (1+\alpha)\hat{\theta}_u$. Show that the range of α where the MSE of $\hat{\theta}_b$ is smaller than that of $\hat{\theta}_u$ is

$$-2 < -\frac{2\text{MSE}(\hat{\theta}_u)}{\text{MSE}(\hat{\theta}_u) + \theta_0^2} < \alpha < 0.$$

Solution: The MSE for the new estimator is

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta}_b - \theta_0)^2\right] &= \mathbb{E}\left[\Big((1+\alpha)\hat{\theta}_u - \theta_0\Big)^2\right] \\
&= \mathbb{E}\left[\Big((1+\alpha)(\hat{\theta}_u - \theta_0) + \alpha\theta_0\Big)^2\right] \\
&= (1+\alpha)^2\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_0^2.
\end{aligned}$$

To obtain smaller MSE for the unbiased estimator we must have

$$(1+\alpha)^2\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_0^2 < \text{MSE}(\hat{\theta}_u),$$

or, after using elementary algebra,

$$\alpha\left[\alpha + \frac{2\,\text{var}(\hat{\theta}_u)}{\theta_0^2 + \text{var}(\hat{\theta}_u)}\right] < 0,$$

where $\text{var}(\cdot)$ denotes the variance, and clearly $\text{MSE}(\hat{\theta}_u) = \text{var}(\hat{\theta}_u)$. The solution of the previous inequality results to the desired interval:

$$-\frac{2\,\text{var}(\hat{\theta}_u)}{\theta_0^2 + \text{var}(\hat{\theta}_u)} < \alpha < 0.$$

3.5. Show that for the setting of the Problem 3.4, the optimal value of α is equal to

$$\alpha_* = -\frac{1}{1+\frac{\theta_0^2}{\text{var}(\hat{\theta}_u)}},$$

where, of course, the variance of the unbiased estimator is equal to the corresponding MSE.

Solution: The minimum value of

$$\text{MSE}(\hat{\theta}_b) = \mathbb{E}\left[(\hat{\theta}_b - \theta_0)^2\right] = (1+\alpha)^2\text{MSE}(\hat{\theta}_u) + \alpha^2\theta_0^2,$$

with respect to α occurs when the derivative becomes zero, that is when

$$2(1+\alpha)\,\text{var}(\hat{\theta}_u) + 2\alpha\theta_0^2 = 0,$$

or, equivalently, when

$$\alpha_* = -\frac{\text{var}(\hat{\theta}_u)}{\theta_0^2 + \text{var}(\hat{\theta}_u)} = -\frac{1}{1+\frac{\theta_0^2}{\text{var}(\hat{\theta}_u)}}.$$

3.6. Show that the regularity condition for the Cramér-Rao bound holds true if the order of integration and differentiation can be interchanged.

Solution: By the definition of the expectation we have

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \ln p(\mathcal{X};\theta)}{\partial\theta}\right] &= \int\cdots\int p(\mathcal{X};\theta)\frac{\partial \ln p(\mathcal{X};\theta)}{\partial\theta}d\boldsymbol{x}_1\cdots d\boldsymbol{x}_N \\
&= \int\cdots\int \frac{\partial p(\mathcal{X};\theta)}{\partial\theta}d\boldsymbol{x}_1\cdots d\boldsymbol{x}_N \\
&= \frac{\partial}{\partial\theta}\int\cdots\int p(\mathcal{X};\theta)d\boldsymbol{x}_1\cdots d\boldsymbol{x}_N \\
&= \frac{\partial 1}{\partial\theta} = 0.
\end{aligned}$$

This is in general true, unless the domain where the pdf is nonzero depends on the unknown parameter θ.

3.7. Derive the Cramér-Rao bound for the LS estimator, when the training data result from the linear model

$$y_n = \theta x_n + \eta_n, \quad n = 1, 2, \ldots,$$

where x_n and η_n are observations of i.i.d random variables, drawn from a zero mean random process, with variance σ_x^2, and a Gaussian white noise process, with zero mean and variance σ_η^2, respectively. Assume, also, that x and η are independent. Then, show that the LS estimator achieves the CR bound only asymptotically.

Solution: First, notice that in this case $\mathcal{X} = \{(x_n, y_n)\}_{n=1}^N$. That is, both y_n as well as x_n change as we change the training set. Define here the quantities $\boldsymbol{x}_N := [x_1, x_2, \dots, x_N]^T$, $\boldsymbol{y}_N := [y_1, y_2, \dots, y_N]^T$, and recall, also, the elementary relations

$$\begin{aligned} p(\mathcal{X};\theta) &= p(\boldsymbol{x}_N, \boldsymbol{y}_N;\theta) \\ &= p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta)p(\boldsymbol{x}_N;\theta) \\ &= p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta)p(\boldsymbol{x}_N). \end{aligned}$$

Hence, $\ln p(\mathcal{X};\theta) = \ln p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta) + \ln p(\boldsymbol{x}_N)$, and eventually,

$$\frac{\partial \ln p(\mathcal{X};\theta)}{\partial\theta} = \frac{\partial \ln p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta)}{\partial\theta}. \tag{2}$$

Now, notice that by our original assumptions on the data model,

$$p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta) = \frac{1}{\left(2\pi\sigma_\eta^2\right)^{N/2}} \exp\left(-\frac{1}{2\sigma_\eta^2}\sum_{n=1}^N (y_n - \theta x_n)^2\right),$$

or

$$\ln p(\boldsymbol{y}_N|\boldsymbol{x}_N;\theta) = -\frac{N}{2}\ln\left(2\pi\sigma_\eta^2\right) - \frac{1}{2\sigma_\eta^2}\sum_{n=1}^N (y_n - \theta x_n)^2.$$

Thus, by (2),

$$\frac{\partial \ln p(\mathcal{X};\theta)}{\partial\theta} = \frac{1}{\sigma_\eta^2}\sum_{n=1}^N (y_n - \theta x_n)\, x_n = \frac{1}{\sigma_\eta^2}\sum_{n=1}^N \eta_n x_n. \tag{3}$$

It can be readily verified by (3) that the regularity condition of the Cramér-Rao Theorem is satisfied.

Now,

$$\frac{\partial^2 \ln p(\mathcal{X};\theta)}{\partial\theta^2} = -\frac{1}{\sigma_\eta^2}\sum_{n=1}^N x_n^2.$$

Therefore

$$\mathbb{E}\left[\frac{\partial^2 \ln p(\mathcal{X};\theta)}{\partial\theta^2}\right] = -N\frac{\sigma_x^2}{\sigma_\eta^2},$$

and the Cramér-Rao bound is given by

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{N}\frac{\sigma_\eta^2}{\sigma_x^2}.$$

We will now show that this bound cannot be achieved by any unbiased estimator. The necessary and sufficient condition for the existence of an

MVU estimator that achieves the Cramér-Rao bound translates, for this case, to the existence of a function $g(\mathcal{X})$ such that

$$\frac{\partial \ln p(\mathcal{X};\theta)}{\partial \theta} = N\frac{\sigma_x^2}{\sigma_\eta^2}(g(\mathcal{X}) - \theta),$$

However, looking at (3), it becomes apparent that such a factorization is not possible.

