ECE 830 Fall 2011 Statistical Signal Processing

instructor: R. Nowak

Lecture 20: Bayesian Linear Estimators

1 Linear Minimum Mean Square Error Estimator

Suppose our data is $x \in \mathbb{R}^n$, a random vector governed by a distribution $p(x|\theta)$, which depends on the parameter θ. Moreover, the parameter $\theta \in \mathbb{R}^k$ is treated as a random variable with $\mathbb{E}[\theta] = 0$ and $\mathbb{E}[\theta\theta^T] = \Sigma_{\theta\theta}$. Also, assume that $\mathbb{E}[x] = 0$ and let $\Sigma_{xx} := \mathbb{E}[xx^T]$ and $\Sigma_{\theta x} := \mathbb{E}[\theta x^T]$. Then, as we saw in the previous lecture, the linear filter that provides the minimum MSE is given by:

$$\widehat{A} = \arg\min_{A \in \mathbb{R}^{n\times k}} \mathbb{E}\big[\|\theta - A^T x\|_2^2\big]$$

$$\widehat{A} = \Sigma_{xx}^{-1}\Sigma_{x\theta}$$

The *linear minimum MSE estimator* (LMMSE) estimator is:

$$\widehat{\theta} = \widehat{A}^T x = \Sigma_{\theta x}\Sigma_{xx}^{-1} x.$$

2 Orthogonality Principle

Let $\widehat{\theta} = \Sigma_{\theta x}\Sigma_{xx}^{-1}x$ be the LMMSE estimator, defined above. Then

$$\begin{aligned}\mathbb{E}\big[(\theta - \widehat{\theta})^T x\big] &= \mathbb{E}\big[\mathrm{tr}\big((\theta - \widehat{\theta})x^T\big)\big] \\ &= \mathrm{tr}\big(\Sigma_{\theta x} - \Sigma_{\theta x}\Sigma_{xx}^{-1}\Sigma_{xx}\big) \\ &= 0\ .\end{aligned}$$

In other words, the error $(\theta - \widehat{\theta})$ is orthogonal to the data x. This is shown graphically in Fig. 1.

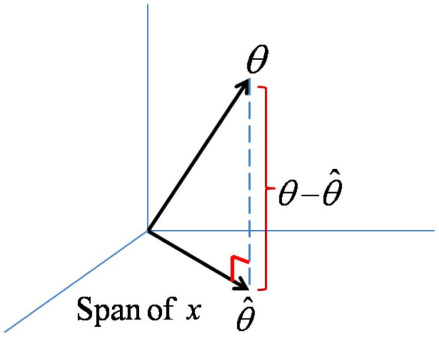

Figure 1: Orthogonality between the estimator $\widehat{\theta}$ and its error $\theta - \widehat{\theta}$.

The orthogonality principle also provides a method for deriving the LMMSE filter. Consider any linear estimator of the form $\widehat{\theta} = B^T x$. If we impose the orthogonality condition

$$0 = \mathbb{E}\big[(\theta - \widehat{\theta})x^T\big] \quad = \quad \Sigma_{\theta x} - B^T\Sigma_{xx}$$

then we see that B^T must be equal to $\Sigma_{\theta x}\Sigma_{xx}^{-1}$.

Example 1 *Linear Signal Model. Suppose we model our detected signal as $x = H\theta + w$, where $x \in \mathbb{R}^n$, $\theta \in \mathbb{R}^k$, $H_{n\times k}$ is a known linear transformation, and w is a noise process. Furthermore assume that*

$$\mathbb{E}[w] = 0,\ \mathbb{E}[ww^T] = \sigma_w^2 I_{n\times n}$$
$$\mathbb{E}[\theta] = 0,\ \mathbb{E}[\theta\theta^T] = \sigma_\theta^2 I_{k\times k}$$

