Chapter 1
Stochastic processes

Chapter 2
Statistical Estimation Theory

Chapter 3
Linear Filtering

Contents

3.1 Introduction

A common challenge in estimation involves determining the coefficients of a linear filter with M coefficients based solely on the observation of its inputs and outputs. This task, along with related ones, falls under the generic term 'linear filtering.' In this chapter, we will demonstrate how the techniques described in Chapter 2 can be applied to design ML, MAP, MAD and MMSE estimators for the coefficients of the aforementioned filter. This will also include the estimation of future filter outputs, provided that the corresponding inputs are known.

3.2 The filtering problem

Assume that a finite impulse response filter (FIR) $s[n]$, with $s[n] = 0$, for n other than $0, 1, \ldots, M-1$ is used to filter a signal[1] $u[n]$. The result is then added to a certain Gaussian noise $\varepsilon[n]$, which is an IID zero-mean stochastic process with variance σ_ε^2 and independent of $s[n]$, giving rise to an observation $x[n]$ (see Figure 3.1). That is, the corresponding entries are

$$
\begin{aligned}
x[n] &= u[n] * s[n] + \varepsilon[n] \\
&= u[n]s[0] + u[n-1]s[1] + \ldots + u[n-M+1]s[M-1] + \varepsilon[n].
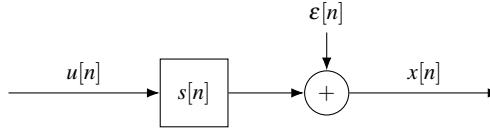\end{aligned}
\tag{3.1}
$$

Fig. 3.1 The signal model used in this chapter.

For simplicity, we will assume that the input signal starts at $n = 0$, that is, $u[n] = 0$ for $n < 0$. Also, we will assume that it is known; thus, a satistical model for $u[n]$ is unnecessary for our analysis).

The filtering problem consists in estimating the filter $s[n]$ from the input signal and a finite set of N samples from the output, $x[0], \ldots, x[N-1]$. Also, we will tackle the output prediction problem: predicting unobserved values of the output for a given input signal $u[n]$.

Our goal is to solve the filtering problem using the tools from estimation theory developed in Chapter 2. To do so, we will first represent the target variables (the filter coefficients) and the observations (the output signal) in vector form, and the relation between them using a vector equation. Thus, we will join the nonzero coefficients in an M-dimensional vector

---

[1] Although the notation $u[n]$ is frequently used to refer to the step function, in this chapter it is used to denote an arbitrary input signal.

$$\mathbf{s} = \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[M-1] \end{bmatrix}_{M \times 1}. \tag{3.2}$$

Also, we will represent any M-length window of consecutive input values in vector form as

$$\mathbf{u}[n] = \begin{bmatrix} u[n] \\ u[n-1] \\ \vdots \\ u[n-M+1] \end{bmatrix}_{M \times 1}, \tag{3.3}$$

we can write

$$x[n] = \mathbf{u}[n]^\top \mathbf{s} + \varepsilon[n]. \tag{3.4}$$

Note that, for any n, $\mathbf{u}[n]$ contains only the input values that are relevant to compute $x[n]$.

Taking $n = 0, 1, \ldots, N-1$ in Eq. (3.4), we get a system of N linear equations relating the filter coefficients and the observations. We can join all of them into a single matrix equation by defining the observation vector

$$\mathbf{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}_{N \times 1}, \tag{3.5}$$

the input matrix

$$\begin{aligned} \mathbf{U} &= [\mathbf{u}[0] \ \ \mathbf{u}[1] \ \ \ldots \ \ \mathbf{u}[M-1] \ \ \ldots \ \ \mathbf{u}[N-1]] \\ &= \begin{bmatrix} u[0] & u[1] & \ldots & u[M-1] & \ldots & u[N-1] \\ 0 & u[0] & \ldots & u[M-2] & \ldots & u[N-2] \\ \vdots & \vdots & \ddots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & u[0] & \ldots & u[N-M] \end{bmatrix}_{M \times N}, \end{aligned} \tag{3.6}$$

and the noise vector

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon[0] \\ \varepsilon[1] \\ \vdots \\ \varepsilon[N-1] \end{bmatrix}_{N \times 1}, \tag{3.7}$$

Using (3.5), (3.6) (3.7), we can write the signal model (3.4) as

$$\boxed{\mathbf{x} = \mathbf{U}^\top \mathbf{s} + \boldsymbol{\varepsilon}} \tag{3.8}$$

The filtering problem reduces to the problem of estimating $\mathbf{s}$ given $\mathbf{x}$ knowing (3.8) and the noise statistics.