Let us now rewrite (3) as

$$\begin{aligned}\frac{\partial \ln p(\mathcal{X};\theta)}{\partial \theta} &= N\frac{\sigma_x^2}{\sigma_\eta^2}\Big(\frac{1}{N\sigma_x^2}\sum_{n=1}^{N}\big(y_n x_n - \theta x_n^2\big)\Big)\\ & N\frac{\sigma_x^2}{\sigma_\eta^2}\Big(\frac{1}{N\sigma_x^2}\sum_{n=1}^{N} y_n x_n - \theta(\frac{1}{N\sigma_x^2}\sum_{n=1}^{N} x_n^2)\Big)\\ & N\frac{\sigma_x^2}{\sigma_\eta^2}\Big(g(\mathcal{X}) - \theta(\frac{1}{N\sigma_x^2}\sum_{n=1}^{N} x_n^2)\Big) \qquad (4)\end{aligned}$$

where

$$g(\mathcal{X}) := \frac{1}{N\sigma_x^2}\sum_{n=1}^{N} y_n x_n. \qquad (5)$$

For a large number N of the training data set, we assume the following approximation: $\sum_{n=1}^{N} x_n^2 \approx N\sigma_x^2$. By embedding this into (4) we obtain a form that allows for an unbiased estimator to attain the Cramér-Rao bound, and the corresponding estimate is given by

$$\hat{\theta} = g(\mathcal{X}) = \frac{1}{N\sigma_x^2}\sum_{n=1}^{N} y_n x_n. \qquad (6)$$

However, (6) is the LS estimator for large values of N. Indeed, by the definition of the LS estimator we have that

$$\frac{1}{N}\left(\sum_{n=1}^{N} x_n^2\right)\hat{\theta} = \frac{1}{N}\sum_{n=1}^{N} x_n y_n, \qquad (7)$$

which results in (6). It is easy to verify that (7) corresponds to an unbiased estimator.

Let us do it for the sake of an exercise. First of all, let us examine if the

LS estimator for this more general case is unbiased. We have

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\sum_n \mathrm{x}_n \mathrm{y}_n\right] = \mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\sum_n \mathrm{x}_n(\theta x_n + \eta_n)\right] \\
&= \mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\left(\theta\sum_n \mathrm{x}_n^2 + \sum_n \mathrm{x}_n \eta_n\right)\right] = \theta + \mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\sum_n \mathrm{x}_n \eta_n\right] \\
&= \theta + \mathbb{E}_x\left[\frac{1}{\sum_n \mathrm{x}_n^2}\,\mathbb{E}_{\eta|x}\left[\sum_n \mathrm{x}_n \eta_n\right]\right] \\
&= \theta + 0 = \theta.
\end{aligned}$$

In other words, the LS estimator is unbiased even for this case, where both output as well as input samples change in the training set and this is true independent of the number of measurements. The corresponding variance is given by

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta}-\theta)^2\right] &= \mathbb{E}\left[\frac{1}{(\sum_n \mathrm{x}_n^2)^2}\left(\sum_n \mathrm{x}_n \eta_n\right)^2\right] \\
&= \mathbb{E}_x\left[\frac{1}{(\sum_n \mathrm{x}_n^2)^2}\,\mathbb{E}_{\eta|x}\left[\sum_n \mathrm{x}_n^2 \eta_n^2 + \sum_i \sum_{j\neq i} \mathrm{x}_i \mathrm{x}_j \eta_i \eta_j\right]\right] \\
&= \sigma_\eta^2\,\mathbb{E}_x\left[\frac{\sum_n \mathrm{x}_n^2}{(\sum_n \mathrm{x}_n^2)^2}\right] = \sigma_\eta^2\,\mathbb{E}\left[\frac{1}{\sum_n \mathrm{x}_n^2}\right].
\end{aligned}$$

Asymptotically, this provides the bound that we have previously derived. However, for finite N, this is different.

3.8. Let us consider the regression model

$$y_n = \boldsymbol{\theta}^T \boldsymbol{x}_n + \eta_n, \quad n = 1, 2, \ldots, N,$$

where the noise vector $\boldsymbol{\eta} := [\eta_1, \ldots, \eta_N]^T$ comprises samples from zero mean Gaussian random variable, with covariance matrix Σ_η. If $X := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T$ stands for the input matrix, and $\boldsymbol{y} = [y_1, \ldots, y_N]^T$, the vector of the observations, then show that the corresponding estimator,

$$\hat{\theta} = \left(X^T \Sigma_\eta^{-1} X\right)^{-1} X^T \Sigma_\eta^{-1} \mathbf{y},$$

is an efficient one.

Notice, here, that the previous estimator coincides with the Maximum Likelihood (ML) one. Moreover, bear in mind that in the case where the noise process is considered to be white, i.e., $\Sigma_\eta = \sigma^2 I_N$, then the ML estimate becomes equal to the LS one.

Solution: In the case where the parameter $\boldsymbol{\theta}$ becomes a k-dimensional vector, the Cramér-Rao bound takes a more general form than the one we have met previously, i.e., the case where the parameter θ is a scalar. For any unbiased estimator $g(\mathcal{X})$ of the unknown parameter vector $\boldsymbol{\theta}$, the Cramér-Rao bound becomes as follows:

$$\mathbb{E}\left[(g(\mathcal{X})-\boldsymbol{\theta})(g(\mathcal{X})-\boldsymbol{\theta})^T\right] \succeq I^{-1}(\boldsymbol{\theta}), \quad \forall\boldsymbol{\theta},$$

where $I(\boldsymbol{\theta})$ is the *Fisher information matrix* defined as

$$I(\boldsymbol{\theta}) := -\,\mathbb{E}\left[\frac{\partial^2 \ln p(\mathcal{X};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\right],$$

and which is known to be a positive semidefinite matrix. Given any matrices A, B, of the same dimensions, the inequality $A \succeq B$ means that the matrix $A-B$ is positive semidefinite. A necessary and sufficient condition for g to be efficient is for the equation

$$\frac{\partial \ln p(\mathcal{X};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = I(\boldsymbol{\theta})\left(g(\mathcal{X})-\boldsymbol{\theta}\right). \tag{8}$$

For the present model, we have that $\mathcal{X}=\boldsymbol{y}$ and

$$p(\boldsymbol{y};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{N}{2}}(\det\Sigma_\eta)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\left(\boldsymbol{y}-X\boldsymbol{\theta}\right)^T\Sigma_\eta^{-1}\left(\boldsymbol{y}-X\boldsymbol{\theta}\right)\right).$$

Hence, $\ln p(\boldsymbol{y};\boldsymbol{\theta}) = -\frac{1}{2}\left(\boldsymbol{y}-X\boldsymbol{\theta}\right)^T\Sigma_\eta^{-1}\left(\boldsymbol{y}-X\boldsymbol{\theta}\right) + \text{constant}$, and

$$\begin{aligned}\frac{\partial \ln p(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} &= X^T\Sigma_\eta^{-1}\boldsymbol{y} - X^T\Sigma_\eta^{-1}X\boldsymbol{\theta} \\ &= X^T\Sigma_\eta^{-1}X\left((X^T\Sigma_\eta^{-1}X)^{-1}X^T\Sigma_\eta^{-1}\boldsymbol{y}-\boldsymbol{\theta}\right).\end{aligned} \tag{9}$$

The second derivative is equal to

$$\frac{\partial^2 \ln p(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2} = -X^T\Sigma_\eta^{-1}X,$$

so that the Fisher information matrix becomes $I(\boldsymbol{\theta}) = X^T\Sigma_\eta^{-1}X$. By this, and by a simple inspection of (9), we can readily verify that (8) is satisfied, if $g(\boldsymbol{y}) = \left(X^T\Sigma_\eta^{-1}X\right)^{-1}X^T\Sigma_\eta^{-1}\boldsymbol{y}$. This establishes the claim. Moreover, note that the covariance matrix of the efficient estimator is given by $(X^T\Sigma_\eta X)^{-1}$.

