Notes on Non-Random Parameter Estimation

Çağatay Candan

January 6, 2011

Abstract

Brief notes on the basics of non-random parameter estimation are presented. These notes have been prepared for the EE 503 course (METU, Electrical Engin.) whose major emphasis is the Bayesian estimation. Therefore the notes may not be as complete as you may wish.

1 Non-Random Parameter Estimation

We examine the following model to illustrate the problem:

$$r = s + n \tag{1}$$

Here n is a random variable showing the contribution of noise whose statistics are at least partially known. The parameter s is a *deterministic* constant which we call, following the terminology of Van Trees [1], as a *non-random parameter*. The variable r is the observation given to us which is also a random quantity due to the additive noise term.

Previously we have considered general Bayesian estimators and also the special case of linear minimum mean square error (MMSE) estimators. Different from the MMSE problem, here we do not have any a-priori information on s. In the linear MMSE problem, first two moments of s and its cross-correlation with the observations are utilized to construct an estimator that is optimal in the sense of minimizing $E\{|s-\widehat{s}|^2\}$. Here by the nature of observations or due to insufficient a-priori information, we do not have any statistical information on s. In spite of this, we proceed similar to the MMSE case and try to produce an estimator that minimizes $E\{(s-\widehat{s})^2\}$.

Minimization of $E\{(s-\widehat{s})^2\}$: The variable $\widehat{s}$ shows the estimate that is produced from the observation. Stated differently, the estimate is a function of the observation r, $\widehat{s} = g(r)$. The goal is to establish a good $g(r)$ function so that the cost is minimized.

It is clear that the estimate $\widehat{s} = g(r)$ is also a random variable. We denote the mean of $\widehat{s}$ with $\eta_{\widehat{s}}$. We can now express the cost function as follows:

$$\begin{aligned} E\{(s-\widehat{s})^2\} &= E\{((s-\eta_{\widehat{s}})-(\widehat{s}-\eta_{\widehat{s}}))^2\} \\ &= E\{(s-\eta_{\widehat{s}})^2 - 2(s-\eta_{\widehat{s}})(\widehat{s}-\eta_{\widehat{s}}) + (\widehat{s}-\eta_{\widehat{s}})^2\} \\ &= \underbrace{E\{(s-\eta_{\widehat{s}})^2\}}_{(s-\eta_{\widehat{s}})^2} - 2(s-\eta_{\widehat{s}})\underbrace{E\{(\widehat{s}-\eta_{\widehat{s}})\}}_{0} + E\{(\widehat{s}-\eta_{\widehat{s}})^2\} \\ &= (s-\eta_{\widehat{s}})^2 + E\{(\widehat{s}-\eta_{\widehat{s}})^2\} \\ &= \text{bias}^2 + \text{estimator-variance} \end{aligned} \tag{2}$$

The last equation shows that the mean square estimation error is a function of the unknown parameter s. This is due to the first term in the MSE relation, i.e. bias^2 term. In the random parameter estimation problem (that is when the parameter s is random), the expectation in (2) is taken with respect to both r and s and this removes the dependence of MSE on the realization of s.

One may proceed as follows:

1. Take $\widehat{s} = c$ for all observations. Then the estimator-variance is zero and the MSE is $(s-c)^2$.

   Comments: This approach eliminates error-variance at the cost of bias. There is no effort done to eliminate the bias which can be significant in many problems. The suggested estimator is a realizable estimator but its performance is questionable to say the least. An estimator making use of the observations would be more pleasing.

2. Choose $\widehat{s} = g(r) = s$ for all observations. The estimator maps the observation to the correct s value. The bias, estimator-variance and the overall MSE is zero.

   Comments: This is the ideal estimator. It reduces the error to zero, but the estimator is not realizable. (It depends on the unknown parameter s.) This is the degenerate solution of the optimization problem and this solution is not realizable.

3. Choose $\widehat{s}$ such that it is guaranteed to be unbiased, then try to minimize MSE (estimator variance) under this constraint.

