Consider the Bayesian statistical model

\[ X = H \cdot \theta + W \]

where

- \( \theta \) is unknown, \( p \times 1 \)
- \( H \) is known, \( N \times p \)
- \( \theta \sim N(\mu_0, R_0) \)
- \( W \sim N(0, R_w) \)
- \( \theta \) and \( W \) are independent
- \( R_0, R_w, \mu_0 \) are known.
This model amounts to a signal subspace with a Gaussian prior on $\Theta$ and a Gaussian conditional distribution of $X$ given $\Theta$.

This formulation is quite general and encompasses many interesting and important examples.

**Example** Suppose $X = S + W$ where

$$s(n) = \cos(2\pi fn + \phi), \quad n = 0, 1, \ldots, N-1$$

and $-\frac{1}{N} \leq \phi \leq \frac{1}{N}$. On the homework we have seen that it is possible to approximate $S = H\Theta$ where the dimension of $\Theta$ is $p = 2L+1$, and $\Theta$ follows a Gaussian distribution.
Result. The posterior distribution of $\theta|\mathbf{x}$ is

$$\theta|\mathbf{x} \sim N \left( \mu_{\theta|\mathbf{x}}, \sigma_{\theta|\mathbf{x}} \right)$$

where

$$\mu_{\theta|\mathbf{x}} = \mu_{\theta} + \Phi_{\theta} H^T (H R_{\theta} H^T + R_w)^{-1} (x - H \mu_{\theta})$$

$$\sigma_{\theta|\mathbf{x}}^2 = R_{\theta} - \Phi_{\theta} H^T (H R_{\theta} H^T + R_w)^{-1} H \Phi_{\theta}$$

Proof. $x$ and $\theta$ are jointly Gaussian:

$$\begin{bmatrix} x \\ \theta \end{bmatrix} \sim N \left( \begin{bmatrix} \mu_x \\ \mu_{\theta} \end{bmatrix}, \begin{bmatrix} R_x & 0 \\ 0 & R_{\theta} \end{bmatrix} \right)$$

where

$$\begin{bmatrix} \theta \\ w \end{bmatrix} \sim N \left( \begin{bmatrix} \mu_{\theta} \\ 0 \end{bmatrix}, \begin{bmatrix} R_{\theta} & 0 \\ 0 & R_w \end{bmatrix} \right)$$

$$\implies \begin{bmatrix} x \\ \theta \end{bmatrix} \sim N \left( \begin{bmatrix} H \mu_{\theta} \\ \mu_{\theta} \end{bmatrix}, \begin{bmatrix} H R_{\theta} H^T + R_w & H R_{\theta} \\ H R_{\theta}^T & R_{\theta} \end{bmatrix} \right)$$

Now apply the Gaussian conditioning principle.
It can be shown using the matrix inversion lemma that

\[ M_{	heta \mid x} = M_{\theta} + R_{\theta} H^T (H R_{\theta} H^T + R_w)^{-1} (x - H M_{\theta}) \]

\[ = M_{\theta} + (H^T R_w^{-1} H + R_{\theta}^{-1}) H^T R_w^{-1} (x - H M_{\theta}) \]

and

\[ R_{\theta \mid x} = R_{\theta} - R_{\theta} H^T (H R_{\theta} H^T + R_w)^{-1} H R_{\theta} \]

\[ = (H^T R_w^{-1} H + R_{\theta}^{-1})^{-1} \]

These alternative formulas are sometimes more convenient to work with.

To verify these formulas is a tedious but manageable exercise.
The posterior distribution is Gaussian, which is symmetric and unimodal. Therefore, the optimal estimator (minimizing the Bayes risk) is

\[
\hat{\theta}(x) = \mu_0 + R_0 H^T (H R_0 H^T + R_w)^{-1} (x - H \mu_0)
\]

\[
= \mu_0 + (H^T R_w^{-1} H + R_\theta^{-1})^{-1} H^T R_w^{-1} (x - H \mu_0)
\]

regardless of the loss function.