3.9. Assume a set of i.i.d $\mathcal{X} := \{x_1, x_2, \ldots, x_N\}$ Gaussian random variables, with mean μ and variance σ^2. Define also the quantities

$$S_\mu := \frac{1}{N}\sum_{n=1}^N x_n, \quad S_{\sigma^2} := \frac{1}{N}\sum_{n=1}^N (x_n - S_\mu)^2,$$

$$\bar{S}_{\sigma^2} := \frac{1}{N}\sum_{n=1}^N (x_n-\mu)^2.$$

Show that if μ is considered to be known, a sufficient statistic for σ^2 is $\bar{S}_{\sigma^2}$. Moreover, in the case where both (μ, σ^2) are unknown, then a sufficient statistic is the pair (S_μ, S_{σ^2}).

Solution: The joint pdf of $\mathcal{X}$ is obviously

$$p(\mathcal{X}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_n (x_n - \mu)^2\right).$$

If only σ^2 is considered to be unknown, then notice that

$$p(\mathcal{X}; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{N}{2\sigma^2}\bar{S}_{\sigma^2}\right) \equiv g_1(\bar{S}_{\sigma^2}, \sigma^2),$$

where g_1 is a function that depends only on $\bar{S}_{\sigma^2}$ and the unknown σ^2. According to the Fisher-Neyman factorization theorem, the statistic $\bar{S}_{\sigma^2}$ is sufficient.

Assume now the case where both (μ, σ^2) are unknown. Notice that by

$$\begin{aligned}\sum_{n=1}^N (x_n - \mu)^2 &= \sum_{n=1}^N (x_n - S_\mu)^2 + N(S_\mu - \mu)^2 \\ &= NS_{\sigma^2} + N(S_\mu - \mu)^2,\end{aligned}$$

the previous joint pdf becomes

$$\begin{aligned}p(\mathcal{X}; \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left(NS_{\sigma^2} + N(S_\mu - \mu)^2\right)\right) \\ &:= g_2\left((S_\mu, S_{\sigma^2}), (\mu, \sigma^2)\right).\end{aligned}$$

It can be readily verified that g_2 depends only on the statistic (S_μ, S_{σ^2}) and the unknowns (μ, σ^2). Hence, once again, by the Fisher-Neyman factorization theorem, the statistic (S_μ, S_{σ^2}) is sufficient.

3.10. Show that solving the task

$$\text{minimize} \qquad L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^N \left(y_n - \theta_0 - \sum_{i=1}^l \theta_i x_{ni}\right)^2 + \lambda \sum_{i=1}^l |\theta_i|^2,$$

is equivalent with minimizing

$$\text{minimize} \qquad L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^N \left((y_n - \bar{y}) - \sum_{i=1}^l \theta_i (x_{ni} - \bar{x}_i)\right)^2 + \lambda \sum_{i=1}^l |\theta_i|^2,$$

and the estimate of θ_0 is given by

$$\hat{\theta}_0 = \bar{y} - \sum_{i=1}^l \hat{\theta}_i \bar{x}_i.$$

Solution: We have that

$$L(\theta_0, \theta_{1:l}) = \sum_{n=1}^{N} \Big(y_n - \theta_0 - \sum_{i=1}^{l} \theta_i x_{ni} \Big)^2 + \sum_{i=1}^{l} \theta_i^2.$$

Taking first the derivative with respect to θ_0 and setting it equal to zero we obtain

$$\frac{\partial L}{\partial \theta_0} = \sum_{n=1}^{N} \Big(-2(y_n - \theta_0) + 2\sum_{i=1}^{l} \theta_i x_{ni} \Big) = 0$$

or

$$N\theta_0 = \sum_{n=1}^{N} y_n - \sum_{i=1}^{l} \theta_i \sum_{n=1}^{N} x_{ni},$$

which results in

$$\theta_0 = \bar{y} - \sum_{i=1}^{l} \theta_i \bar{x}_i.$$

That is, the optimum value for θ_0, is given in terms of the rest components. Thus, optimizing with respect to $\theta_i,\ i = 1, 2, \ldots, l$, this has to be taken into account. Substituting the above in the Lagrangian, we get

$$L(\hat{\theta}_0, \theta_{1:l}) = \sum_{n=1}^{N} \Big(y_n - \bar{y} - \sum_{i=1}^{l} \theta_i (x_{ni} - \bar{x}_i) \Big)^2 + \sum_{i=1}^{l} \theta_i^2,$$

which proves the claim.

Note that the exact form of the regularizer does not enter into the game, since does not depend on θ_0. Hence, this technique of centering the data is also applicable to other forms of regularization.

3.11. This problem refers to Example 3.4, where a linear regression task with a real valued unknown parameter θ is considered. Show that $\text{MSE}(\hat{\uptheta}_b(\lambda)) < \text{MSE}(\hat{\uptheta}_{\text{MVU}})$, i.e., the ridge regression estimate shows a lower MSE performance than the one for the MVU estimate, if

$$\begin{cases} \lambda \in (0, \infty), & \theta^2 \leq \frac{\sigma_\eta^2}{N}, \\ \lambda \in \left(0, \frac{2\sigma_\eta^2}{\theta^2 - \frac{\sigma_\eta^2}{N}}\right), & \theta^2 > \frac{\sigma_\eta^2}{N}. \end{cases}$$

Moreover, the minimum MSE performance for the ridge regression estimate is attained at $\lambda_* = \sigma_\eta^2/\theta^2$.

Solution: Theory suggests that our estimate $\hat{\theta}_b$ is the solution of the task of minimizing the following loss function with respect to $\theta \in \mathbb{R}$:

$$L(\theta, \lambda) = \sum_{n=1}^{N} (y_n - \theta)^2 + \lambda\theta^2, \quad \lambda \geq 0.$$

The minimizer $\hat{\theta}_b$ will be obtained if we set the gradient $dL(\theta, \lambda)/d\theta$ equal to zero, or equivalently,

$$\hat{\theta}_b(\lambda) = \frac{N}{N+\lambda} \frac{1}{N} \sum_{n=1}^{N} y_n := \frac{N}{N+\lambda} \hat{\theta}_{\mathrm{MVU}},$$

where we used the notation $\hat{\theta}_b(\lambda)$ in order to highlight the dependence of the estimate $\hat{\theta}_b$ on the parameter λ. Notice here that $\mathbb{E}[\hat{\theta}_b(\lambda)] = \frac{N}{N+\lambda}\theta_0$, where θ_0 is the estimandum.

Elementary calculus helps us to express $\mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right)$ as

$$\begin{aligned} \mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right) &= \mathbb{E}\left[\left(\hat{\theta}_b(\lambda) - \mathbb{E}[\hat{\theta}_b(\lambda)]\right)^2\right] + \left(\mathbb{E}[\hat{\theta}_b(\lambda)] - \theta_0\right)^2 \\ &= \frac{N^2 \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right) + \lambda^2\theta_0^2}{(N+\lambda)^2}, \end{aligned} \tag{10}$$

and

$$\frac{d}{d\lambda}\mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right) = \frac{2\theta_0^2\lambda(N+\lambda)^2 - 2\left(N^2\mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right) + \lambda^2\theta_0^2\right)(N+\lambda)}{(N+\lambda)^4}. \tag{11}$$

Let us first examine the range of values of $\lambda > 0$ which guarantee that $\mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right) < \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right)$. The case of $\lambda = 0$ is excluded from the discussion, since in such a case, $\hat{\theta}_b(0) = \hat{\theta}_{\mathrm{MVU}}$. It is easy to verify by (10) that

$$\begin{aligned} &\mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right) < \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right) \\ \Leftrightarrow\ &\lambda\left(\theta_0^2 - \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right)\right) < 2N\mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right). \end{aligned}$$

In the case where $\theta_0^2 > \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right)$, then the desired λ belongs to the interval $(0, 2\sigma_\eta^2/(\theta_0^2 - \sigma_\eta^2/N))$, where we have used the fact that

$$\mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right) = \frac{\sigma_\eta^2}{N}.$$

In the case where $\theta_0^2 \leq \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right)$, notice that $\forall\lambda > 0$, we have that

$$\lambda\left(\theta_0^2 - \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right)\right) \leq 0 < 2N\mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right),$$

i.e., the desired λ belongs to the interval $(0, \infty)$.