   Comments: This approach, if it is feasible, guarantees that bias = 0 and proceeds to minimize the estimator-variance under the zero bias constraint. This is indeed a reasonable approach. In these notes, we examine the feasibility of this approach and construction of the estimator for the linear observation model. (More information about the unbiased minimum variance estimators is given in [2, chap.2].)

We should remember that the the linear (not affine) MMSE estimator for *random parameters* can be biased. The amount of bias can be calculated and it is possible to subtract the calculated bias off from every estimate making the estimator unbiased.

To understand the significance of bias a little further, let's consider a simple communication channel such as $r = s + n$. Here s is +1 or -1 with equal probability. ($E\{s\} = 0$ and its variance is 1.) Since we have information on the joint statistics of s and n, we can construct an estimator for s. Let's say that this estimator produces either +1 or -1 with equal probability. Then $E\{\widehat{s}\} = 0$ showing that the estimator is unbiased. The question on the goodness of this estimator is not yet answered. The unbiasedness condition says that the center of mass of the histogram of the estimates and the quantity to be estimated are the same. As an example, lets say that the estimator produces wrong estimates all the time, that is its probability of error is 100%. In this case, it is still unbiased; but its MSE is the largest possible value.

For the same problem, we may choose to interpret s in $r = s + n$ as a non-random parameter. In this case, the unbiasedness condition guarantees that the estimates are on the average on the correct s, that is the center of mass of the estimate histogram is on the value that we are looking for. This is a minor but an important difference in the bias considerations.

2 Maximum Likelihood Estimation

The maximum likelihood (ML) estimation does not aim to minimize a cost over the observation space, i.e. it does not aim to minimize a cost such as $E\{(s - \widehat{s})^2\}$. Instead, the aim is the solution of the following problem:

$$\widehat{s} = \arg\max_{s} f_r(r; s) \tag{3}$$

Here $f_r(r)$ is the density function of the observation and s is the non-random parameter that we are trying to estimate. The function $f_r(r; s)$ should not be interpreted as a conditional density such as $f_r(r|A)$ where A is an event. The function $f_r(r; s)$ should be interpreted as an ordinary function with a parameter s whose value is not known. With this interpretation, it is clear that the maximum likelihood estimation problem is a deterministic optimization problem.

It should also be noted that we are not maximizing the probability of finding correct s. The parameter s is a non-random quantity, therefore it does not have an associated probability space. For this reason, the approach is called maximum likelihood not maximum probability.

We study the following problem to illustrate the maximum likelihood procedure. We assume that a cosine function with unknown amplitude (α) and frequency (ω where $0 < \omega < 1/2$) is observed under additive white Gaussian noise:

$$r[n] = s[n] + v[n] = \alpha \cos(2\pi\omega n) + v[n], \qquad n = \{0, \ldots, N-1\}$$

Here $v[n]$ is zero-mean i.i.d. Gaussian distributed random variables with variance σ_v^2. The average signal power is $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} s^2[n] = \frac{\alpha^2}{2}$ provided that ω is not close to 0 or 1 (If that is the case, the average signal power becomes α^2.) The SNR is defined as the ratio of average signal power over σ_v^2.

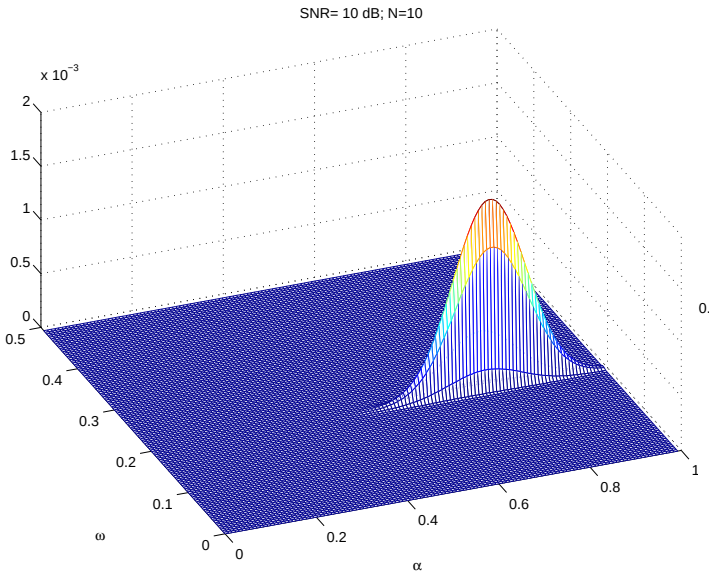


Figure 1: SNR = 10 dB, $N = 10$. The cost function has some spread around the true values.