Observations:

1. \(\hat{\theta}(x)\) is an affine function of \(x\).
2. \(\hat{\theta}(x)\) is again multivariate Gaussian.
3. Consider the case where \(R_\theta = \sigma^2 I_p\) and \(\sigma^2 \to \infty\). This can be thought of as a "non-committal" prior. Then \(R_\theta^{-1} \to 0_p\) and

\[
\hat{\theta}(x) = \mu_0 + (H^T R_w^{-1} H)^{-1} H^T R_w^{-1} (x - H \mu_0)
\]

\[
= (H^T R_w^{-1} H)^{-1} H^T R_w^{-1} x
\]

\[
= \text{MLE/MVUE}
\]
Exercise

Suppose we observe

\[ X_i = A + W_i, \quad i = 1, \ldots, N \]

where \( A \) is an unknown scalar and

\[ A \sim N(\mu_A, \sigma_A^2) \]

\[ W_i \text{ iid } \sim N(0, \sigma_w^2) \]

with \( \mu_A, \sigma_A^2, \sigma_w^2 \) known. Find the Bayesian estimate \( \hat{A} \).

Interpret your result. Analyze limiting cases.
Solution: The problem falls within the linear model with

\[ H = \frac{1}{N} \quad (N \times 1) \]

\[ \theta = A \quad (1 \times 1) \]

\[ M_\theta = M_A \quad (1 \times 1) \]

\[ R_\theta = \sigma_A^2 \quad (1 \times 1) \]

\[ R_w = \sigma^2 I_N \quad (N \times N) \]

Using the second formula for \( M_A \mid x \) (the one that comes from the matrix inversion lemma) we obtain

\[
\hat{A}(x) = M_A \mid x = M_A + \left( \frac{1}{\sigma^2} + \frac{1}{\sigma_w^2} \right)^{-1} \cdot \frac{1}{\sigma_w^2} (x - \frac{1}{N} M_A)
\]

\[
= M_A + \left( \frac{N}{\sigma_w^2} + \frac{1}{\sigma_A^2} \right)^{-1} \cdot \frac{1}{\sigma_A^2} (\sum x_i - N M_A)
\]

\[
= M_A + \frac{1}{\frac{N}{\sigma_w^2} + \frac{1}{\sigma_A^2}} \cdot \frac{N}{\sigma_A^2} (\bar{x} - M_A)
\]

\[
= M_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_w^2}{N}} (\bar{x} - M_A)
\]
Thus

\[ \hat{A}(\alpha) = (1 - \alpha) \mu_k + \alpha \cdot \bar{x} \]

where

\[ \alpha = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_w^2} \]

controls the tradeoff between prior knowledge and data.

**Limiting cases:**

(a) \( N \to \infty \Rightarrow \alpha \to 0 \Rightarrow \hat{A} \to \)

\( N = 0 \Rightarrow \alpha = 0 \Rightarrow \hat{A} = \)

\( \sigma_A^2 \to \infty \Rightarrow \alpha \to 1 \Rightarrow \hat{A} \to \)

\( \sigma_A^2 \to 0 \Rightarrow \alpha \to 0 \Rightarrow \hat{A} \to \)
It suffices to focus on the case $\mu_0 = 0$. Then the Bayesian estimator is

\[ \hat{\mu}(x) = \mu_0|x = R_0 H^T (H R_0 H^T + R_w)^{-1} x \]

\[ = (H^T R_w^{-1} H + R_0^{-1})^{-1} H^T R_0^{-1} x \]

If ever $\mu_0 \neq 0$, we may apply the above estimator to $x = H \mu_0$ and add $\mu_0$ to the result.

Simultaneously Diagonalizable Covariance Matrices

Consider the problem of estimating a signal in additive Gaussian noise

\[ x = s + w \]

where

\[ x = \text{observed noisy signal} \]
\[ s = \text{clean signal} \]
\[ w = \text{noise} \]
This can be modeled using the general linear model with

$$\theta = \xi$$

$$H = I_n$$

and adopting a Gaussian prior for $\xi$:

$$\xi \sim N(0, R_{ss}).$$

The Bayesian estimate for $\xi$ is

$$\hat{\xi} = \text{...}$$
Application: Bandpass Filtering

Suppose we observe

\[ x = s + w \]

and we know a priori that the signal of interest occupies a certain passband. In other words, \( |y_k^H x| \) is large on average for certain DFT basis vectors \( y_k \), and small for others.