It is also easy to verify by equating the numerator of (11) to zero that the λ_* which minimizes $\mathrm{MSE}\left(\hat{\theta}_b(\lambda)\right)$ becomes equal to σ_η^2/θ_0^2. To leave no place for ambiguity, we remark here that in the case where $\theta_0^2 > \mathrm{MSE}\left(\hat{\theta}_{\mathrm{MVU}}\right) = \sigma_\eta^2/N$, this λ_* belongs to the interval $(0, 2\sigma_\eta^2/(\theta_0^2 - \sigma_\eta^2/N))$, since

$$0 < \lambda_* = \frac{\sigma_\eta^2}{\theta_0^2} < 2\frac{\sigma_\eta^2}{\theta_0^2} < \frac{2\sigma_\eta^2}{\theta_0^2 - \frac{\sigma_\eta^2}{N}}.$$

3.12. Assume that the model that generates the data is

$$y_n = A \sin\left(\frac{2\pi}{N}kn + \phi\right) + \eta_n, \tag{12}$$

where $A > 0$, and $k \in \{1, 2, \ldots, N-1\}$. Assume that η_n are samples from a Gaussian white noise, of variance σ_η^2. Show that there is no unbiased estimator for the phase, ϕ, based on N measurement points, y_n, $n = 0, 1, \ldots N-1$, that attains the Cramér-Rao bound.

Solution: The joint pdf of the measurements $\boldsymbol{y} := [y_0, y_1, \ldots, y_{N-1}]^T$ is given by

$$p(\boldsymbol{y}; \phi) = \frac{1}{(2\pi\sigma_\epsilon^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{n=0}^{N-1}\left(y_n - A\sin\left(\frac{2\pi}{N}kn + \phi\right)\right)^2\right).$$

The two derivatives of the ln of this function can be easily shown to be

$$\frac{\partial \ln p(\boldsymbol{y};\phi)}{\partial \phi} = \frac{A}{\sigma_\epsilon^2}\sum_{n=0}^{N-1}\left(y_n \cos\left(\frac{2\pi}{N}kn + \phi\right) - \frac{A}{2}\sin\left(\frac{4\pi}{N}kn + 2\phi\right)\right), \tag{13}$$

and

$$\frac{\partial^2 \ln p(\boldsymbol{y};\phi)}{\partial \phi^2} = -\frac{A}{\sigma_\epsilon^2}\sum_{n=0}^{N-1}\left(y_n \sin\left(\frac{2\pi}{N}kn + \phi\right) + A\cos\left(\frac{4\pi}{N}kn + 2\phi\right)\right),$$

and by substituting the value of y_n from (12), the mean value becomes

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{y};\phi)}{\partial \phi^2}\right] &= -\frac{A^2}{\sigma_\epsilon^2}\sum_{n=0}^{N-1}\left(\frac{1}{2} + \frac{1}{2}\cos\left(\frac{4\pi}{N}kn + 2\phi\right)\right) \\ &= -\frac{NA^2}{2\sigma_\epsilon^2}.\end{aligned}$$

To derive the above, we used the trigonometric formula $\sin^2\alpha = (1 -$

$\cos(2\alpha))/2$, and also the fact

$$\begin{aligned}\sum_{n=0}^{N-1}\cos\left(\frac{4\pi}{N}kn+2\phi\right) &= \frac{1}{2}\sum_{n=0}^{N-1}\left(e^{j\left(\frac{4\pi}{N}kn+2\phi\right)}+e^{-j\left(\frac{4\pi}{N}kn+2\phi\right)}\right)\\ &= \frac{1}{2}e^{2j\phi}\sum_{n=0}^{N-1}e^{j\frac{4\pi}{N}kn}+\frac{1}{2}e^{-2j\phi}\sum_{n=0}^{N-1}e^{-j\frac{4\pi}{N}kn}\\ &= \frac{1}{2}e^{2j\phi}\frac{1-e^{j\frac{4\pi}{N}kN}}{1-e^{j\frac{4\pi}{N}k}}+\frac{1}{2}e^{-2j\phi}\frac{1-e^{-j\frac{4\pi}{N}kN}}{1-e^{-j\frac{4\pi}{N}k}}\\ &= 0,\end{aligned}$$

since $e^{-j\frac{4\pi}{N}kN}=1$, where $j:=\sqrt{-1}$.

Hence, if $\hat{\phi}$ stands for an unbiased estimator of ϕ, then

$$\mathrm{var}(\hat{\phi}) \geq \frac{2\sigma_\epsilon^2}{NA^2}.$$

However, looking back at (13), we can verify that there does not exist a function g such that $\forall \boldsymbol{y}\in\mathbb{R}^N$,

$$g(\boldsymbol{y})-\phi=\frac{2}{NA}\sum_{n=0}^{N-1}\left(y_n\cos\left(\frac{2\pi}{N}kn+\phi\right)-\frac{A}{2}\sin\left(\frac{4\pi}{N}kn+2\phi\right)\right).$$

Thus, even if an unbiased estimator exists, this cannot achieve the Cramér-Rao bound.

3.13. Show that if $(\mathbf{y},\mathbf{x})$ are two jointly distributed random vectors, with values in $\mathbb{R}^k\times\mathbb{R}^l$, then the MSE optimal estimator of $\mathbf{y}$ given the value $\mathbf{x}=\boldsymbol{x}$ is the regression of $\mathbf{y}$ conditioned on $\boldsymbol{x}$, i.e., $\mathbb{E}[\mathbf{y}|\boldsymbol{x}]$.

*S*olution: The proof follows a similar line as the scalar case. Let

$$\boldsymbol{f}(\boldsymbol{x}):=[f_1(\boldsymbol{x}),\ldots,f_k(\boldsymbol{x})]^T$$

be the vector estimator. Then the MSE optimal one should minimize the sum of square errors per component, i.e.,

$$\mathbb{E}[\sum_{i=1}^{k}(\mathrm{y}_i-f_i(\boldsymbol{x}))^2=\sum_{i=1}^{k}\mathbb{E}[(\mathrm{y}_i-f_i(\boldsymbol{x}))^2].$$

This is equivalent with minimizing l scalar terms individually, which can be carried out as in the text in the Chapter. The result of the ith problem is that the respective ith component of the MSE optimal estimator is given by,

$$\hat{g}_i(\boldsymbol{x})=\mathbb{E}[\mathrm{y}_i|\boldsymbol{x}],$$

or

$$\hat{\boldsymbol{g}}(\boldsymbol{x})=\mathbb{E}[\mathbf{y}|\boldsymbol{x}].$$

Note that minimizing the sum of square errors per component is equivalent with minimizing the trace of the error covariance,

$$\mathbb{E}[\mathbf{e}\mathbf{e}^T] = \mathbb{E}[(\mathbf{y} - \boldsymbol{f}(\boldsymbol{x}))(\mathbf{y} - \boldsymbol{f}(\boldsymbol{x}))^T].$$

3.14. Assume that $\mathbf{x}$, $\mathbf{y}$ are jointly Gaussian random vectors, with covariance matrix

$$\Sigma := \mathbb{E}\left[\begin{bmatrix}\mathbf{x} - \boldsymbol{\mu}_x \\ \mathbf{y} - \boldsymbol{\mu}_y\end{bmatrix}\left[(\mathbf{x} - \boldsymbol{\mu}_x)^T, (\mathbf{y} - \boldsymbol{\mu}_y)^T\right]\right] = \begin{bmatrix}\Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y\end{bmatrix}.$$

Assuming also that the matrices Σ_x and $\bar{\Sigma} := \Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}$ are non-singular, then show that the optimal MSE estimator $\mathbb{E}[\mathbf{y}|\boldsymbol{x}]$ takes the following form

$$\mathbb{E}[\mathbf{y}|\boldsymbol{x}] = \mathbb{E}[\mathbf{y}] + \Sigma_{yx}\Sigma_x^{-1}(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]).$$

Notice that $\mathbb{E}[\mathbf{y}|\boldsymbol{x}]$ is an affine function of $\boldsymbol{x}$. In other words, for the case where $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, the optimal estimator of $\mathbf{y}$, in the MSE sense, which is in general a non-linear function, becomes an affine function of $\boldsymbol{x}$.