3.3 ML solution

Eq. (3.4) shows that, given $\mathbf{s}$, $x[n]$ is Gaussian with mean

$$\mathbb{E}\{x[n] \mid \mathbf{s}\} = \mathbf{u}[n]^\top \mathbf{s} + \mathbb{E}\{\varepsilon[n]\} = \mathbf{u}[n]^\top \mathbf{s} \tag{3.9}$$

and variance

$$\mathbb{E}\{(x[n] - \mathbf{u}[n]^\top \mathbf{s})^2 \mid \mathbf{s}\} = \sigma_\varepsilon^2 \tag{3.10}$$

that is, the likelihood function is

$$p(x[n] \mid \mathbf{s}) = \mathcal{N}(x[n] \mid \mathbf{u}[n]^\top \mathbf{s}, \sigma_\varepsilon^2), \tag{3.11}$$

where the notation $\mathcal{N}(x \mid \mu, v)$ is used to refer to the *normal* (Gaussian) pdf of a random variable with mean μ and variance v, evaluated at x.

Given $\mathbf{s}$, $x[n]$ depends on $\varepsilon[n]$ only, which is an IID process. Thus, all samples from $x[n]$ are independent given $\mathbf{s}$, that is,

$$p_{\mathbf{X}|\mathbf{S}}(\mathbf{x} \mid \mathbf{s}) = \prod_{n=0}^{N-1} \mathcal{N}(x[n] \mid \mathbf{u}[n]^\top \mathbf{s}, \sigma_\varepsilon^2) = \mathcal{N}(\mathbf{x} \mid \mathbf{U}^\top \mathbf{s}, \sigma_\varepsilon^2 \mathbf{I}). \tag{3.12}$$

The value of $\mathbf{s}$ that maximizes $p(\mathbf{x}|\mathbf{s})$ is

$$
\begin{aligned}
\hat{\mathbf{s}}_{\text{ML}} &= \underset{\mathbf{s}}{\arg\max}\, p_{\mathbf{X}|\mathbf{S}}(\mathbf{x} \mid \mathbf{s}) = \underset{\mathbf{s}}{\arg\max}\, \log p_{\mathbf{X}|\mathbf{S}}(\mathbf{x} \mid \mathbf{s}) \\
&= \underset{\mathbf{s}}{\arg\min}\, \frac{1}{2}(\mathbf{x} - \mathbf{U}^\top \mathbf{s})^\top (\sigma_\varepsilon^2 \mathbf{I})^{-1}(\mathbf{x} - \mathbf{U}^\top \mathbf{s}) + \frac{1}{2}\log|\sigma_\varepsilon^2 \mathbf{I}| + \frac{N}{2}\log(2\pi) \\
&= \underset{\mathbf{s}}{\arg\min}\, ||\mathbf{x} - \mathbf{U}^\top \mathbf{s}||^2
\end{aligned}
$$

$$\tag{3.13}$$

This minimum can be easily obtained by taking the gradient with respect to $\mathbf{s}$, equalizing to zero and clearing, leading to

$$\hat{\mathbf{s}}_{\text{ML}} = (\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{U}\mathbf{x}. \tag{3.14}$$

3.4 Bayesian Solution

To obtain a Bayesian estimator of $\mathbf{s}$ it is necessary to know its prior probability distribution, $p(\mathbf{s})$. A common choice for this distribution is

$$p_{\mathbf{S}}(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\mathbf{0}, \sigma_s^2 \mathbf{I}), \tag{3.15}$$

since it accommodates any set of real coefficients and assumes they have a zero mean and a dispersion determined by σ_s^2. It is also possible to set $\sigma_s^2 \to \infty$ to approach a uniform distribution. In any case, the use of this prior distribution enables the analytic derivation of the posterior distribution.

Given the likelihood, $p(\mathbf{x}|\mathbf{s})$, and the prior distribution $p(\mathbf{s})$, the posterior distribution $p(\mathbf{s}|\mathbf{x})$ can be obtained. While it is feasible to directly apply Bayes' theorem and simplify the expression as much as possible, this process can be quite tedious. Instead, we will achieve the result in two steps.