How can we incorporate this prior knowledge into the prior for \( s \)? In other words, what should we take for \( R_{ss} \)?

Let us assume we can specify

\[ \sigma_k^2 = \mathbb{E} \left\{ |y_k^H S|^2 \right\} \]

the average signal energy at frequency \( k/N \).

Let's also assume that signal content at different frequencies are independent.
In essence we are taking

\[ U^H \Sigma = N(0, \Sigma) \]

as a prior, where

\[ \Sigma = \begin{pmatrix}
\sigma_1^2 \\
\sigma_2^2 \\
\vdots \\
\sigma_N^2
\end{pmatrix} \]

Equivalently, the prior on \( \Sigma \) is

\[ \Sigma \sim N(0, \sqrt{R_{SS}} U \Sigma U^H) \]

For example

\[ \sigma_k^2 = \frac{\sigma_k^2}{\sigma_{N-k}} \text{ by conjugate symmetry} \]

is a lowpass model.
Notice that the energy of $s$ is

$$E\{s^{T}s\} = E\{ (u^H s)(u^H s) \}$$

$$= \sum_{k=0}^{N-1} \sigma_k^2$$

So to specify the $\sigma_k^2$ it suffices to know the signal energy and the \underline{shape} of the frequency response.

Assume the noise is i.i.d:

$$R_{ww} = \sigma^2 I_N,$$

$\sigma^2$ known. Then the MMSE estimator is

$$\hat{s} = R_{ss} (R_{ss} + R_{ww})^{-1} \alpha$$

$$= u \Sigma u^H (u [\Sigma + \sigma^2 I] u^H)^{-1} \alpha$$

$$= u [\Sigma (\Sigma + \sigma^2 I)^{-1}] u^H \alpha$$
Note that
\[
\sum (\Sigma + \sigma^2 I)^{-1} = \begin{bmatrix}
\frac{\sigma_1^2}{\sigma_1^2 + \sigma_0^2} \\
\frac{\sigma_2^2}{\sigma_2^2 + \sigma_0^2} \\
\frac{\sigma_N^2}{\sigma_N^2 + \sigma^2}
\end{bmatrix}
\]

Therefore, the estimator is a bandpass filter.

**Interpretation:**
- \(\sigma_k^2 \gg \sigma^2\) \(\implies\) keep most of signal
- \(\sigma_k^2 \ll \sigma^2\) \(\implies\) kill most of signal
- \(\sigma_k^2 \approx \sigma^2\) \(\implies\) keep some of signal
The preceding example generalizes easily to any situation where $R_{ss}$ and $R_{sw}$ are simultaneously diagonalizable.

If $R_{ss} = U \Lambda_s U^T$, $R_{sw} = U \Lambda_w U^T$
for a unitary matrix $U$, then

$$\hat{\Sigma} = U \left[ \Lambda_s (\Lambda_s + \Lambda_w)^{-1} \right] U^T \Lambda$$

where

$$\Lambda = \begin{bmatrix}
\frac{\lambda_s}{\lambda_s + 1/w} & 0 & \cdots & 0 \\
0 & \ddots & \ddots & 0 \\
0 & \cdots & \frac{\lambda_s}{\lambda_s + 1/w}
\end{bmatrix}$$

$\Rightarrow$ "transform domain shrinkage"
Summary

- Extension of signal subspace model to Bayesian setting.
- When subspace coefficients (prior) and observation noise (likelihood) are jointly Gaussian, posterior is also Gaussian (conjugate prior).
- Posterior mean (mode) is a linear/affine function.
- Classical estimators fall out in limiting cases.
- When $R_\theta$, $R_w$ are simultaneously diagonalizable
  $\Rightarrow$ transform domain "shrinkage"
  e.g., bandpass filtering.

Key

- $a. \quad 1, \bar{x}$
- $b. \quad R_{ss} (R_{ss} + R_{ww})^{-1} \bar{x}$
- $0, \mu_A$
- $1, \bar{x}$
- $0, \mu_A$