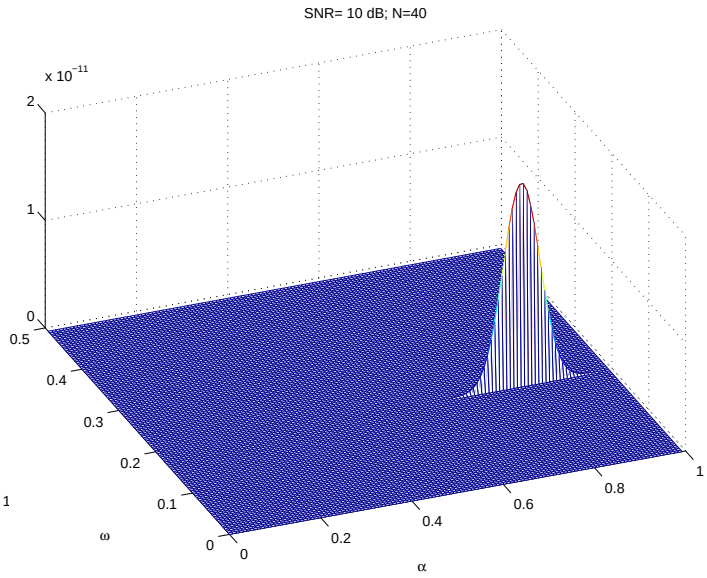


Figure 2: SNR = 10 dB, $N = 40$. The cost function has less spread around the true values in comparison to the SNR = 10 dB, $N = 10$ case.

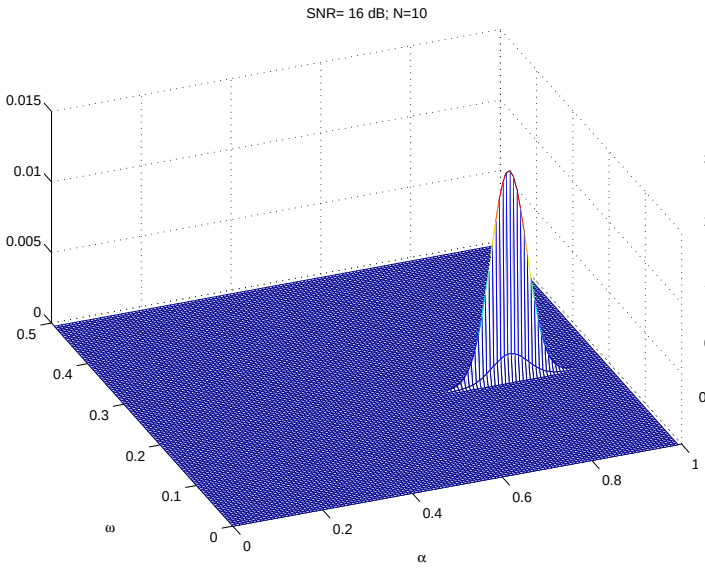


Figure 3: SNR = 16 dB, $N = 10$. The cost function has almost the same spread around the true values when compared with the SNR = 10 dB, $N = 40$ case. (Why ?)