In addition, assume we know that the parameter and the noise process are uncorrelated, i.e., $\mathbb{E}[\theta w^T] = \mathbb{E}[w\theta^T] = 0$. As demonstrated before, the LMMSE estimator is

$$\widehat{\theta} = \Sigma_{x\theta}\Sigma_{xx}^{-1}x$$

where Σ_{xx}^{-1} and $\Sigma_{x\theta}$ can be obtained as follows:

$$\Sigma_{x\theta} = \mathbb{E}[\theta x^T] = \mathbb{E}[\theta(H\theta + w)^T] = \sigma_\theta^2 H^T$$
$$\Sigma_{xx} = \mathbb{E}[xx^T] = \mathbb{E}[(H\theta + w)(H\theta + w)^T] = \sigma_\theta^2 HH^T + \sigma_w^2$$

Therefore, the LMMSE estimator is given by

$$\begin{aligned}\widehat{\theta} &= \sigma_\theta^2 H^T(\sigma_\theta^2 HH^T + \sigma_w^2 I_{n\times n})^{-1}x \\ &= H^T(HH^T + \frac{\sigma_w^2}{\sigma_\theta^2} I_{n\times n})^{-1}x\ .\end{aligned}$$

3 Gauss-Markov Theorem

It is natural to ask when does the LMMSE estimator minimize the Bayes MSE amnog all possible estimators? When is the linear estimar optimal? Based on our previous discussion of Bayesian estimators, we know that the LMMSE esteimator is optimal, i.e., it is the minimum Bayesian MSE estimator, when the posterior mean estimator is linear. This happens to be the case when both data and parameter are modeled as jointly Gaussian.

Theorem 1 Gauss-Markov Theorem. *Let x and y be jointly Gaussian random vectors, whose joint distribution can be expressed as*

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)$$

Then the conditional distribution of y given x is

$$y|x \sim \mathcal{N}\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right).$$

According to the Gauss-Markov Theorem, the posterior mean of the density $p(y|x)$ is a linear function of x, and therefore in this case the minimum MSE estimator is linear. Notice that the matrix $\Sigma_{yx}\Sigma_{xx}^{-1}$ is precisely the optimal A^T derived above (if we take $y = \theta$). In general, when the joint distribution is non-Gaussian, the minimum MSE estimator is *nonlinear*.

Example 2 *Application to the Linear Signal Model. We model the detected signal as $x = H\theta + w$ where $w \sim \mathcal{N}(0, \sigma_w^2 I_{n\times n})$ and $\theta \sim \mathcal{N}(0, \sigma_\theta^2 I_{k\times k})$. Then the vector $[x\ \theta]^T$ is a multivariate Gaussian random vector. As we saw in previous lectures, the Bayesian MSE is minimized by the posterior mean $\mathbb{E}[\theta|x]$ which, in this case, using the Gauss-Markov theorem, is*

$$\begin{aligned}\mathbb{E}[\theta|x] &= \mu_\theta + \Sigma_{\theta x}\Sigma_{xx}^{-1}(x - \mu_x) \\ &= 0 + \sigma_\theta^2 H^T(\sigma_\theta^2 HH^T + \sigma_w^2 I_{n\times n})^{-1}(x - 0) \\ &= \sigma_\theta^2 H^T(\sigma_\theta^2 HH^T + \sigma_w^2 I_{n\times n})^{-1}x\ ,\end{aligned}$$

which is the previously derived LMMSE estimator.

3.1 Proof of the Gauss-Markov theorem

Without loss of generality assume that x and y are zero-mean random vectors. Therefore

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{(2\pi)^{-n/2}(2\pi)^{-n/2}|\Sigma|^{-1}\exp\{-\frac{1}{2}\begin{bmatrix}x^T & y^T\end{bmatrix}\Sigma^{-1}\begin{bmatrix}x\\ y\end{bmatrix}\}}{(2\pi)^{-n/2}|\Sigma_{xx}|^{-1}\exp\{-\frac{1}{2}x^T\Sigma_{xx}^{-1}x\}}$$

where

$$\Sigma = \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}.$$

To simplify the formula we need to determine Σ^{-1}. The inverse can be written as:

$$\begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}^{-1} = \begin{bmatrix}\Sigma_{xx}^{-1} & 0\\ 0 & 0\end{bmatrix} + \begin{bmatrix}-\Sigma_{xx}^{-1}\Sigma_{xy}\\ I\end{bmatrix} Q^{-1}\begin{bmatrix}-\Sigma_{yx}\Sigma_{xx}^{-1} & I\end{bmatrix}$$

where

$$Q := \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}.$$

This formula for the inverse is easily verified by multiplying it by Σ to get the identity matrix. Substituting the inverse into $p(y|x)$ yields

$$p(x|y) = (2\pi)^{-n/2}|Q|^{-1}\exp\{-\frac{1}{2}(y-\Sigma_{yx}\Sigma_{xx}^{-1}x)^T Q^{-1}(y-\Sigma_{yx}\Sigma_{xx}^{-1}x)\}$$

which shows that $y|x \sim \mathcal{N}(\Sigma_{yx}\Sigma_{xx}^{-1}x, Q)$. For the general case when $\mathbb{E}[x] = \mu_x$ and $\mathbb{E}[y] = \mu_y$ then

$$(y-\mu_y)|(x-\mu_x) \sim \mathcal{N}(\Sigma_{yx}\Sigma_{xx}^{-1}(x-\mu_x), Q)$$

$$y|x \sim \mathcal{N}(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x-\mu_x), Q)$$

4 The Wiener Filter

When the expected values of the parameter $\theta \in \mathbb{R}^k$ and the data $x \in \mathbb{R}^n$ are zero, then the Wiener filter A_{opt} is obtained by minimizing the mean square error between the parameter and estimator:

$$A_{opt} = \arg\min_{A\in\mathbb{R}^{k\times n}} \mathbb{E}\big[\|\theta - Ax\|_2^2\big]$$

which results in $A_{opt} = \Sigma_{\theta x}\Sigma_{xx}^{-1}$, involving second order moments and which becomes the optimal estimator when both the data and the parameter are jointly Gaussian distributed.

Example 3 *Signal + Noise Model. We model our detected signal as $x = s + w$ where the noiseless signal s (our parameter) follows a Gaussian distribution $\mathcal{N}(0, \Sigma_{ss})$ and $w \sim \mathcal{N}(0, \Sigma_{ww})$. In addition, s and w are uncorrelated. Therefore, the observation vector $x \sim \mathcal{N}(0, \Sigma_{ss} + \Sigma_{ww})$ and $\mathbb{E}[s\,x^T] = \mathbb{E}[s(s+w)^T] = \Sigma_{ss}$. From here, the LMMSE estimator $\widehat{s}$ becomes:*

$$\widehat{s} \;=\; \Sigma_{ss}(\Sigma_{ss} + \Sigma_{ww})^{-1}x.$$

Now assume that the observation is modeled $x = H\theta + w$ where now $\theta \sim \mathcal{N}(0, \Sigma_{\theta\theta})$ and $w \sim \mathcal{N}(0, \Sigma_{ww})$ where $\theta \in \mathbb{R}^k$ and $w \in \mathbb{R}^n$ (which are uncorrelated) and H is a known $n \times k$ linear transformation matrix. Therefore $x \sim \mathcal{N}(0, H\Sigma_{\theta\theta}H^T + \Sigma_{ww})$. In addition, $\mathbb{E}[\theta\,x^T] = \mathbb{E}[\theta(H\theta + w)^T] = \Sigma_{\theta\theta}H^T$ and the estimator is

$$\widehat{\theta} \;=\; \Sigma_{\theta\theta}H^T(H\Sigma_{\theta\theta}H^T + \Sigma_{ww})^{-1}x.$$

Now suppose that $\Sigma_{\theta\theta} = \sigma_\theta^2 I_{k\times k}$ and $\Sigma_{ww} = \sigma_w^2 I_{n\times n}$. Then

$$\widehat{\theta} \;=\; \sigma_\theta^2 H^T(\sigma_\theta^2 HH^T + \sigma_w^2 I_{n\times n})^{-1}x$$

and the LMMSE estimator of the signal is given by

$$\widehat{s} \;=\; H\widehat{\theta} \;=\; \sigma_\theta^2 HH^T(\sigma_\theta^2 HH^T + \sigma_w^2 I_{n\times n})^{-1}x \;.$$

