**ECE 830 Fall 2010 Statistical Signal Processing**

**instructor:** R. Nowak , **scribe:** I. Rosado-Mendez

Lecture 19: Bayesian Linear Estimators

# 1 Linear Minimum Mean-Square Estimator

Suppose our data is set $X \in \mathbb{R}^n$ , which is considered to be a random vector governed by a distribution $p(x|\theta)$, which depends on the parameter $\theta$. 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 best linear estimator of $\theta$ is given by:

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

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

As a consequence, our **Linear Minimum Mean Square Error Estimator** or **LMMSE** estimator becomes:

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

# 2 Orthogonality Principle

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

$$\mathbb{E}\big[(\theta - \hat{\theta})^T x\big] = \mathbb{E}\big[tr\big(\theta - \hat{\theta})x^T\big)\big]$$

$$\mathbb{E}\big[(\theta - \hat{\theta})^T x\big] = tr\big(\Sigma_{\theta x} - \Sigma_{\theta x}\Sigma_{xx}^{-1}\Sigma_{xx}\big)$$

$$\mathbb{E}\big[(\theta - \hat{\theta})^T x\big] = 0$$

In other words, the error $(\theta - \hat{\theta})$ is orghogonal to the data $x$. This is shown graphically in Fig. 1. It is important to note that if $\hat{\theta} = A^T x$, then, by the orthogonality principle

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

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

which is the previously derived LMMSE estimator.

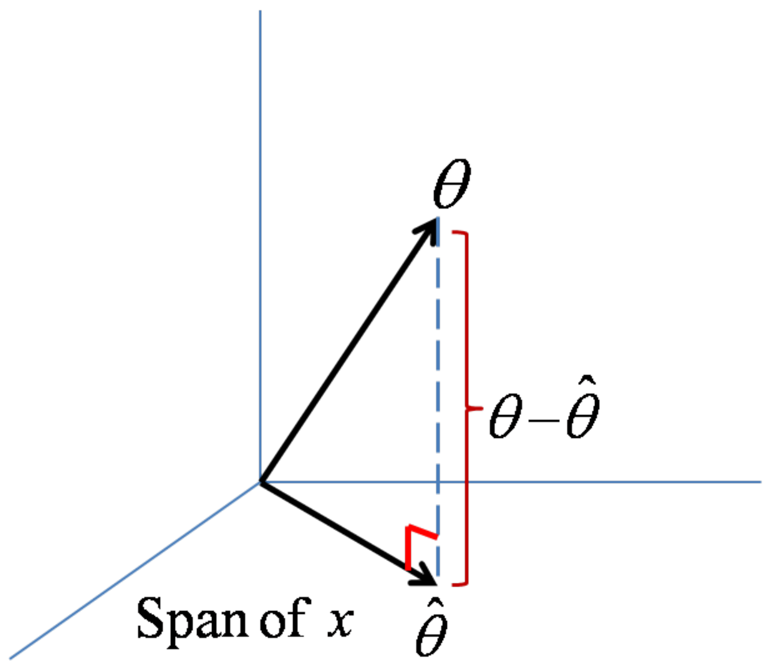


Figure 1: Orthogonality between the estimator $\hat{\theta}$ and its error $\theta - \hat{\theta}$.

## 2.1 Linear signal model

Suppose we model our detected signal as $X = H\theta + W$, where $X$ and $W \in \mathbb{R}^n$, $\theta \in \mathbb{R}^k$, $H_{n\times k}$ is a known linear transformation, and $W$ is a noise process. Furhtermore we know 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, 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

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

where $\Sigma_{xx}^{-1}$ 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

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

When does the LMMSE estimator minimize the Bayes MSE amog all possible estimators? When is the linear estimar optimal? **The LMMSE esteimator is optimal, i.e., it is the minimum Bayesian MSE estimator when the Maximum A Posteriori estimator is linear**.

# 3 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

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

## 3.1 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 multivarian 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

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

which is the previously derived LMMSE estimator. Therefore, **linear estimators are optimal in the Gaussian case**.