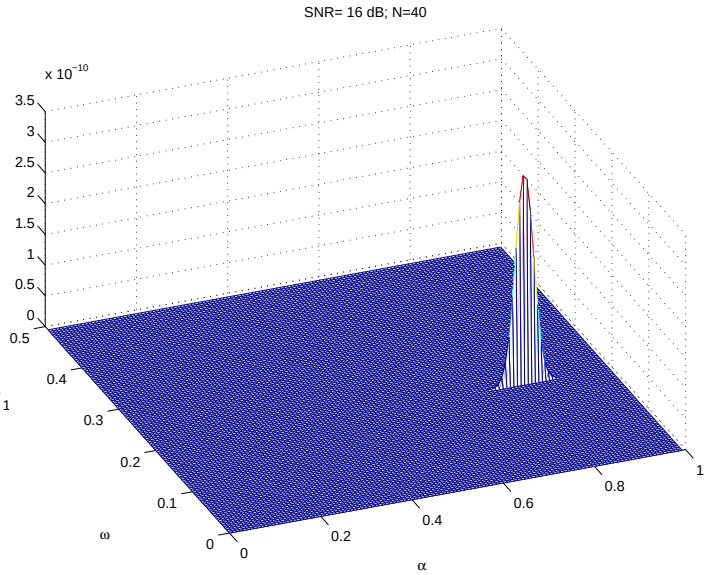


Figure 4: SNR = 16 dB, $N = 40$. The cost function has the least spread around the true values.

The density of $f_{\mathbf{r}}(\mathbf{r};\alpha,\omega)$ can be written as follows:

$$f_{\mathbf{r}}(\mathbf{r};\alpha,\omega)=\frac{1}{(2\pi\sigma_v^2)^{\frac{N}{2}}}\exp\left(\frac{-1}{2\sigma_v^2}\sum_{n=0}^{N-1}(r[n]-\alpha\cos(2\pi\omega n)^2\right) \tag{4}$$

We treat the density function as an ordinary function and by taking the partial derivatives of (4) with respect to α and ω, it is possible to find the peak points of the cost function. Note that the $r[n]$ values in the cost function are the observations which are provided to us, therefore the only unknowns in this function are α and ω. For this problem ω is periodic with 1, that is if ω_x is a peak location, so is ω_x+k (k is an integer). Furthermore cosine function is an even function, therefore, $-\omega_x$ and $-\omega_x+k$ are also peak locations. Because of this reason, we choose to the range of $0\leq\omega_x\leq\frac{1}{2}$ and $-\infty<\alpha<\infty$ for the search domain of ω and α.

In these notes, we do not pursue the mechanics of the optimization process, but try to understand the estimation procedure. Figures 1 to 4 show the objective function in the optimization as a function of α and ω at different operational conditions. For the generation of these figures, an $r[n]$ sequence of length N at an operational SNR is generated. In all figures, the true α and ω values are 0.8 and 0.2 respectively. The coordinates of the peak observed in all figures are the ML estimates of α and ω for that $r[n]$ sequence. (You can examine the Matlab code given in Appendix A.2 for the details of search process for ML estimation.) It can be noted that ML estimate works well for the examples shown here. It can be also noted that frequency estimation is much more accurate than the amplitude estimate. (In these notes, we do not provide any details on the performance of ML procedure. Interested students can examine [2] and other texts for this important topic. This topic is mostly discussed in EE535 in our curriculum.)

The ML estimate in general has no known optimality properties. But in some special cases, it is known to be the efficient estimator. (The efficient estimator is the estimator meeting the Cramer-Rao lower bound. [1, 2].) Also, it is known that (under fairly general conditions) as SNR $\rightarrow\infty$ and/or $N\rightarrow\infty$, the ML estimate is unbiased, efficient and normal distributed. Hence the ML estimate is asymptotically unbiased and efficient.

3 Unbiased, Minimum Variance Estimators For the Linear Observation Model

We examine the details of a setting frequently emerging in applications. The non-random vector $\mathbf{x}$ of the dimensions $L\times 1$ is observed with a linear observation system as shown below:

$$\mathbf{r}=\mathbf{H}\mathbf{x}+\mathbf{v} \tag{5}$$

The matrix $\mathbf{H}$ is a $N\times L$ matrix denoting the observation system. The vector $\mathbf{v}$ is due to noise. For the unbiased minimum variance estimators, it is assumed that $N\geq L$, that is the number of observations is greater than the number of unknowns. (This condition is not necessary for Bayesian estimators.) The error vector is assumed to be zero-mean and has the auto-correlation matrix of $\mathbf{R_v}$. An estimator in the form of

$$\widehat{\mathbf{x}}=\mathbf{K^H}\mathbf{r} \tag{6}$$

is to be designed to minimize the total estimation error variance, $E\{||\mathbf{x}-\widehat{\mathbf{x}}||^2\}$. (As we have noted earlier, the problem is not well defined under this general setting. We introduce the unbiasedness condition shortly which makes the solution realizable.)