First we will determine the joint pdf of $\mathbf{s}$ and $\mathbf{x}$. A simple way to do this is to observe that

$$\begin{bmatrix} \mathbf{s} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{U}^\top \end{bmatrix} \mathbf{s} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\varepsilon} \end{bmatrix} \tag{3.16}$$

that is, vector $[\mathbf{s}^\top \ \mathbf{x}^\top]^\top$ is a linear combination of two Gaussian random vectors and, thus, is jointly Gaussian:

$$p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{s} \\ \mathbf{x} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m_S} \\ \mathbf{m_X} \end{bmatrix}, \begin{bmatrix} \mathbf{V_S} & \mathbf{V_{SX}} \\ \mathbf{V_{SX}^\top} & \mathbf{V_X} \end{bmatrix} \right) \tag{3.17}$$

where

$$\mathbf{m_S} = \mathbb{E}\{\mathbf{s}\} = \mathbf{0} \tag{3.18}$$

$$\mathbf{m_X} = \mathbb{E}\{\mathbf{x}\} = \mathbf{U}^\top \mathbb{E}\{\mathbf{s}\} + \mathbb{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0} \tag{3.19}$$

$$\mathbf{V_S} = \mathrm{Var}\{\mathbf{s}\} = \sigma_s^2 \mathbf{I} \tag{3.20}$$

$$\mathbf{V_{SX}} = \mathbb{E}\{\mathbf{s}\mathbf{x}^\top\} = \mathbb{E}\{\mathbf{s}\mathbf{s}^\top\}\mathbf{U} + \mathbb{E}\{\mathbf{s}\boldsymbol{\varepsilon}^\top\} = \sigma_s^2 \mathbf{U} \tag{3.21}$$

$$\mathbf{V_X} = \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\} = \mathbf{U}^\top \mathbb{E}\{\mathbf{s}\mathbf{s}^\top\}\mathbf{U} + \mathbb{E}\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\} = \sigma_s^2 \mathbf{U}^\top \mathbf{U} + \sigma_\varepsilon^2 \mathbf{I} \tag{3.22}$$

Therefore,

$$p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{s} \\ \mathbf{x} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_s^2 \mathbf{I} & \sigma_s^2 \mathbf{U} \\ \sigma_s^2 \mathbf{U}^\top & \sigma_s^2 \mathbf{U}^\top \mathbf{U} + \sigma_\varepsilon^2 \mathbf{I} \end{bmatrix} \right) \tag{3.23}$$

From previous chapter, we know that if $p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x})$ is Gaussian, the conditional distribution $p_{\mathbf{S}|\mathbf{X}}(\mathbf{s}\,|\,\mathbf{x})$ is also Gaussian, with mean and covariance

$$\mathbf{m}_{\mathbf{S}|\mathbf{X}} = \mathbf{m}_{\mathbf{S}} + \mathbf{V}_{\mathbf{SX}} \mathbf{V}_{\mathbf{X}}^{-1} (\mathbf{x} - \mathbf{m}_{\mathbf{X}})$$

$$= \mathbf{U} \left(\mathbf{U}^{\top} \mathbf{U} + \frac{\sigma_{\varepsilon}^2}{\sigma_s^2} \mathbf{I} \right)^{-1} \mathbf{x} \tag{3.24}$$

$$\mathbf{V}_{\mathbf{S}|\mathbf{X}} = \mathbf{V}_{\mathbf{S}} - \mathbf{V}_{\mathbf{SX}} \mathbf{V}_{\mathbf{X}}^{-1} \mathbf{V}_{\mathbf{SX}}^{\top}$$

$$= \sigma_s^2 \mathbf{I} - \sigma_s^2 \mathbf{U} \left(\mathbf{U}^{\top} \mathbf{U} + \frac{\sigma_{\varepsilon}^2}{\sigma_s^2} \mathbf{I} \right)^{-1} \mathbf{U}^{\top} \tag{3.25}$$

The above expressions involve the computation of the inverse of an $N \times N$ matrix, which may be not feasible for large signal records (the computational cost is $\mathscr{O}(N^3)$). However, using the *matrix inversion lemma* (see the Appendix in Sec. 3.7), we can obtain the following alternative expressions:

$$\mathbf{m}_{\mathbf{S}|\mathbf{X}} = \mathbf{PUx} \tag{3.26}$$

$$\mathbf{V}_{\mathbf{S}|\mathbf{X}} = \sigma_{\varepsilon}^2 \mathbf{P} \tag{3.27}$$

where

$$\mathbf{P} = (\mathbf{U}\mathbf{U}^{\top} + \frac{\sigma_{\varepsilon}^2}{\sigma_s^2} \mathbf{I})^{-1}. \tag{3.28}$$

Eq. (3.28) involves the inversion of a matrix $M \times M$, which is usually much smaller than $N \times N$.