## 3.2 Proof of the Gauss-Markov theorem

Without loss of generality assume that $X$ and $Y$ are zero-mean random vectors. Therefore

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

where

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

To simplify the formula we need to determine $\Sigma^{-1}$. 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 $\Sigma$ 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:

$$A_{opt} = \underset{A:\hat{\theta}=Ax}{\arg\min}\ \mathbb{E}\big[\|\theta - Ax\|_2^2\big]$$

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.

## 4.1 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 data vector $X \sim \mathcal{N}(0, \Sigma_{ss} + \Sigma_{ww})$ and $\mathbb{E}[sx^T] = \mathbb{E}[s(s+w)^T] = \Sigma_{ss}$. From here, the LMMSE estimator $\hat{s}$ becomes:

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

## 4.2 Linear Signal + Noise Model

Now we assume that the detected signal can be model as $X = H\theta + W$ 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 an $n \times k$ 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 becomes

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

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

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

and the LMMSE estimator of the noisless signal becomes

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

In some cases $HH^T$ can be diagonalized under an orthonormal transformation $U$ (its columns are orthonormal to each other) in such a way that only the first $k$ elements of the diagonal are nonzero

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

As a consequence

$$\hat{s} = \sigma_\theta^2 UDU^T(\sigma_\theta^2 UDU^T + \sigma_w^2 I_{n\times n})^{-1}x$$

$$\hat{s} = \sigma_\theta^2 UDU^T(\sigma_\theta^2 UDU^T + \sigma_w^2 UU^T)^{-1}x$$

$$\hat{s} = \sigma_\theta^2 UDU^T(U\big[\sigma_\theta^2 D + \sigma_w^2 I_{n\times n}\big]U^T)^{-1}x$$

$$\hat{s} = \sigma_\theta^2 UD\big[\sigma_\theta^2 D + \sigma_w^2 I_{n\times n}\big]^{-1}U^Tx$$

$$\hat{s} = U(\sigma_\theta^2 D\big[\sigma_\theta^2 D + \sigma_w^2 I_{n\times n}\big]^{-1})U^Tx$$

Note that the term in parenthesis reduces to a diagonal matrix of the form

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

which, as $\sigma_w^2/\sigma_\theta^2$ tends to zero, converges to

$$\begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix}$$

and $\hat{s} \longrightarrow P_H x$.

## 4.3 Frequency Domain Wiener Filter

In tihs part we review the signal + noise example (4.1) but approaching the problem from Fourier domain. Again, our model is $X = S + W$ and now we take the DFT of both sides:

$$U^T X = U^T S + U^T W$$
$$\tilde{X} = \tilde{S} + \tilde{W}$$

where $\tilde{S} \sim \mathcal{N}(0, \Lambda_s)$ and $\tilde{W} \sim \mathcal{N}(0, \Lambda_w)$. Equivalently

$$X = UU^T S + UU^T W = S + W$$

so $S \sim \mathcal{N}(0, U\Lambda_s U^T)$ and $W \sim \mathcal{N}(0, U\Lambda_w U^T)$
Therefore, the Wiener-Fielter estimator becomes

$$\hat{s} = \Sigma_{ss}(\Sigma_{ss} + \Sigma_{ww})^{-1} x$$
$$\hat{s} = U\Lambda_s U^T (U[\Lambda_s + \Lambda_w]U^T)^{-1} x$$
$$\hat{s} = U\Lambda_s U^T U[\Lambda_s + \Lambda_w]^{-1} U^T x$$
$$\hat{s} = U\Lambda_s[\Lambda_s + \Lambda_w]^{-1} U^T x$$

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

where $\sigma_j^2$ and $\gamma_j^2$ are the $j^th$ elements of the diagonal matrices $\Lambda_s$ and $\Lambda_w$, 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 $\omega_j$, where $\mathrm{SNR}_j = \sigma_j^2/\gamma_j^2$.
3. Take the inverse DFT of the attenuated spectrum.