The error can be expressed as shown below:

$$\begin{aligned}
E\{||\mathbf{x}-\widehat{\mathbf{x}}||^2\} &= E\{||(\mathbf{I}-\mathbf{K^H}\mathbf{H})\mathbf{x}-\mathbf{K^H}\mathbf{v}||^2\}\\
&= ||(\mathbf{I}-\mathbf{K^H}\mathbf{H})\mathbf{x}||^2+E\{||\mathbf{K^H}\mathbf{v}||^{\mathbf{2}}\}\\
&\overset{(a)}{=} ||(\mathbf{I}-\mathbf{K^H}\mathbf{H})\mathbf{x}||^2+E\{\mathrm{trace}\{\mathbf{K^H}\mathbf{v}\mathbf{v^H}\mathbf{K}\}\}\\
&= ||(\mathbf{I}-\mathbf{K^H}\mathbf{H})\mathbf{x}||^2+\mathrm{trace}\{\mathbf{K^H}\mathbf{R_v}\mathbf{K}\}
\end{aligned} \tag{7}$$

In the line shown with (a) the identity trace$\{\mathbf{AB}\}$ = trace$\{\mathbf{BA}\}$ is utilized. The first term in the last equation depends on $\mathbf{x}$ and it is eliminated if $\mathbf{K^H H = I}$ condition is satisfied. As expected, with this condition $E\{\widehat{\mathbf{x}}\} = \mathbf{x}$, therefore the estimator becomes unbiased.

It should be noted that $\mathbf{I}$ is a $L \times L$ identity matrix and $\mathbf{K^H}$ is a $L \times N$ (fat and short) matrix. If $L = N$ (the number of observations is equal to the number of unknowns), $\mathbf{K}$ matrix is unique and it is $\mathbf{K^H = H^{-1}}$, assuming $\mathbf{H}$ is invertible. If $N > L$, there are infinitely many $\mathbf{K}$ matrices satisfying the condition $\mathbf{K^H H = I}$. (Why?)

The problem of minimizing $E\{||\mathbf{x} - \widehat{\mathbf{x}}||^2\}$ for the unbiased estimator can be expressed as follows:

$$\min_{\mathbf{K}} \text{trace}\{\mathbf{K^H R_v K}\} \quad \text{such that} \quad \mathbf{K^H H = I} \tag{8}$$

It should be noted that $\mathbf{R_v}$ matrix is positive semi-definite therefore the optimization problem is well defined. The same optimization problem can be written as follows:

$$\min_{\mathbf{K}} \sum_{j=1}^{L} \mathbf{k_j^H R_v k_j} \quad \text{such that} \quad \mathbf{K^H H = I} \tag{9}$$

In the last relation $\mathbf{k_j^H}$ vector is the j'th row of the $\mathbf{K^H}$ matrix. If we focus on the optimization of $\mathbf{k_j^H}$, that is the estimator for the j'th unknown variable; then the optimization problem becomes

$$\min_{\mathbf{k_j}} \mathbf{k_j^H R_v k_j} \quad \text{such that} \quad \mathbf{k_j^H H = e_j^H} \tag{10}$$

or

$$\min_{\mathbf{k_j}} \mathbf{k_j^H R_v k_j} \quad \text{such that} \quad \mathbf{H^H k_j = e_j} \tag{11}$$

Here $\mathbf{e_j}$ denotes the canonical $L \times 1$ vector whose j'th entry is 1 and the rest are 0. (In the very last equation, we have taken the Hermitian of the constraint equation.)