Using these expressions, the MMSE, MAP and MAD estimates of $\mathbf{s}$ are:

$$\hat{\mathbf{s}}_{\mathrm{MSE}} = \hat{\mathbf{s}}_{\mathrm{MAP}} = \hat{\mathbf{s}}_{\mathrm{MAD}} = \mathbf{PUx} \tag{3.29}$$

Also, note that taking $\sigma_{\varepsilon}^2 \to 0$ (negligible noise) and/or $\sigma_s^2 \to \infty$ (which can be interpreted as assuming an infinitely wide uniform prior) these Bayesian solutions become equivalent to the ML in (3.14).

3.4.1 Probabilistic prediction of the filter output

Once we have resolved several estimators of filter $\mathbf{s}$, we now begin to consider the problem of predicting a new output $x[k]$ at some time $k > N$. Continuing with the Bayesian perspective, we will obtain the posterior pdf of the target variable, $x[k]$, in light of the outputs already observed, $\mathbf{x}$. That is, we aim to calculate $p(x[k] \mid \mathbf{x})$.

First, it should be noted that $\mathbf{x}, x[k]$ and $\mathbf{s}$ are jointly Gaussian. This follows from Eq. (3.23), which can be extended to any arbitrary number of outputs, including $x[k]$. This

necessarily implies that $\mathbf{x}$ and $x[k]$ are jointly Gaussian (when marginalizing $\mathbf{s}$) and finally that $p(x[k]|\mathbf{x})$ must be Gaussian. Given that

$$x[k] = \mathbf{u}[k]^\top \mathbf{s} + \varepsilon[k] \tag{3.30}$$

is a linear transformation of $\mathbf{s}$ with independent white noise, we can easily compute the posterior mean and variance, respectively, as follows:

$$\mathbb{E}\{x[k] \mid \mathbf{x}\} = \mathbf{u}[k]^\top \mathbb{E}\{\mathbf{s} \mid \mathbf{x}\} + \mathbb{E}\{\varepsilon[k] \mid \mathbf{x}\} = \mathbf{u}[k]^\top \hat{\mathbf{s}}_{\text{MSE}} \tag{3.31}$$

$$\begin{aligned}
\text{Var}\{x[k] \mid \mathbf{x}\} &= \mathbb{E}\left\{ \left(\mathbf{u}[k]^\top (\mathbf{s} - \hat{\mathbf{s}}_{\text{MSE}}) + \varepsilon[k] \right)^2 \mid \mathbf{x} \right\} \\
&= \mathbf{u}[k]^\top \mathbb{E}\{(\mathbf{s} - \hat{\mathbf{s}}_{\text{MSE}})(\mathbf{s} - \hat{\mathbf{s}}_{\text{MSE}})^\top \mid \mathbf{x}\}\mathbf{u}[k] + \mathbb{E}\{\varepsilon[k]^2 \mid \mathbf{x}\} \\
&= \mathbf{u}[k]^\top \mathbf{V}_{\mathbf{S}|\mathbf{X}}\mathbf{u}[k] + \sigma_\varepsilon^2 \\
&= \sigma_\varepsilon^2 \mathbf{u}[k]^\top \mathbf{P}\mathbf{u}[k] + \sigma_\varepsilon^2 \tag{3.32}
\end{aligned}$$

Therefore, the Bayesian prediction of the output is succinctly given by

$$\hat{x}_{\text{MSE}}[k] = \hat{x}_{\text{MAP}}[k] = \hat{x}_{\text{MAD}}[k] = \mathbf{u}[k]^\top \hat{\mathbf{s}}_{\text{MSE}}. \tag{3.33}$$

Eq. (3.33) shows that the optimal prediction of the output at any time $k > N$ can be computed by passing the input signal $u[n]$ though the filter with the coefficients in the Bayesian estimation of $\mathbf{s}$ (see Fig. 3.2).
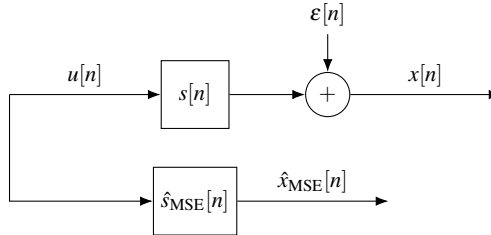

Fig. 3.2 The best prediction of future values of the output (knowing the input) can be computed by passing the input signal through a filter with the coefficients of the Bayesian estimation of $\mathbf{s}$.

