Estimation Theory

Zhenrong
March 19, 2021

1 Parameter Estimation

If we have N-point data set $\{x[0], x[1], \ldots, x[N-1]\}$ which depends on an unknown parameter θ and we wish to determine θ based on the data or to define an estimator

$$\hat{\theta} = g(x[0], x[1], \ldots, x[N-1]) \tag{1}$$

where g is the estimation function. This is the problem of parameter estimation.

The mathematical form of data is represented by PDF $p(x[0], x[1], \ldots, x[N-1]; \theta)$, which is parameterized by θ. In classical estimation, the parameter θ are assumed to be deterministic but unknown. In Bayesian estimation, the parameter are no longer deterministic but a random variables and was assigned a PDF. Bayesian estimation helps us to incorporate the prior knowledge of the parameter.

1.1 Unbiased Estimator

For an estimator to be unbiased we mean that on the average the estimator will yield the true value of the unknown parameter. Mathematically, an estimator is unbiased if

$$E(\hat{\theta}) = \theta \quad a < \theta < b, \tag{2}$$

where (a, b) denotes the range of possible values of θ.

1.2 Minimum Variance Criterion

In searching for optimal estimators we need to adopt some optimality criterion. A natural one is the mean square error (MSE), defined as

$$\text{mse}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]. \tag{3}$$

This measures the average mean squared deviation of the estimator from the true value. Rewrite the MSE as

$$\begin{aligned} \text{mse}(\hat{\theta}) &= E\left[(\hat{\theta} - \theta)^2\right], \\ &= E\left\{[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2\right\}, \\ &= \text{var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2, \\ &= \text{var}(\hat{\theta}) + b^2(\theta), \end{aligned} \tag{4}$$

which shows that the MSE is composed of errors due to the variance of the estimator as well as the bias. MSE criterion will leads to unrealizable estimators since the bias term in (4) is a function of unknown parameter θ. An alternative is minimum variance unbiased estimator (MVU).

1.3 Minimum Variance Unbiased Estimator

- Estimator is unbiased.
- Estimator has to have minimum variance

$$\hat{\theta}_{MVU} = \arg\min_{\hat{\theta}}\{\text{var}(\hat{\theta})\} = \arg\min_{\hat{\theta}}\left\{E(\hat{\theta} - E(\hat{\theta}))^2\right\}. \tag{5}$$

1.3.1 Cramer-Rao Lower Bound (CRLB)

Assumed that the PDF $p(\mathbf{x};\theta)$ satisfies the "regularity" condition

$$E\left[\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right] = 0, \text{for all } \theta, \tag{6}$$

where the expectation is taken with respect to $p(\mathbf{x};\theta)$. Then the variance of any unbiased estimater $\hat{\theta}$ is lower bounded by the CLRB, with the variance of the MVU estimator attaining the CLRB. That is:

$$\text{var}(\hat{\theta}) \geq \frac{1}{-E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right]}, \tag{7}$$

and

$$\text{var}\left(\hat{\theta}_{MVU}\right) = \frac{1}{-E\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right]}, \tag{8}$$

where the derivative is evaluated at the true value of θ and the expectation is taken with respect to $p(\mathbf{x};\theta)$. Furthermore, an unbiased estimator may be found that attains the bound for all θ if and only if

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta} = I(\theta)(g(\mathbf{x}) - \theta) \tag{9}$$

for some functions and $I(\theta)$. That estimator, which is the MVU estimator, is $\hat{\theta} = g(\mathbf{x})$ and the minimum variance is $\frac{1}{I(\theta)}$.

2 General MVU Estimation

The evaluation of the CRLB sometimes results in an efficient and hence MVU estimator (§1.3.1). If an efficient estimator does not exist, it is still of interest to be able to find the MVU estimator. To do so requires the concept of sufficient statistics and the important Rao-Blackwell-Lehmann-Scheffe theorem.