In the special case where x, y are scalar random variables, then

$$\mathbb{E}[\mathrm{y}|x] = \mu_y + \frac{\alpha\sigma_y}{\sigma_x}(x - \mu_x),$$

where α stands for the *correlation coefficient*, defined as

$$\alpha := \frac{\mathbb{E}\left[(\mathrm{x} - \mu_x)(\mathrm{y} - \mu_y)\right]}{\sigma_x\sigma_y},$$

with $|\alpha| \leq 1$. Notice, also, that the previous assumption on the non-singularity of Σ_x and $\bar{\Sigma}$ translates, in this special case, to $\sigma_x \neq 0 \neq \sigma_y$, and $|\alpha| < 1$.

Solution: First, it is easy to verify that $\Sigma_{yx} = \Sigma_{xy}^T$. Moreover, since Σ_x and $\bar{\Sigma}$ are assumed to be non-singular, then it can be verified [Magn 99] that the determinant $\det \Sigma = \det \Sigma_x \det \bar{\Sigma}$, and that

$$\Sigma^{-1} = \begin{bmatrix}\Sigma_x^{-1} + \Sigma_x^{-1}\Sigma_{xy}\bar{\Sigma}^{-1}\Sigma_{yx}\Sigma_x^{-1} & -\Sigma_x^{-1}\Sigma_{xy}\bar{\Sigma}^{-1} \\ -\bar{\Sigma}^{-1}\Sigma_{yx}\Sigma_x^{-1} & \bar{\Sigma}^{-1}\end{bmatrix}.$$

Observe that $\bar{\Sigma}$ is the Schur complement of $\Sigma_{\boldsymbol{x}}$ in Σ. Also, the previous formula is the matrix inversion formula in terms of the Schur complement, as provided in the Appendix A of the book. To save space, let $\bar{\boldsymbol{x}} := \boldsymbol{x} - \boldsymbol{\mu}_x$

and $\bar{\boldsymbol{y}} := \boldsymbol{y} - \boldsymbol{\mu}_y$. Then, the joint pdf of $\mathbf{x}$ and $\mathbf{y}$ becomes

$$
\begin{aligned}
p(\boldsymbol{x},\boldsymbol{y}) &= \frac{1}{(2\pi)^l (\det \varSigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \left[\bar{\boldsymbol{x}}^T, \bar{\boldsymbol{y}}^T \right] \varSigma^{-1} \begin{bmatrix} \bar{\boldsymbol{x}} \\ \bar{\boldsymbol{y}} \end{bmatrix} \right) \\
&= \frac{1}{(2\pi)^l (\det \varSigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \bar{\boldsymbol{x}}^T \varSigma_x^{-1} \bar{\boldsymbol{x}} - \frac{1}{2} \bar{\boldsymbol{x}}^T \varSigma_x^{-1} \varSigma_{xy} \bar{\varSigma}^{-1} \varSigma_{yx} \varSigma_x^{-1} \bar{\boldsymbol{x}} \right. \\
&\quad \left. + \bar{\boldsymbol{x}}^T \varSigma_x^{-1} \varSigma_{xy} \bar{\varSigma}^{-1} \bar{\boldsymbol{y}} - \frac{1}{2} \bar{\boldsymbol{y}}^T \bar{\varSigma}^{-1} \bar{\boldsymbol{y}} \right) \\
&= \frac{1}{(2\pi)^l (\det \varSigma_x)^{\frac{1}{2}} (\det \bar{\varSigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \bar{\boldsymbol{x}}^T \varSigma_x^{-1} \bar{\boldsymbol{x}} \right. \\
&\quad \left. -\frac{1}{2} \left(\bar{\boldsymbol{y}} - \varSigma_{yx} \varSigma_x^{-1} \bar{\boldsymbol{x}} \right)^T \bar{\varSigma}^{-1} \left(\bar{\boldsymbol{y}} - \Sigma_{yx} \varSigma_x^{-1} \bar{\boldsymbol{x}} \right) \right).
\end{aligned}
$$

As a result, the marginal pdf $p(\boldsymbol{x})$ becomes

$$
\begin{aligned}
p(\boldsymbol{x}) = \int p(\boldsymbol{x},\boldsymbol{y}) d\boldsymbol{y} &= \frac{1}{(2\pi)^{l/2} (\det \varSigma_x)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \bar{\boldsymbol{x}}^T \varSigma_x^{-1} \bar{\boldsymbol{x}} \right) \\
&\times \frac{1}{(2\pi)^{l/2} (\det \bar{\varSigma})^{\frac{1}{2}}} \int \exp\left(-\frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{\mu}_y - \varSigma_{yx} \varSigma_x^{-1} \bar{\boldsymbol{x}} \right)^T \right. \\
&\quad \times \left. \bar{\varSigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}_y - \varSigma_{yx} \varSigma_x^{-1} \bar{\boldsymbol{x}} \right) \right) d\boldsymbol{y} \\
&= \frac{1}{(2\pi)^{l/2} (\det \varSigma_x)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \bar{\boldsymbol{x}}^T \varSigma_x^{-1} \bar{\boldsymbol{x}} \right).
\end{aligned}
$$

Using the previous relations, we can easily see that

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})} \\
&= \frac{1}{(2\pi)^{\frac{l}{2}} (\det \bar{\varSigma})^{\frac{1}{2}}} \exp\left(- \frac{\left(\bar{\boldsymbol{y}} - \varSigma_{yx} \varSigma_x^{-1} \bar{\boldsymbol{x}} \right)^T \bar{\varSigma}^{-1} \left(\bar{\boldsymbol{y}} - \varSigma_{yx} \varSigma_x^{-1} \bar{\boldsymbol{x}} \right)}{2} \right).
\end{aligned}
$$

A simple inspection of this relation shows that the conditional pdf $p(\boldsymbol{y}|\boldsymbol{x})$ is Gaussian with covariance matrix $\bar{\varSigma}$ and conditional mean $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{\mu}_y - \varSigma_{yx} \varSigma_x^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x)$.

3.15. Assume a number l of jointly Gaussian random variables $\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_l\}$, and a non-singular matrix $A \in \mathbb{R}^{l \times l}$. If $\mathbf{x} := [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_l]^T$, then show that the components of the vector $\mathbf{y}$, obtained by $\mathbf{y} = A\mathbf{x}$, are also jointly Gaussian random variables.

A direct consequence of this result is that any linear combination of jointly Gaussian variables is also Gaussian.