The last problem can be recognized as the minimum norm solution problem for the under-determined equation system of $\mathbf{H^H k_j = e_j}$ under the weighted Euclidean norm of $\mathbf{k_j^H R_v k_j}$. The minimum norm solution for $\mathbf{k_j}$ is

$$\mathbf{k_j = R_v^{-1} H (H^H R_v^{-1} H)^{-1} e_j} \tag{12}$$

This concludes the derivation for $\mathbf{k_j}$. Readers can find the details of the minimum norm solution in Appendix A.1.

If we return to the original problem given in (9), it can be noted that the individual optimization of $\mathbf{k_j}$ vectors is sufficient to optimize the overall cost function. That is the cost function function can be expressed as the summation of L smaller dimensional cost functions such as the one given in (11).

Finally by taking the Hermitian of $\mathbf{k_j}$ given in (12), we can get the j'th row of $\mathbf{K^H}$ matrix. By concatenating the rows of $\mathbf{K^H}$ matrix, we get the unbiased minimum variance estimator as follows:

$$\widehat{\mathbf{x}} = \mathbf{K^H r} = \underbrace{(\mathbf{H^H R_v^{-1} H})^{-1} \mathbf{H^H R_v^{-1}}}_{\mathbf{K^H}} \mathbf{r} \tag{13}$$

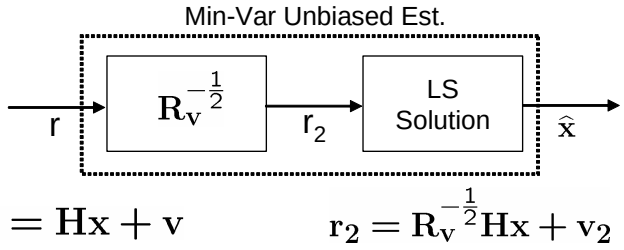
It should be noted that the minimum variance estimator is nothing but the the weighted least square solution of the equation system $\mathbf{r = Hs}$ where the norm is calculated according to $||\mathbf{e}||^2 = \mathbf{e^H R_v^{-1} e}$.

The error covariance matrix for the minimum variance unbiased estimator can be found as follows:

$$\begin{aligned}
E\{\mathbf{ee^H}\} &= E\{(\mathbf{K^H v})(\mathbf{K^H v})^\mathbf{H}\} \\
&= E\{\mathbf{K^H v v^H K^H}\} \\
&= \mathbf{K^H R_v K} \\
&= (\mathbf{H^H R_v^{-1} H})^{-1} \underbrace{\mathbf{H^H R_v^{-1} R_v R_v^{-1} H} (\mathbf{H^H R_v^{-1} H})^{-1}}_{\mathbf{I}} \\
&= (\mathbf{H^H R_v^{-1} H})^{-1}
\end{aligned} \tag{14}$$

An important special case is the case of white noise, that is when $\mathbf{R_v} = \sigma_v^2\mathbf{I}$. Under this condition, the estimate is $\hat{\mathbf{x}} = \mathbf{K^H r} = (\mathbf{H^H H})^{-1}\mathbf{H^H r}$, i.e the classical least solution. Note that the estimate does not depend on noise variance. Under this condition the error covariance matrix is $\sigma_v^2(\mathbf{H^H H})^{-1}$.

As a last note, we would like to mention that the minimum variance, unbiased estimator for the linear observations can be interpreted as the least square solution of the whitened input as shown in Figure 5.



$$\mathbf{r = Hx + v} \qquad \mathbf{r_2 = R_v^{-\frac{1}{2}} Hx + v_2}$$

Figure 5: Minimum variance unbiased estimator interpreted as the whitened least square solution.

The first block, shown in Figure 5, whitens the process using $\mathbf{R_v^{-\frac{1}{2}}}$ matrix. In an earlier part of the course, we have studied different methods of whitening or decorrelating random vectors. The $\mathbf{R_v^{-\frac{1}{2}}}$ can be calculated using the methods discussed in that part of the course.