2.1 Neyman-Fisher Factorization

- (Neyman-Fisher Factorization) If the PDF $p(\mathbf{x};\theta)$ can be factorized as

$$p(\mathbf{x};\theta) = g(T(\mathbf{x}),\theta)h(\mathbf{x}), \tag{10}$$

  where g is a function depending on $\mathbf{x}$ only through $T(\mathbf{x})$ and h is a function depending only on $\mathbf{x}$, then $T(\mathbf{x})$ is a **sufficient statistic** for θ. Conversely, if $T(\mathbf{x})$ is a sufficient statistic for θ, then the PDF can be factored as in (10).

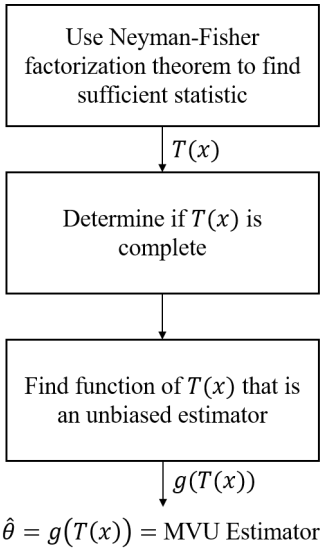


Figure 1: Procedure for finding MVU estimator.

2.2 Rao-Blackwell-Lehmann-Scheffe Theorem

If $\check{\theta}$ is an unbiased estimator of θ and $T(\mathbf{x})$ is a sufficient statistic for θ, then $\hat{\theta} = E(\check{\theta} \mid T(\mathbf{x}))$ is

- a valid estimator for θ (not dependent on θ),
- unbiased,
- of lesser or equal variance than that of $\check{\theta}$, for all θ.

Additionally, if the sufficient statistic is **complete**, then $\hat{\theta}$ is the MVU estimator.

For example:

$$p(x;A) = \underbrace{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}\left(NA^2 - 2A\sum_{n=0}^{N-1}x[n]\right)\right]}_{g(T(\mathbf{x}),A)}\underbrace{\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]}_{h(\mathbf{x})} \tag{11}$$

$T(x)$ is a sufficient and complete statistic. By inspection this is $g(x) = x/N$, which yields

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1}x[n]. \tag{12}$$

as the MVU estimator.

3 Best Linear Unbiased Estimator (BLUE)

In practice, the MVU estimator cannot be found even if it exists. For example, we may not know the PDF of the data. In this case our previous methods, which rely on the CRLB and the theory of sufficient statistics, cannot be applied. Even if the PDF is known, the latter approaches are not guaranteed to produce the MVU estimator. Faced with inability of finding the MVU estimator, it is reasonable to resort to a suboptimal estimator. One of the approach is to restrict the estimator to be linear in the data and find the linear estimator that is unbiased and has minimum variance. This estimator, which is termed the best linear unbiased estimator (BLUE), can be determined with knowledge of only the first and second moments of the PDF.

Finding the BLUE, we want our estimator to be a linear function of the data, that is:

$$\hat{\theta} = \mathrm{Ax}. \tag{13}$$

The requirement is that the estimator be unbiased, that is:

$$E(\hat{\boldsymbol{\theta}}) = \mathbf{A}E(\mathbf{x}) = \theta. \tag{14}$$

which can only be satisfied if:

$$E(\mathbf{x}) = \mathbf{H}\theta. \tag{15}$$

i.e. AH = I.

The BLUE is derived by finding the A which minimises the variance, $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{ACA}^T$, where $\mathbf{C} = E\left[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T\right]$ is the covariance of the data x, subject to the constraint AH = I. Carrying out this minimisation yields the following for the BLUE:

$$\hat{\boldsymbol{\theta}} = \mathbf{Ax} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}, \tag{16}$$

where $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}$.

The form of the BLUE is identical to the MVU estimator for the general linear model. The crucial difference is that the BLUE **does not make any assumptions on the PDF of the data (or noise)** whereas the MVU estimator was derived assuming Gaussian noise. If the data is Gaussian then the BLUE is also the MVU estimator. The BLUE for the general linear model can be stated as follows:

Theorem 3.1 (Gauss-Markov Theorem)**.** Consider a general linear model of the form:

$$\mathrm{x} = \mathrm{H}\theta + \mathrm{w}$$

where **H** is known, and w is noise with covariance C (the PDF of w is otherwise arbitrary), then the BLUE of θ is:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$$

where $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}$ is the minimum covariance.