Solution: The Jacobian matrix of a linear transform $\mathbf{y} = A\mathbf{x}$ is easily shown to be

$$
J := J(\mathbf{y}; \mathbf{x}) = A.
$$

Also, since A is non-singular, we have that $\mathbf{x} = A^{-1}\mathbf{y}$. Without any loss of generality, assume that $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, which results into $\mathbb{E}[\mathbf{y}] = \mathbf{0}$. Hence,

$$\Sigma_y = \mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right] = \mathbb{E}\left[A\mathbf{x}\mathbf{x}^T A^T\right] = A\Sigma_x A^T.$$

Clearly, $\det \Sigma_y = (\det A)^2 \det \Sigma_x$. Then, by the theorem of transformation for random variables, e.g., [Papo 02], we have the following:

$$\begin{aligned} p(\boldsymbol{y}) &= \frac{p(\boldsymbol{x})}{|\det J|} = \frac{p(A^{-1}\boldsymbol{y})}{|\det A|} \\ &= \frac{1}{(2\pi)^{l/2}|\det A|(\det \Sigma_x)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T A^{-T}\Sigma_x^{-1}A^{-1}\boldsymbol{y}\right) \\ &= \frac{1}{(2\pi)^{l/2}(\det \Sigma_y)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\Sigma_y^{-1}\boldsymbol{y}\right), \end{aligned}$$

which establishes the first claim.

For the second claim, assume a non-zero vector $\boldsymbol{a} \in \mathbb{R}^l$, and define the linear combination of $\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_l\}$ as $\mathrm{y} = \boldsymbol{a}^T\mathbf{x}$. Elementary linear algebra guarantees that there always exists a set of non-zero l-dimensional vectors $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{l-1}\}$ such that the collection $\{\boldsymbol{a}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{l-1}\}$ constitutes a basis of $\mathbb{R}^l$ [Magn 99]. Thus, the matrix $A := [\boldsymbol{a}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{l-1}]^T \in \mathbb{R}^{l\times l}$ is non-singular, and the first component of the vector $\mathbf{y} = A\mathbf{x}$ is the quantity $\mathrm{y} = \boldsymbol{a}^T\mathbf{x}$. We have already seen by the first claim that the components of $\mathbf{y}$ are jointly Gaussian random variables. Moreover, a classical result states that if a number of random variables are jointly Gaussian, then each one of them, and thus y, is also Gaussian (the opposite is not always true) [Papo 02]. This establishes the second claim of the problem.

3.16. Let $\mathbf{x} \in \mathbb{R}^l$ be a vector of jointly Gaussian random variables, of covariance matrix Σ_x. Consider the general linear regression model

$$\mathbf{y} = \Theta\mathbf{x} + \boldsymbol{\eta},$$

where $\Theta \in \mathbb{R}^{k\times l}$ is a parameter matrix and $\boldsymbol{\eta}$ is the vector of noise samples, which are considered to be Gaussian, with zero mean, and with covariance matrix Σ_η, independent of $\mathbf{x}$. Then show that $\mathbf{y}$ and $\mathbf{x}$ are jointly Gaussian, with covariance matrix given by

$$\Sigma = \begin{bmatrix} \Theta\Sigma_x\Theta^T + \Sigma_{eta} & \Theta\Sigma_x \\ \Sigma_x\Theta^T & \Sigma_x \end{bmatrix}.$$

Solution: The combined vector is given by

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \Theta\mathbf{x} + \boldsymbol{\eta} \\ \mathbf{x} \end{bmatrix} = \underbrace{\begin{bmatrix} \Theta & I_k \\ I_l & 0_{l\times k} \end{bmatrix}}_{A} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\eta} \end{bmatrix}.$$

However, since $\mathbf{x}$ and $\boldsymbol{\eta}$ are both Gaussian vector variables, and mutually independent, then they are also jointly Gaussian. Notice also that the matrix A is non-singular; indeed, a simple permutation of the columns of A leads to the matrix $\begin{bmatrix} I_k & \Theta \\ 0_{l\times k} & I_l \end{bmatrix}$, whose determinant can be easily seen to be equal to 1.

Therefore, according to Problem 3.15, $[\mathbf{y}^T, \mathbf{x}^T]^T$ is also jointly Gaussian. The covariance matrix is a straightforward result following the definitions of the involved variables.

3.17. Show that a linear combination of Gaussian independent variables is also Gaussian.

Solution: This is a direct consequence of Problem 3.15, since independent Gaussian variables can be readily checked out that they are also jointly Gaussian.

3.18. Show that if a sufficient statistic $T(\mathcal{X})$ for a parameter estimation problem exists, then $T(\mathcal{X})$ suffices to express the respective ML estimate.

Solution: This is direct consequence of the Fisher-Neyman factorization theorem. Indeed, recall that $T(\mathcal{X})$ is sufficient iff the respective joint pdf can be factored as: $p(\mathcal{X};\boldsymbol{\theta}) = h(\mathcal{X})g(T(\mathcal{X}),\boldsymbol{\theta})$, where h and g are appropriate functions. Hence, by the definition of the ML estimate,

$$\hat{\boldsymbol{\theta}}_{\text{ML}} \equiv \arg\max_{\boldsymbol{\theta}} p(\mathcal{X};\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} g(T(\mathcal{X}),\boldsymbol{\theta}).$$

In other words, $T(\mathcal{X})$ is sufficient, via g, to obtain the ML estimate.

3.19. Show that if an efficient estimator exists then it is also optimal in the ML sense.

Solution: Assume the existence of an efficient estimator, i.e., a function g which achieves the Cramér-Rao bound. A necessary and sufficient condition for g to be efficient, is for (8) to hold true for all values of $\boldsymbol{\theta}$. Since (8) holds for all values of $\boldsymbol{\theta}$, then it holds true for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{ML}}$. However, for this value, the left-hand-side of (8) becomes zero, and since $I(\hat{\boldsymbol{\theta}}_{\text{ML}})$ is non-singular, we obtain that $g(\mathcal{X}) = \hat{\boldsymbol{\theta}}_{\text{ML}}$. This establishes the claim.

3.20. Let the observations resulting from an experiment be x_n, $n = 1, 2, \ldots, N$. Assume that they are independent and that they originate from a Gaussian pdf $\mathcal{N}(\mu, \sigma^2)$. Both, the mean and the variance, are unknown. Prove that the ML estimates of these quantities are given by

$$\hat{\mu}_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \hat{\sigma}^2_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu}_{\text{ML}})^2.$$

Solution: The log-likelihood function is given by

$$L(\mu, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2.$$

Taking the gradient with respect to μ, σ^2, and equating it to zero we obtain the following system of equations

$$\begin{aligned} \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) &= 0 \\ -\frac{N}{2}\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 &= 0. \end{aligned}$$

The solution of this system leads trivially to the required result.

3.21. Let the observations $x_n,\ n=1,2,\dots,N$, come from the uniform distribution

$$p(x;\theta)=\begin{cases}\frac{1}{\theta}, & 0<x\le\theta,\\ 0, & \text{otherwise.}\end{cases}$$

Obtain the ML estimate of θ.

Solution: The likelihood function is given by

$$L(\boldsymbol{x};\theta)=\prod_{n=1}^{N}\frac{1}{\theta}=\frac{1}{\theta^N}.$$

We know that $\theta\ge x_n$, $n=1,\dots,N$, or equivalently, $\theta\ge\max_{n=1,\dots,N}x_n$. Hence, the likelihood function is maximized by taking the minimum value of θ, which is

$$\hat{\theta}_{\mathrm{ML}}=\max\{x_1,x_2,\dots,x_N\}.$$

3.22. Obtain the ML estimate of the parameter $\lambda>0$ of the exponential distribution

$$p(x)=\begin{cases}\lambda\exp(-\lambda x), & x\ge 0,\\ 0, & x<0,\end{cases}$$

based on a set of measurements, x_n, $n=1,2,\dots,N$.