3.5 Online calculus

It is possible to obtain the above solutions online, that is, as new input-output pairs are obtained. While complete calculations could be repeated each time a new sample arrives, there are often more efficient ways to do this.

Note that estimating $\mathbf{s}$ using Eqs. (3.14) or (3.29) requires inverting an $M \times M$ matrix. This has a cost $\mathcal{O}(M^3)$, that is, if we double the size of the filter, M we multiply its computational cost by eight. Suppose now that you want to estimate $\mathbf{s}$ as new input-output pairs

are received, that is, we are given first $\{u[0], x[0]\}$, then $\{u[1], x[1]\}$ and so on. In this case, we could reuse the results of the previous estimate to calculate the new updated estimate of $\mathbf{s}$, thus reducing the cost $\mathcal{O}(M^3)$ that would have a *naive* method that simply recalculates everything again every time a sample arrives.

3.5.1 Bayesian solution

$\hat{\mathbf{s}}_{\mathrm{MSE}}$ can be obtained exactly as more samples are available (i.e. as N increases) without redoing all calculations, by reusing the previous solution. To do this, it is defined

$$\mathbf{P}_N = (\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I})^{-1}, \tag{3.34}$$

$$\mathbf{r}_N = \mathbf{U}\mathbf{x} \tag{3.35}$$

and the following recursive calculation is used (the first equation corresponds to the direct application of the matrix inversion lemma to the $\mathbf{P}$ update):

$$\mathbf{P}_{N+1} = \mathbf{P}_N - \frac{\mathbf{P}_N \mathbf{u}[N+1]\mathbf{u}[N+1]^\top \mathbf{P}_N}{1 + \mathbf{u}[N+1]^\top \mathbf{P}_N \mathbf{u}[N+1]}$$

$$\mathbf{r}_{N+1} = \mathbf{r}_N + \mathbf{u}[N+1]x[N+1]$$

$$\mathbf{s}_{N+1} = \mathbf{P}_{N+1}\mathbf{r}_{N+1},$$

which only has a cost $\mathcal{O}(M^2)$ per step (as opposed to applying the complete original equation at each step, which would cost $\mathcal{O}(M^3)$). This algorithm is called *recursive least squares* (RLS).

3.5.2 ML solution

An online approximation to $\hat{\mathbf{s}}_{\mathrm{ML}}$ with computational cost $\mathcal{O}(M)$ can be obtained just by noting that

$$\hat{\mathbf{s}}_{\mathrm{ML}} = \underset{\mathbf{s}}{\operatorname{argmax}}\, p(\mathbf{x}|\mathbf{s}) = \underset{\mathbf{s}}{\operatorname{argmin}}\, ||\mathbf{x} - \mathbf{U}^\top \mathbf{s}||^2 \tag{3.36}$$

and then use stochastic gradient to minimize $||\mathbf{x} - \mathbf{U}^\top \mathbf{s}||^2$.

Notice that

$$||\mathbf{x} - \mathbf{U}^\top \mathbf{s}||^2 = \sum_{n=0}^{N-1} (x[n] - \mathbf{u}[n]^\top \mathbf{s})^2, \tag{3.37}$$

so a gradient descent method would calculate the gradient of that expression and iteratively shift the estimate of the minimum in the opposite direction of the gradient in each step. A descent by stochastic gradient performs the same operation, but considering only one of the additions of the mentioned sum in each step. So, the updating of coefficients that must be iterated to perform the minimization is in this case

$$\hat{\mathbf{s}}_{n+1} = \hat{\mathbf{s}}_n + \mu \left(x[n] - \mathbf{u}[n]^\top \hat{\mathbf{s}}_n \right) \mathbf{u}[n], \tag{3.38}$$

where μ is an adaptation step that should be "small enough". This algorithm is called *least mean squares* (LMS).