4 Maximum Likelihood Estimation (MLE)

In some cases the MVU estimator may not exist or it cannot be found by any of the methods discussed so far. The MLE approach is an alternative method in cases where the PDF is known. With MLE the unknown parameter is estimated by maximising the PDF. That is define $\hat{\theta}$ such that:

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{x}; \theta), \tag{17}$$

where x is the vector of observed data.

4.1 Properties

Theorem 4.1 (Asymptotic Properties of the MLE)**.** If the PDF $p(x; \theta)$ of the data x satisfies some "regularity" conditions, then the MLE of the unknown parameter is asymptotically distributed (for large data records) according to

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}\left(\theta, I^{-1}(\theta)\right),$$

where $I(\theta)$ is the Fisher information evaluated at the true value of the unknown parameter.

From the asymptotic distribution, the MLE is seen to be asymptotically unbiased and asymptotically attains the CRLB. It is therefore asymptotically efficient, and hence asymptotically optimal.

5 Least Squares

The MSE, MVU and MLE estimators developed previously required an expression for the PDF $p(\mathbf{x}; \theta)$ in order to estimate the unknown parameter θ in some optimal criterion. The BLUE does not required an expression for the PDF, but the mean and the variance of the data are required and it only works for linear models. An alternative approach is to assume a signal model (rather than probabilistic assumptions about the data) and achieve a design goal assuming this model.

With the Least Squares (LS) approach we assume that the signal model is a function of the unknown parameter θ and produces a signal:

$$s[n] \equiv s(n; \theta), \tag{18}$$

where $s(n; \theta)$ is a function of n and parameterised by θ. Due to noise and model inaccuracies, $w[n]$, the signal $s[n]$ can only be observed as:

$$x[n] = s[n] + w[n]. \tag{19}$$

The "error": $e[n] = x[n] - s[n]$ should be minimised in a least-squares sense with the appropriate choice of θ . That is we choose $\theta = \hat{\theta}$ so that the criterion:

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - s[n])^2 \tag{20}$$

is minimised over the N observation samples of interest and we call this the LSE of θ. More precisely we have:

$$\hat{\theta} = \arg\min_{\theta} J(\theta). \tag{21}$$

and the minimum LS error is given by:

$$J_{\min} = J(\hat{\theta}). \tag{22}$$

6 The Bayesian Philosophy

In stead of the classical approach to statistical estimation in which the parameter θ of interest is assumed to be a deterministic, we assume that θ is a random variable whose particular realization we must estimate. The motivation for doing so is twofold. First, if we have available some prior knowledge about θ, we can incorporate it into our estimator. The mechanism for doing this requires us to assume that θ is a random variable with a given prior PDF. Second, Bayesian estimation is useful in situations where an MVU estimator cannot be found, as for example, when the variance of an unbiased estimator may not be uniformly less than that of all other unbiased estimators. In this instance, it may be true that for most values of the parameter an estimator can be found whose mean square error may be less than that of all other estimators. The resultant estimator can then be said to be optimal "on the average," or with respect to the assumed prior PDF of θ.

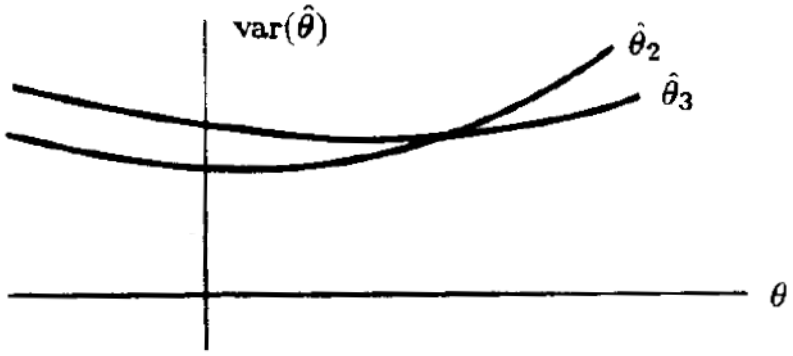


Figure 2: No estimator's variance is uniformly less than that of all other unbiased estimators, so the MVU estimator do not exist.