Solution: The log-likelihood function is

$$L(\boldsymbol{x};\lambda)=N\ln\lambda-\lambda\sum_{n=1}^{N}x_n.$$

Taking the derivative and equating to zero we obtain

$$\frac{N}{\lambda}-\sum_{n=1}^{N}x_n=0,$$

which leads to

$$\hat{\lambda}_{\mathrm{ML}}=\frac{N}{\sum_{n=1}^{N}x_n}.$$

3.23. Assume an $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, and a stochastic process $\{x_n\}_{n=-\infty}^{\infty}$, consisting of i.i.d random variables, such that $p(x_n|\mu) = \mathcal{N}(\mu, \sigma^2)$. Consider a number of N members of the process $\{x_n\}_{n=-\infty}^{\infty}$, i.e., $\mathcal{X} \equiv \{x_1, x_2, \ldots, x_N\}$, and prove that the posterior $p(x|\mathcal{X})$, of any $x = x_{n_0}$ conditioned on $\mathcal{X}$, turns out to be Gaussian with mean μ_N and variance $\sigma^2 + \sigma_N^2$, where

$$\mu_N \equiv \frac{N\sigma_0^2\bar{x} + \sigma^2\mu_0}{N\sigma_0^2 + \sigma^2}, \quad \sigma_N^2 \equiv \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}.$$

Solution: From basic theory we have that

$$p(\mu|\mathcal{X}) = \frac{p(\mathcal{X}|\mu)p(\mu)}{\int p(\mathcal{X}|\mu)p(\mu)d\mu} = \alpha p(\mu)\prod_{k=1}^{N} p(x_k|\mu),$$

or

$$\begin{aligned}
p(\mu|\mathcal{X}) &= \frac{\alpha}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)\prod_{k=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x_k-\mu)^2}{\sigma^2}\right)\\
&= \alpha_1\exp\left(-\frac{1}{2}\left(\sum_{k=1}^{N}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right)\\
&= \alpha_2\exp\left(-\frac{1}{2}\left(\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N}x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right)\right)\\
&= \alpha_2\exp\left(-\frac{1}{2}\left(\mu^2\frac{N\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2} - 2\mu\frac{N\sigma_0^2\bar{x}+\sigma^2\mu_0}{\sigma^2\sigma_0^2}\right)\right)\\
&= \alpha_2\exp\left(-\frac{N\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}\left(\mu^2 - 2\mu\frac{N\sigma_0^2\bar{x}+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2}\right)\right)\\
&= \alpha_3\exp\left(-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right),
\end{aligned}$$

where $\alpha, \alpha_1, \alpha_2, \alpha_3$ are factors independent of μ, and

$$\begin{aligned}
\bar{x} &:= \frac{1}{N}\sum_{k=1}^{N}x_k,\\
\mu_N &:= \frac{N\sigma_0^2\bar{x}+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2},\\
\sigma_N^2 &:= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}.
\end{aligned}$$

Since $p(\mu|\mathcal{X})$ is a pdf, then necessarily

$$\alpha_3 = \frac{1}{\sqrt{2\pi}\sigma_N}.$$

Hence, $\lim_{N\to\infty} \sigma_N^2 = 0$, and for large N, $p(\mu|\mathcal{X})$ behaves like a δ function centered around μ_N. Thus,

$$p(x|\mathcal{X}) = \int p(x|\mu)p(\mu|\mathcal{X})d\mu \simeq p(x|\mu_N) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu_N)^2}{2\sigma^2}\right).$$

Therefore, $p(x|\mathcal{X})$ tends to a Gaussian pdf with mean μ_N and variance σ^2. Furthermore, $\lim_{N\to\infty} \mu_N = \bar{x}$.

For the general case of any value of N, and not only the case of large N, we have

$$\begin{aligned} p(x|\mathcal{X}) &= \int p(x|\mu)p(\mu|\mathcal{X})d\mu \\ &= \frac{1}{2\pi\sigma\sigma_N} \int \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right) d\mu. \end{aligned} \tag{14}$$

Hence, in order to obtain $p(x|\mathcal{X})$, the previous integration has to take place. Here we will follow another path, which avoids any direct integration. Assume a random variable y defined as y := $\xi + \nu$, where $\xi \sim \mathcal{N}(0, \sigma^2)$ and $\nu \sim \mathcal{N}(\mu_N, \sigma_N^2)$, independent of each other. It is well-known [Papo 02] that the pdf of y is given by the joint pdf of ξ and ν as follows:

$$p(y) = \int p_{\xi\nu}(y-\nu, \nu)d\nu.$$

However, since ξ and ν are assumed to be independent, then $p_{\xi\nu}(\xi, \nu) = p_\xi(\xi)p_\nu(\nu)$, and

$$\begin{aligned} p(y) &= \int p_\xi(y-\nu)p_\nu(\nu)d\nu \\ &= \frac{1}{2\pi\sigma\sigma_N} \int \exp\left(-\frac{(y-\nu)^2}{2\sigma^2}\right) \exp\left(-\frac{(\nu-\mu_N)^2}{2\sigma_N^2}\right) d\nu, \end{aligned}$$

which is identical to (14). However, recall from basic statistics [Papo 02] that y, being the sum of two independent Gaussians is also Gaussian (see, also, Problem 3.17), with mean the sum of the mean values and variance the sum of the variances. Therefore, (14) becomes

$$p(x|\mathcal{X}) = \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_N^2)}} \exp\left(-\frac{(x-\mu_N)^2}{2(\sigma^2+\sigma_N^2)}\right).$$

3.24. Show that for the linear regression model,

$$\boldsymbol{y} = X\boldsymbol{\theta} + \boldsymbol{\eta},$$

the a-posteriori probability $p(\boldsymbol{\theta}|\boldsymbol{y})$ is a Gaussian one, if the prior distribution probability is given by $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}, \Sigma_0)$, and the noise samples follow the multivariate Gaussian distribution $p(\boldsymbol{\eta}) = \mathcal{N}(\mathbf{0}, \Sigma_\eta)$. Compute the mean vector and the covariance matrix of the posterior distribution.

Solution: It can be easily checked that $p(\boldsymbol{\theta}|\boldsymbol{y}) = \text{const} \times \exp\left(-\frac{1}{2}\Psi\right)$, where

$$\begin{aligned}\Psi &= (\boldsymbol{y} - X\boldsymbol{\theta})^T \Sigma_\eta^{-1} (\boldsymbol{y} - X\boldsymbol{\theta}) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \Sigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= \boldsymbol{y}^T \Sigma_\eta^{-1} \boldsymbol{y} - 2\boldsymbol{y}^T \Sigma_\eta^{-1} X\boldsymbol{\theta} + \boldsymbol{\theta}^T X^T \Sigma_\eta^{-1} X\boldsymbol{\theta} + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \Sigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0).\end{aligned}$$

From now on, all terms that will be independent of $\boldsymbol{\theta}$ will be collected in constant terms. Hence

$$\begin{aligned}\Psi = \alpha_1 &- 2\boldsymbol{y}^T \Sigma_\eta^{-1} X\boldsymbol{\theta} + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \Sigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &+ (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T X^T \Sigma_\eta^{-1} X (\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \boldsymbol{\theta}_0^T X^T \Sigma_\eta^{-1} X\boldsymbol{\theta}_0 + 2\boldsymbol{\theta}_0^T X^T \Sigma_\eta^{-1} X\boldsymbol{\theta}.\end{aligned}$$

As a result,

$$\begin{aligned}\Psi &= \alpha_2 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left(\Sigma_0^{-1} + X^T \Sigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + 2\boldsymbol{\theta}_0^T X^T \Sigma_\eta^{-1} X\boldsymbol{\theta} - 2\boldsymbol{y}^T \Sigma_\eta^{-1} X\boldsymbol{\theta} \\ &= \alpha_3 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left(\Sigma_0^{-1} + X^T \Sigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad - 2\left(\boldsymbol{y} - X\boldsymbol{\theta}_0\right)^T \Sigma_\eta^{-1} X \left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right). \end{aligned} \tag{15}$$