The matrix HH^T can be diagonalized by an orthonormal transformation U (i.e., HH^T is a symmetric, positive-semidefinite matrix and UDU^T is its eigendecomposition. Since H is rank $k < n$, only the first k elements of the diagonal are nonzero (and non-negative)

$$HH^T \;=\; UDU^T \;=\; U\begin{bmatrix} \lambda_1^2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_k^2 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix} U^T \;.$$

As a consequence

$$\begin{aligned} \widehat{s} &= \sigma_\theta^2 UDU^T(\sigma_\theta^2 UDU^T + \sigma_w^2 I_{n\times n})^{-1}x \\ &= \sigma_\theta^2 UDU^T(\sigma_\theta^2 UDU^T + \sigma_w^2 UU^T)^{-1}x \\ &= \sigma_\theta^2 UDU^T(U\big[\sigma_\theta^2 D + \sigma_w^2 I_{n\times n}\big]U^T)^{-1}x \\ &= \sigma_\theta^2 UD\big[\sigma_\theta^2 D + \sigma_w^2 I_{n\times n}\big]^{-1}U^T x \\ &= U(\sigma_\theta^2 D\big[\sigma_\theta^2 D + \sigma_w^2 I_{n\times n}\big]^{-1})U^T x \;. \end{aligned}$$

Note that the term in parenthesis is the diagonal matrix

$$\begin{bmatrix} \frac{\sigma_\theta^2\lambda_1^2}{\sigma_\theta^2\lambda_1^2+\sigma_w^2} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \frac{\sigma_\theta^2\lambda_k^2}{\sigma_\theta^2\lambda_k^2+\sigma_w^2} & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix}$$

which, as $\sigma_w^2/\sigma_\theta^2$ tends to zero, converges to the identity matrix. Therefore, as the SNR $\to \infty$, the Wiener filter output

$$\widehat{s} \to UU^T x \;=\; P_H x \;,$$

where $P_H = H(H^TH)^{-1}H^T$, the orthogonal projection matrix onto the subspace spanned by H.

4.1 Frequency Domain Wiener Filter

Now let us consider a Wiener filter designed in the frequency domain. Again, our model is $x = s + w$ and now we take the DFT of both sides of this equation. Let U denote the DFT.

$$\widetilde{x} \;=\; U^T x \;=\; U^T s + U^T s \;=\; \widetilde{s} + \widetilde{w}$$

where $\widetilde{x}$, $\widetilde{s}$, and $\widetilde{w}$ denote the DFTs of the observation, signal and noise, respectively.

Let's specify our signal and noise models in the frequency domain as follows. Let $\widetilde{s} \sim \mathcal{N}(0, \Lambda_s)$ and $\widetilde{w} \sim \mathcal{N}(0, \Lambda_w)$. Equivalently $s \sim \mathcal{N}(0, U\Lambda_s U^T)$ and $w \sim \mathcal{N}(0, U\Lambda_w U^T)$. In this case the Wiener filter is given by

$$\begin{aligned}
\widehat{s} &= \Sigma_{ss}(\Sigma_{ss} + \Sigma_{ww})^{-1}x \\
&= U\Lambda_s U^T (U[\Lambda_s + \Lambda_w]U^T)^{-1}x \\
&= U\Lambda_s U^T U[\Lambda_s + \Lambda_w]^{-1}U^T x \\
&= U\Lambda_s[\Lambda_s + \Lambda_w]^{-1}U^T x
\end{aligned}$$

$$\widehat{s} = U \begin{bmatrix}
\frac{\sigma_1^2}{\sigma_1^2+\gamma_1^2} & \dots & 0 & 0 & \dots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & 0 \\
0 & \ddots & \frac{\sigma_i^2}{\sigma_i^2+\gamma_i^2} & 0 & \dots & 0 \\
0 & \ddots & 0 & 0 & \dots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & 0 \\
0 & \dots & 0 & 0 & \dots & \frac{\sigma_n^2}{\sigma_n^2+\gamma_n^2}
\end{bmatrix} U^T x$$

where σ_j^2 and γ_j^2 are the jth diagonal elements of the diagonal matrices Λ_s and Λ_w, respectively. Therefore the filtering process can be synthesized by the following algorithm:

1. Take the DFT of the measured signal.
2. Attenuate each frequency component according to $\frac{1}{1+\mathrm{SNR}_j^{-1}}$ at frequency ω_j, where $\mathrm{SNR}_j = \sigma_j^2/\gamma_j^2$.
3. Take the inverse DFT of the attenuated spectrum.

4.2 Classical derivation of the Wiener Filter

Again, we start with the model $x = s + w$ where x, s, and w are wide-sense stationary processes. We express them as time series

$$x[n] = s[n] + w[n] \ .$$

We aim at defining a filter $h[n]$ that will be convolved with $x[n]$ to etstimate $s[n]$

$$\widehat{s}[n] \ = \ \sum_k h[k]x[n-k] \ .$$

Our filter should minimize the MSE:

$$\begin{aligned}
\mathrm{MSE}(\widehat{s}[n]) &= \mathbb{E}\big[(s[n] - \widehat{s}[n])^2\big] \\
&= \mathbb{E}\big[(s[n]^2 - 2s[n]\textstyle\sum_k h[k]x[n-k] + (\sum_k h[k]x[n-k])^2\big] \ .
\end{aligned}$$

Differentiating with respect to $h[m]$ and making the derivative equal to zero

$$\begin{aligned}
\frac{\partial \mathrm{MSE}(\widehat{s}[n])}{\partial h[m]} &= \mathbb{E}\big[-2s[n]x[n-m] + 2(\sum_k h[k]x[n-k])x[n-m]\big] \\
&= -2R_{sx}[m] + 2(\sum_k h[k]R_{xx}[m-k]) \\
&= -2R_{ss}[m] + 2\bigg(\sum_k h[k](R_{ss}[m-k] + R_{ww}[m-k])\bigg) = 0 \ .
\end{aligned}$$

Therefore the optimal filter satisfies $R_{ss}[m] = \sum_k h[k](R_{ss}[m-k] + R_{ww}[m-k])$, which is the Wiener-Hopf equation. Taking the Discrete-Time Fourier Transform (DTFT) of both sides, we get

$$S_{ss}(\omega) = H(\omega)\big(S_{ss}(\omega) + S_{ww}(\omega)\big)$$

where $S_{ss}(\omega)$ and $S_{ww}(\omega)$ are the power spectra of the signal and the noise process, respectively. Therefore, the frequency response of the Wiener filter is

$$H(\omega) = \frac{S_{ss}(\omega)}{S_{ss}(\omega) + S_{ww}(\omega)} \ .$$

5 Deconvolution

The final topic of this lecture is deconvolution. We model the detected signal as $x = Gs + w$ where G is a circular convolution operator (a blurring transformation, shown in Fig. 2). As in the previous sections $s \sim \mathcal{N}(0, U\Lambda_s U^T)$ and $w \sim \mathcal{N}(0, U\Lambda_w U^T)$. Furthermore, since G is circulant, $G = UDU^T$, where D is a diagonal matrix, which is the frequency response of G.