6.1 Minimum Mean Square Estimator (MMSE)

In the classic approach we derived the MVU estimator by first considering mimisation of the mean square eror, i.e. $\hat{\theta} = \arg\min_{\hat{\theta}} \text{mse}(\hat{\theta})$ where:

$$\text{mse}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = \int (\hat{\theta} - \theta)^2 p(\mathbf{x}; \theta) d\mathbf{x}, \tag{23}$$

and $p(\mathbf{x};\theta)$ is the pdf of x parametrised by θ. In the Bayesian approach we similarly derive an estimator by minimising $\hat{\theta} = \arg\min_{\hat{\theta}} \text{Bmse}(\hat{\theta})$ where:

$$\text{Bmse}(\hat{\theta}) = E\left[(\theta - \hat{\theta})^2\right] = \iint (\theta - \hat{\theta})^2 p(\mathbf{x}, \theta) d\mathbf{x} d\theta, \tag{24}$$

is the Bayesian mse and $p(\mathbf{x}, \theta)$ is the joint pdf of $\mathbf{x}$ and θ (since θ is now a random variable). It should be noted that the Bayesian squared error $(\theta - \hat{\theta})^2$ and classic squared error $(\hat{\theta} - \theta)^2$ are the same. The minimum Bmse $(\hat{\theta})$ estimator or MMSE is derived by differentiating the expression for $\text{Bmse}(\hat{\theta})$ with respect to $\hat{\theta}$ and setting this to zero to yield:

$$\hat{\theta} = E(\theta \mid \mathbf{x}) = \int \theta p(\theta \mid \mathbf{x}) d\theta. \tag{25}$$

where the posterior $pdf, p(\theta \mid \mathbf{x})$, is given by:

$$p(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x}, \theta)}{p(\mathbf{x})} = \frac{p(\mathbf{x}, \theta)}{\int p(\mathbf{x}, \theta) d\theta} = \frac{p(\mathbf{x} \mid \theta) p(\theta)}{\int p(\mathbf{x} \mid \theta) p(\theta) d\theta}. \tag{26}$$

6.2 Maximum A Posteriori (MAP) Estimator

In the MAP estimation approach we choose $\hat{\theta}$ to maximize the posterior PDF or

$$\hat{\theta} = \arg\max_{\theta} p(\theta \mid \mathbf{x}). \tag{27}$$

which was shown to minimize the Bayes risk for a "hit-or-miss" cost function. In finding the maximum of $p(\theta \mid \mathbf{x})$ we observe that

$$p(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \theta) p(\theta)}{p(\mathbf{x})}, \tag{28}$$

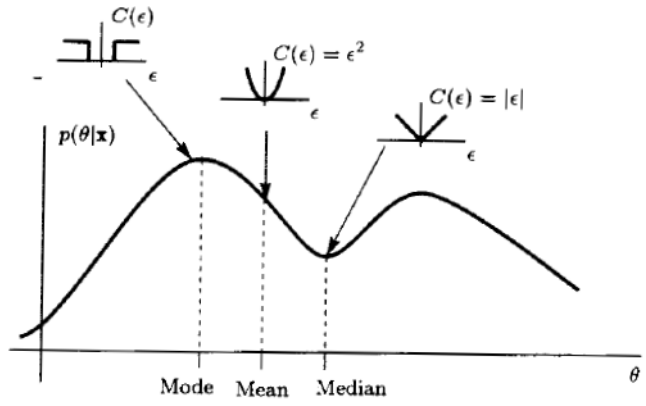
so an equivalent maximization is of $p(\mathbf{x} \mid \theta) p(\theta)$. This is reminiscent of the MLE except for the presence of the prior PDF. Hence, the MAP estimator is

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{x} \mid \theta) p(\theta). \tag{29}$$