The relation between the desired vector $\mathbf{x}$ and the whitened observations (output of $\mathbf{R_v^{-\frac{1}{2}}}$ block) can be written as

$$\mathbf{r_2} = \underbrace{\mathbf{R_v^{-\frac{1}{2}} H}}_{\mathbf{H_2}}\mathbf{x} + \mathbf{v_2}. \tag{15}$$

Calling $\mathbf{H_2} = \mathbf{R_v^{-\frac{1}{2}} H}$, we can immediately write the least square solution as $\hat{\mathbf{x}} = (\mathbf{H_2^H H_2})^{-1}\mathbf{H_2^H r_2}$, which is identical to the estimator given in (13). This result shows that the whitened LS solution is the minimum variance, unbiased estimator for the linear observation model.

A Appendices

A.1 Minimum Norm Solution of Underdetermined Equation System

The problem is finding the solution of an under-determined equation of $\mathbf{Ax} = \mathbf{b}$ that minimizes the cost of $||\mathbf{x}||_{\mathbf{R}}^2 = \mathbf{x^H R x}$. (The dimensions of the matrix $\mathbf{A}$ is assumed to be $L \times N$ where $L < N$.)

The problem can be expressed as follows:

$$\min \mathbf{x^H R x} \quad \text{such that} \quad \mathbf{Ax = b} \tag{16}$$

The constraint optimization problem can be formulated as unconstrainted optimization problem with the Lagrange multipliers.

$$J(\mathbf{x}) = \mathbf{x^H R x} + \boldsymbol{\lambda}^T(\mathbf{Ax - b}) \tag{17}$$

Here $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers, that is $\boldsymbol{\lambda} = [\lambda_1\ \lambda_2\ \ldots\ \lambda_L]$. L constraints to be satisfied are imposed in the cost function $J(\mathbf{x})$ through the Lagrange multipliers.

By taking gradient with respect to $\mathbf{x}$ and $\boldsymbol{\lambda}$ and equating them to zero, we get the following equations:

$$\nabla_{\mathbf{x}} J(\mathbf{x}) = 2\mathbf{Rx} + \mathbf{A^T}\boldsymbol{\lambda} = 0 \tag{18}$$

$$\nabla_{\boldsymbol{\lambda}} J(\mathbf{x}) = \mathbf{Ax - b} = 0 \tag{19}$$

Solving $\mathbf{x}$ from (18), we get $\mathbf{x} = \frac{1}{2}\mathbf{R^{-1}A^T}\boldsymbol{\lambda}$. Replacing $\mathbf{x}$ into (19), we get a solution for $\boldsymbol{\lambda}$, $\boldsymbol{\lambda} = 2(\mathbf{AR^{-1}A^T})^{-1}\mathbf{b}$. By substituting $\boldsymbol{\lambda}$ back into the first equation, we get the solution as follows:

$$\mathbf{x}_{\text{opt}} = \mathbf{R^{-1}A^T}(\mathbf{AR^{-1}A^T})^{-1}\mathbf{b} \tag{20}$$

A.2 Matlab Code of the Illustrations Given in ML Estimation Section

```
omega=0.2;
alpha=0.8;
N=40;
n=0:N-1;
SNR_dB=16;

SNR=10^(SNR_dB/10);
signalvar=alpha^2*1/2;   %average signal power
noisevar= signalvar/SNR;

x=alpha*cos(2*pi*omega*n)+sqrt(noisevar)*randn(1,N);  %observations

%construct the grid
[ag,og]=meshgrid(0:1/128:1-1/128,0:1/256:1/2-1/256);

%sketch the cost function for ML search
dum=zeros(size(ag));
for thisn=n,
    dum=dum+(x(thisn+1)-ag.*cos(2*pi*og*(thisn))).^2;
end;
cost1=exp(-1/2/noisevar*(dum));  %unnecessary constants are not included
mesh(ag,og,cost1);
xlabel('\alpha','fontsize',10);
ylabel('\omega','fontsize',10);
title(['SNR= ' num2str(SNR_dB) ' dB; N=' num2str(N)])
view(-22,46);
```

References

[1] H. L. V. Trees, *Detection, Estimation and Modulation Theory, part 1.* John Wiley - Sons, 1971.

[2] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory* . Prentice Hall, 1993.

[3] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications and Control.* Prentice Hall, 1995.