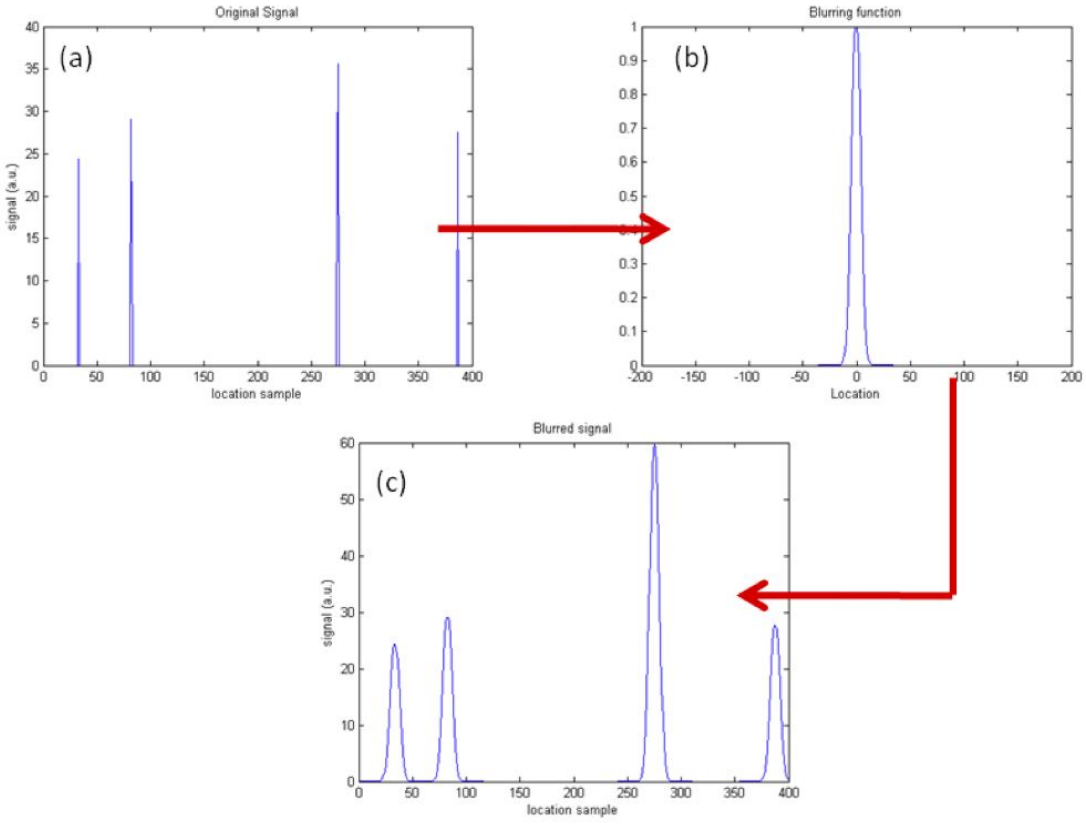

Figure 2: Blurring process. (a) Original impulse signal. (b) Blurring function. (c). Blurred signal

In this case, the Wiener filter solution is computed as follows:

$$\begin{aligned}
\widehat{s} &= \Sigma_{ss} G^T (G\Sigma_{ss} G^T + \Sigma_{ww})^{-1} x \\
&= U\Lambda_s U^T G^T (GU\Lambda_s U^T G^T + U\Lambda_w U^T)^{-1} x \\
&= U\Lambda_s U^T U D^T U^T (UDU^T U\Lambda_s U^T UD^T U^T + U\Lambda_w U^T)^{-1} x \\
&= U\Lambda_s D^T (D\Lambda_s D^T + \Lambda_w)^{-1} U^T x \\
&= U\widetilde{D} U^T x
\end{aligned}$$

where $\widetilde{D}(k,k) = \frac{D^T(k,k)}{|D(k,k)|^2 + P^{-1}(k,k)}$ and $P(k,k) = \frac{\Lambda_s(k,k)}{\Lambda_w(k,k)}$. Do not forget that the transpose operator works as the conjugate transpose operator when the matrix has complex elements.

5.1 Classical Wiener Filter

Following a derivation simmilar to that of Section 4.2. In the case of a blurred, noise time series modeled as

$$x[n] = g[n] * s[n] + w[n]$$

we aim at obtaining a filter $h[n]$ such that the estimator of the deblurred, noisless signal is computed from $\widehat{s}[n] = \sum_k h[k]x[n-k]$. The resulting filter in Fourier domain is:

$$H(\omega) = \frac{G^*(\omega) S_{ss}(\omega)}{|G(\omega)|^2 S_{ss}(\omega) + S_{ww}(\omega)}$$

where $G(\omega)$ is the transfer function of the blurring filter $g[n]$ and G^* is its complex conjugate.