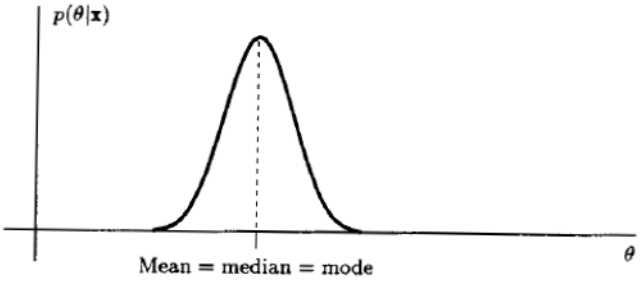
For MAP estimation, as $N \to \infty$, the MAP estimator becomes the Bayesian MLE (if $N \to \infty$ so that the data PDF dominates the prior PDF). If $p(\mathbf{x} \mid \boldsymbol{\theta})$ has the same form as the PDF family $p(\mathbf{x}; \boldsymbol{\theta})$, then the Bayesian MLE and the classical MLE will have the same form.

6.3 Risk Function

In derivation of the MMSE estimator, we minimizing $E\left[(\theta - \hat{\theta})^2\right]$, where the expectation is with respect to the PDF $p(\mathbf{x}, \theta)$. If we let $\epsilon = \theta - \hat{\theta}$ denote the error of the estimator for a particular realization of $\mathbf{x}$ and θ, and also let $\mathcal{C}(\epsilon) = \epsilon^2$, then the MSE criterion minimizes



(a) General posterior PDF



(b) Gaussian posterior PDF

Figure 3: Examples of cost functions and the estimators for different cost functions.

$E[\mathcal{C}(\epsilon)]$. The deterministic function $\mathcal{C}(\epsilon)$ as shown in Figure 3 is termed the cost function. Also, the average cost or $E[\mathcal{C}(\epsilon)]$ is termed the Bayes risk $\mathcal{R}$ or

$$\mathcal{R} = E[\mathcal{C}(\epsilon)]. \tag{30}$$

The estimators that minimize the Bayes risk for the cost functions of Figure 3 are the mean, median, and mode of the posterior PDF. For some posterior PDFs these three estimators are identical (Gaussian posterior PDF). The MAP estimator minimizes the Bayes risk for the "hit-or-miss" cost function is therefore the mode (location of the maximum) of the posterior PDF. For Gaussian posterior PDFs, the MMSE and MAP estimators will be identical.

6.4 Linear MMSE

The optimal Bayesian estimators are difficult to determine in closed form, and in practice too computationally intensive to implement. Although under the jointly Gaussian assumption these estimators are easily found, in general, they are not. To fill this gap we can choose to retain the MMSE criterion but constrain the estimator to be linear.

Consider the class of all linear estimators of the form

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] + a_N. \tag{31}$$

The resultant LMMSE estimator is

$$\hat{\theta} = E(\theta) + \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}(\mathbf{x} - E(\mathbf{x})). \tag{32}$$

Note that it is identical in form to the MMSE estimator for jointly Gaussian $\mathbf{x}$ and θ. This is because in the Gaussian case the MMSE estimator happens to be linear.

7 Application

7.1 Expectation-Maximization Algorithm

Expectation-maximization (EM) is a method of numerically determining the MLE. EM algorithm is guaranteed under certain mild conditions to converge, and at convergence to produce at least a local maximum. It has the desirable property of increasing the likelihood at each step. The EM algorithm exploits the observation that some data sets may allow easier determination of the MLE than the given one.

In general, we suppose that there is a complete to incomplete data transtormation given as

$$\mathbf{x} = \mathbf{g}\left(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M\right) = \mathbf{g}(\mathbf{y}). \tag{33}$$

The function g is a many-to-one transformation. In stead of maximizing $\ln p_x(\mathbf{x};\boldsymbol{\theta})$, we maximize $\ln p_y(\mathbf{y};\boldsymbol{\theta})$. Since y is unavailable, we replace the log-likelihood function by its conditional expectation or

$$E_{y|x}\left[\ln p_y(\mathbf{y};\boldsymbol{\theta})\right] = \int \ln p_y(\mathbf{y};\boldsymbol{\theta}) p(\mathbf{y} \mid \mathbf{x};\boldsymbol{\theta}) d\mathbf{y}. \tag{34}$$