3.6 Problems

3.1 Consider the sequence

$$u[1]\ldots u[7] \equiv 0.7,\ -0.1,\ 0.7,\ -0.2,\ -0.1,\ 1.5,\ -1.1$$

which is fed as input to a linear filter of three coefficients, $\mathbf{s} = [s_1, s_2, s_3]^\top$. The following elements of the output sequence are known, (corrupted with Gaussian noise of variance 0.25):

$$x[1]\ldots x[6] \equiv -0.60,\ 1.13,\ 0.57,\ 0.42,\ 1.25,\ -2.58$$

a) What is the ML estimate of $\mathbf{s}$ given the data?
b) Use the obtained filter to predict $x[7]$, $\hat{x}_{\text{ML}}$.
c) Calculate the MSE, MAP and MAD estimates of $\mathbf{s}$ assuming that the prior pdf of its components is $s_i \sim \mathcal{N}(0,1)$ and that they are independent.
d) Get the MSE estimate of $x[7]$, $\hat{x}_{\text{MSE}}$.
e) Calculate the MSE in prediction b). (That is, the mean of $(\hat{x}_{\text{ML}} - x[7])^2$ given the data).
f) Calculate the MSE in prediction d). (That is, the mean of $(\hat{x}_{\text{MMSE}} - x[6])^2$ given the data)

3.7 Appendix: the matrix inversion lemma

In this section we apply the matrix inversion lemma (also known as the Woodbury matrix identity) to obtain expression. This lemma states that, for any matrices $\mathbf{U}$, $\mathbf{C}$, bfV and any non-singular matrix $\mathbf{A}$,

$$(\mathbf{A} + \mathbf{VCU})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{V}\left(\mathbf{C}^{-1} + \mathbf{UA}^{-1}\mathbf{V}\right)^{-1}\mathbf{UA}^{-1} \tag{3.39}$$

Proving this equality is not difficult: multiplying the right-hand side of the equality by $\mathbf{A} + \mathbf{UCV}$, it is easy to see that the result is the identity matrix.

Taking

$$\mathbf{A} = \frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I} \tag{3.40}$$

$$\mathbf{C} = \mathbf{I} \tag{3.41}$$

$$\mathbf{V} = \mathbf{U}^\top \tag{3.42}$$

we get

$$\left(\mathbf{U}^\top \mathbf{U} + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1} = \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{I} - \left(\tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\right)^2 \mathbf{U}^\top \left(\mathbf{I} + \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top\right)^{-1}\mathbf{U}$$

$$= \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{I} - \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}^\top \left(\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1}\mathbf{U}$$

$$= \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{I} - \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}^\top \mathbf{P}\mathbf{U} \tag{3.43}$$

Where $\mathbf{P}$ is given by (3.28). Note that, by definition,

$$\left(\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)\mathbf{P} = \mathbf{P}\left(\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right) = \mathbf{I} \tag{3.44}$$

so that

$$\mathbf{U}\mathbf{U}^\top\mathbf{P} = \mathbf{P}\mathbf{U}\mathbf{U}^\top = \mathbf{I} - \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{P} \tag{3.45}$$

We will use these equalities below. Using (3.43) into (3.24), we get

$$\mathbf{m}_{\mathbf{S}|\mathbf{X}} = \mathbf{U}\left(\mathbf{U}^\top \mathbf{U} + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1}\mathbf{x}$$

$$= \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{x} - \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top\mathbf{P}\mathbf{U}\mathbf{x}$$

$$= \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{x} - \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\left(\mathbf{I} - \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{P}\right)\mathbf{U}\mathbf{x}$$

$$= \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{x} - \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{x} + \mathbf{P}\mathbf{U}\mathbf{x}$$

$$= \mathbf{P}\mathbf{U}\mathbf{x} \tag{3.46}$$

$$\mathbf{V}_{\mathbf{S}|\mathbf{X}} = \sigma_s^2\mathbf{I} - \sigma_s^2\mathbf{U}\left(\mathbf{U}^\top \mathbf{U} + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1}\mathbf{U}^\top$$

$$= \sigma_s^2\left[\mathbf{I} - \mathbf{U}\left(\tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{I} - \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}^\top\mathbf{P}\mathbf{U}\right)\mathbf{U}^\top\right]$$

$$= \sigma_s^2\left[\mathbf{I} - \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top\mathbf{P}\mathbf{U}\mathbf{U}^\top\right]$$

$$= \sigma_s^2\left[\mathbf{I} - \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_s^2}{\sigma_\varepsilon^2}\left(\mathbf{I} - \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{P}\right)\mathbf{U}\mathbf{U}^\top\right]$$

$$= \sigma_\varepsilon^2\mathbf{P} \tag{3.47}$$