## 4.4 Classical derivation of the Wiener Filter

Again, we start with the model $X = S+W$ where $X, S, W$ are wide-sense stationary processes. We re-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 $s[n]$

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

Our filter should minimize the MSE:

$$\mathrm{MSE}(\hat{s}[n]) = \mathbb{E}\big[(s[n] - \hat{s}[n])^2\big]$$

$$\mathrm{MSE}(\hat{s}[n]) = \mathbb{E}\big[(s[n]^2 - 2s[n]\sum_k h[k]x[n-k] + (\sum_k h[k]x[n-k])^2\big]$$

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

$$\frac{\partial \mathrm{MSE}(\hat{s}[n])}{\partial h[m]} = \mathbb{E}\big[-2s[n]h[m] + 2(\sum_k h[k]x[n-k])x[n-m]\big]$$

$$\frac{\partial \mathrm{MSE}(\hat{s}[n])}{\partial h[m]} = -2R_{sx}[m] + 2(\sum_k h[k]R_{xx}[m-k])$$

$$\frac{\partial \mathrm{MSE}(\hat{s}[n])}{\partial h[m]} = -2R_{ss}[m] + 2\left(\sum_k h[k](R_{ss}[m-k] + R_{ww}[m-k])\right) = 0$$

Therefore the optimal filter satisfies $R_{ss}[m] = \sum_k h[k](R_{ss}[m-k] + R_{ww}[m-k])$, which is just the familiar Wiener-Hopf equation. Taking the DFT 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 filter becomes:

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

# 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 basically the frequency response of $G$.
In this case, the Wiener filter solution is computed as follows:

$$\hat{s} = \Sigma_{ss}G^T(G\Sigma_{ss}G^T + \Sigma_{ww})^{-1}x$$

$$\hat{s} = U\Lambda_s U^T G^T(GU\Lambda_s U^T G^T + U\Lambda_w U^T)^{-1}x$$

$$\hat{s} = U\Lambda_s U^T UD^T U^T(UDU^T U\Lambda_s U^T UD^T U^T + U\Lambda_w U^T)^{-1}x$$

$$\hat{s} = U\Lambda_s D^T(D\Lambda_s D^T + \Lambda_w)^{-1}U^T x$$

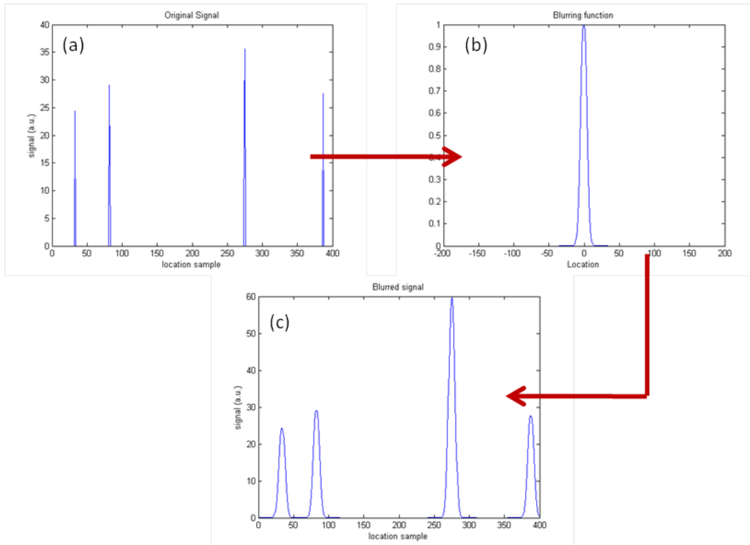


Figure 2: Blurring process. (a) Original impulse signal. (b) Blurring function. (c). Blurred signal

$$\hat{s} = U\tilde{D}U^T x, \text{ where } \tilde{D}_{kk} = \frac{D_{kk}^T}{|D_{kk}|^2 + P_{kk}^{-1}} \text{ and } P_{kk} = \frac{\Lambda_s(k,k)}{\Lambda_w(k,k)}$$

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 4.4, 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 $\hat{s}[n] = \sum_k h[k]x[n-k]$. 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.