Finally, since we need to know θ to determine $p(\mathbf{y} \mid \mathbf{x};\boldsymbol{\theta})$ and hence the expected loglikelihood function, we use the current guess. Letting θ_k denote the k th guess of the MLE of θ, we then have the following iterative algorithm:

- Expectation (E): Determine the average log-likelihood of the complete data

$$U\left(\boldsymbol{\theta}, \boldsymbol{\theta}_k\right) = \int \ln p_y(\mathbf{y};\boldsymbol{\theta}) p\left(\mathbf{y} \mid \mathbf{x};\boldsymbol{\theta}_k\right) d\mathbf{y}. \tag{35}$$

- Maximization (M): Maximize the average log-likelihood function of the complete data

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} U\left(\boldsymbol{\theta}, \boldsymbol{\theta}_k\right). \tag{36}$$

EM algorithm iteratively decouples the original data set into **several separate data sets**. **The maximization given above corresponds to the MLE of a single data set given by the estimated complete data**. Its disadvantages are the difficulty of determining the conditional expectation in closed form and the arbitrariness in the choice of the complete data. Nonetheless, for the Gaussian problem this method can easily be applied. The problem of estimating parameters in the absence of labels is known as unsupervised learning. The EM algorithm is a kind of unsupervised learning.

7.2 Wiener Filters

Wiener filter is one of the important applications of the LMMSE estimator. We assume that the data $\{x[0], x[1], \ldots, x[N-1]\}$ is WSS with zero mean. As such, the $N \times N$ covariance matrix $\mathbf{C}_{xx}$ takes the symmetric Toeplitz form

$$\begin{aligned}\mathbf{C}_{xx} &= \begin{bmatrix} r_{xx}[0] & r_{xx}[1] & \ldots & r_{xx}[N-1] \\ r_{xx}[1] & r_{xx}[0] & \ldots & r_{xx}[N-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[N-1] & r_{xx}[N-2] & \ldots & r_{xx}[0] \end{bmatrix}, \\ &= \mathbf{R}_{xx}.\end{aligned}$$

where $r_{xx}[k]$ is the autocorrelation function of the $x[n]$ process and $\mathbf{R}_{xx}$ denotes the autocorrelation matrix. Furthermore, the parameter θ to be estimated is also assumed to be zero mean.

Consider the smoothing problem. We wish to estimate $\boldsymbol{\theta} = \mathbf{s} = [s[0]s[1] \ldots s[N-1]]^T$ based on $\mathbf{x} = [x[0]x[1] \ldots x[N-1]]^T$. We will make the reasonable assumption that the signal and noise processes are uncorrelated. Hence,

$$r_{xx}[k] = r_{ss}[k] + r_{ww}[k]. \tag{37}$$

Then, we have

$$\mathbf{C}_{xx} = \mathbf{R}_{xx} = \mathbf{R}_{ss} + \mathbf{R}_{ww}. \tag{38}$$

Also,

$$\mathbf{C}_{\theta x} = E\left(\mathbf{s}\mathbf{x}^T\right) = E\left(\mathbf{s}(\mathbf{s}+\mathbf{w})^T\right) = \mathbf{R}_{ss}. \tag{39}$$

Therefore, the Wiener estimator of the signal is, from (32),

$$\hat{\mathbf{s}} = \mathbf{R}_{ss}\left(\mathbf{R}_{ss} + \mathbf{R}_{ww}\right)^{-1}\mathbf{x}. \tag{40}$$

The $N \times N$ matrix

$$\mathbf{W} = \mathbf{R}_{ss}\left(\mathbf{R}_{ss} + \mathbf{R}_{ww}\right)^{-1}, \tag{41}$$

is referred to as the Wiener smoothing matrix. The corresponding minimum MSE matrix is,

$$\begin{aligned}\mathbf{M}_{\hat{s}} &= \mathbf{R}_{ss} - \mathbf{R}_{ss}\left(\mathbf{R}_{ss} + \mathbf{R}_{ww}\right)^{-1}\mathbf{R}_{ss} \\ &= (\mathbf{I} - \mathbf{W})\mathbf{R}_{ss}.\end{aligned} \tag{42}$$