In the sequel, we will follow a standard trick that we do in situations like that. We introduce an auxiliary variable $\bar{\boldsymbol{\theta}}$, whose value is to be determined so that to make the following to be true,

$$\begin{aligned}\Psi &= \alpha_4 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}})^T \left(\Sigma_0^{-1} + X^T \Sigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}) \\ &= \alpha_4 + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left(\Sigma_0^{-1} + X^T \Sigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + \bar{\boldsymbol{\theta}}^T \left(\Sigma_0^{-1} + X^T \Sigma_\eta^{-1} X\right) \bar{\boldsymbol{\theta}} \\ &\quad - 2\bar{\boldsymbol{\theta}}^T \left(\mathrm{\Sigma}_0^{-1} + X^T \Sigma_\eta^{-1} X\right) (\boldsymbol{\theta} - \boldsymbol{\theta}_0). \end{aligned} \tag{16}$$

Inspection of (15) and (16) indicates that this can happen if we choose

$$\bar{\boldsymbol{\theta}} = \left(\Sigma_0^{-1} + X^T \Sigma_\eta^{-1} X\right)^{-1} X^T \Sigma_\eta^{-1} \left(\boldsymbol{y} - X\boldsymbol{\theta}_0\right).$$

Then, we can finally write that

$$\Psi = \alpha_4 + \left(\boldsymbol{\theta} - \mathbb{E}[\theta|\boldsymbol{y}]\right)^T \Sigma_{\boldsymbol{\theta}|\boldsymbol{y}}^{-1} \left(\boldsymbol{\theta} - \mathbb{E}[\theta|\boldsymbol{y}]\right),$$

where

$$\mathbb{E}[\theta|\boldsymbol{y}] = \boldsymbol{\theta}_0 + \left(\Sigma_0^{-1} + X^T \Sigma_\eta^{-1} X\right)^{-1} X^T \Sigma_\eta^{-1} \left(\boldsymbol{y} - X\boldsymbol{\theta}_0\right),$$

and

$$\Sigma_{\theta|y} = \left(\Sigma_0^{-1} + X^T \Sigma_\eta^{-1} X\right)^{-1}.$$

3.25. Assume that x_n, $n = 1, 2 \ldots, N$, are i.i.d observations from a Gaussian $\mathcal{N}(\mu, \sigma^2)$. Obtain the MAP estimate of μ, if the prior follows the exponential distribution

$$p(\mu) = \lambda \exp\left(-\lambda\mu\right), \quad \lambda > 0,\ \mu \geq 0.$$

Solution: Upon defining $\mathcal{X} := \{x_1, x_2, \ldots, x_N\}$, the posterior distribution is given by

$$p(\mu|\mathcal{X}) \propto p(\mathcal{X}|\mu)p(\mu) = \frac{\lambda \exp\left(-\lambda\mu\right)}{(2\pi)^{N/2}\sigma^N} \prod_{n=1}^{N} \exp\left(-\frac{(x_n - \mu)^2}{2\sigma^2}\right).$$

Taking the ln, differentiating with respect to μ, and equating to zero we obtain

$$\frac{\partial\left(-\lambda\mu - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right)}{\partial\mu} = 0,$$

or

$$-\lambda + \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = 0.$$

Finally,

$$\hat{\mu}_{\text{MAP}} = \frac{\sum_{n=1}^{N} x_n - \lambda\sigma^2}{N},$$

for nonnegative values of the numerator.

3.26. Consider, once more, the same regression model as that of Problem 3.8, but with $\Sigma_\eta = I_N$. Compute the MSE of the predictions $\mathbb{E}[(\mathrm{y}-\hat{\mathrm{y}})^2]$, where y is the true response and $\hat{\mathrm{y}}$ is the predicted value, given a test point $\boldsymbol{x}$ and using the LS estimator,

$$\hat{\boldsymbol{\theta}} = \left(X^T X\right)^{-1} X^T \mathbf{y}.$$

The LS estimator has been obtained via a set of N measurements, collected in the input matrix X and $\mathbf{y}$, where the notation has been introduced previously in this chapter. The expectation $\mathbb{E}[\cdot]$ is taken with respect to to y, the training data, $\mathcal{D}$ and the test points $\mathbf{x}$. Observe the dependence of the MSE on the dimensionality of the space.

(Hint: Consider, first, the MSE, given the value of a test point $\boldsymbol{x}$, and then take the average over all the test points.)

Solution: From the theory, we have that given a point $\boldsymbol{x}$, the LS estimator is given by

$$\hat{\mathrm{y}} = \mathbf{y}^T X(X^T X)^{-1}\boldsymbol{x}.$$

Moreover,

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\theta}}(\boldsymbol{x})) &= \mathbb{E}\left[(\mathrm{y}-\hat{\mathrm{y}})^2\right] = \mathbb{E}\left[(\boldsymbol{\theta}^T\boldsymbol{x}+\eta-\hat{\boldsymbol{\theta}}^T\boldsymbol{x})^2\right] \\
&= \sigma_\eta^2 + \mathbb{E}_{\mathcal{D}}\left[\boldsymbol{x}^T(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{x}\right] + 2\,\mathbb{E}_{\mathcal{D}|\eta}\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{x}\,\mathbb{E}_\eta[\eta]\right] \\
&= \sigma_\eta^2 + \mathbb{E}_{\mathcal{D}}\left[\boldsymbol{x}^T(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{x}\right] \\
&= \sigma_\eta^2 + \left[\boldsymbol{x}^T\Sigma_{\hat{\boldsymbol{\theta}}}\boldsymbol{x}\right] = \sigma_\eta^2 + \sigma_\eta^2\left[\boldsymbol{x}^T(X^TX)^{-1}\boldsymbol{x}\right],
\end{aligned}$$

where the result of Problem 3.8 for the covariance matrix of the LS estimator has been used, i.e.,

$$\Sigma_{\hat{\boldsymbol{\theta}}} = \sigma_\eta^2(X^TX)^{-1}.$$

Also, we used the fact that the LS is an unbiased estimator, hence $\mathbb{E}_{\mathcal{D}|\eta}[\hat{\boldsymbol{\theta}}] = \mathbb{E}_{\mathcal{D}}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$.

We can now make the following approximation, for large values of N:

$$\Sigma := \mathbb{E}_x\left[\mathbf{x}\mathbf{x}^T\right] \approx \frac{1}{N}X^TX,$$

where Σ is the covariance matrix of the (zero mean) input vectors. Then, we have

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\theta}}) &\approx \sigma_\eta^2 + \frac{\sigma_\eta^2}{N}\,\mathbb{E}_x\left[\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right] \\
&= \frac{\sigma_\eta^2}{N}\,\mathbb{E}_x\left[\mathrm{trace}\left\{\Sigma^{-1}\mathbf{x}\mathbf{x}^T\right\}\right] \\
&= \frac{\sigma_\eta^2}{N}\,\mathrm{trace}\left\{\Sigma^{-1}\,\mathbb{E}_x\left[\mathbf{x}\mathbf{x}^T\right]\right\} \\
&= \frac{\sigma_\eta^2}{N}\,\mathrm{trace}\left\{\Sigma^{-1}\Sigma\right\} = \frac{\sigma_\eta^2}{N}l.
\end{aligned}$$

In other words, the MSE is proportional to the dimensionality of the space as well as the variance of the noise, and inversely proportional to the number of data points. That is, for given number of points and noise variance, the error depends on the dimensionality, which is a manifestation of the curse of dimensionality.

Bibliography

[Magn 99] Magnus, J. R., and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons, revised Ed., 1999.

[Papo 02] Papoulis, A., and Unnikrishna, P. *Probability, Random Variables, and Stochastic Processes.* McGraw Hill, 4th Ed., 2002.