BAYESIAN ESTIMATION: THE GAUSSIAN LINEAR MODEL

Consider the Bayesian statistical model

$$\underline{X} = H \cdot \underline{\theta} + \underline{W}$$

where

$\underline{\theta}$ is unknown, $p \times 1$

H is known, $N \times p$

$\underline{\theta} \sim \mathcal{N}(\underline{\mu}_\theta, R_\theta)$

$\underline{W} \sim \mathcal{N}(\underline{0}, R_w)$

$\underline{\theta}$ and $\underline{W}$ are independent

$R_\theta, R_w, \underline{\mu}_\theta$ are known.

This model amounts to a signal subspace with a Gaussian prior on $\underline{\theta}$ and a Gaussian conditional distribution of $\underline{X}$ given $\underline{\theta}$.

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

Example Suppose $\underline{X} = \underline{S} + \underline{W}$ where

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

and $-\frac{L}{N} \leq f \leq \frac{L}{N}$. On the homework we have seen that it is possible to approximate $\underline{S} = H\underline{\theta}$ where the dimension of $\underline{\theta}$ is $p = 2L+1$, and $\underline{\theta}$ follows a Gaussian distribution.

Result The posterior distribution of $\underline{\theta}/\underline{x}$ is

$$\underline{\theta}/\underline{x} \sim \mathcal{N}(\underline{\mu}_{\theta/x}, R_{\theta/x})$$

where

$$\underline{\mu}_{\theta/x} = \underline{\mu}_\theta + R_\theta H^T (H R_\theta H^T + R_w)^{-1} (\underline{x} - H\underline{\mu}_\theta)$$

$$R_{\theta/x} = R_\theta - R_\theta H^T (H R_\theta H^T + R_w)^{-1} H R_\theta$$

Proof $\underline{X}$ and $\underline{\theta}$ are jointly Gaussian:

$$\begin{bmatrix} \underline{X} \\ \underline{\theta} \end{bmatrix} = \begin{bmatrix} H & I_N \\ I_p & 0 \end{bmatrix} \begin{bmatrix} \underline{\theta} \\ \underline{w} \end{bmatrix}$$

where

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

$$\Longrightarrow \begin{bmatrix} \underline{X} \\ \underline{\theta} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} H\underline{\mu}_\theta \\ \underline{\mu}_\theta \end{bmatrix}, \begin{bmatrix} H R_\theta H^T + R_w & H R_\theta \\ R_\theta H^T & R_\theta \end{bmatrix} \right)$$

Now apply the Gaussian conditioning principle.

It can be shown using the matrix inversion lemma that

$$\begin{aligned}\underline{\mu}_{\theta|x} &= \underline{\mu}_\theta + R_\theta H^T (H R_\theta H^T + R_\omega)^{-1} (\underline{x} - H\underline{\mu}_\theta) \\ &= \underline{\mu}_\theta + (H^T R_\omega^{-1} H + R_\theta^{-1})^{-1} H^T R_\omega^{-1} (\underline{x} - H\underline{\mu}_\theta)\end{aligned}$$

and

$$\begin{aligned}R_{\theta|x} &= R_\theta - R_\theta H^T (H R_\theta H^T + R_\omega)^{-1} H R_\theta \\ &= (H^T R_\omega^{-1} H + R_\theta^{-1})^{-1}\end{aligned}$$

These alternative formulas are sometimes more convenient to work with.

To verify these formulas is a tedious but manageable exercise

Estimation

The posterior distribution is Gaussian, which is symmetric and unimodal. Therefore, the optimal estimator (minimizing the Bayes risk) is

$$\hat{\underline{\theta}}(\underline{x}) = \underline{\mu}_{\theta|x} = \underline{\mu}_\theta + R_\theta H^T (H R_\theta H^T + R_w)^{-1} (\underline{x} - H\underline{\mu}_\theta)$$

$$= \underline{\mu}_\theta + (H^T R_w^{-1} H + R_\theta^{-1})^{-1} H^T R_w^{-1} (\underline{x} - H\underline{\mu}_\theta)$$

regardless of the loss function.

Observations

1. $\hat{\underline{\theta}}(\underline{x})$ is an affine function of $\underline{x}$.
2. $\hat{\underline{\theta}}(\underline{X})$ 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

$$\hat{\underline{\theta}}(\underline{x}) = \underline{\mu}_\theta + (H^T R_w^{-1} H)^{-1} H^T R_w^{-1} (\underline{x} - H\underline{\mu}_\theta)$$

$$= (H^T R_w^{-1} H)^{-1} H^T R_w^{-1} \underline{x}$$

$$= \text{MLE} / \text{MVUE}$$

Exercise Suppose we observe

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

where A is an unknown scalar and

$$\left.\begin{aligned} A &\sim N(\mu_A, \sigma_A^2) \\ W_i &\overset{iid}{\sim} N(0, \sigma_w^2) \end{aligned}\right\} \text{independent}$$

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

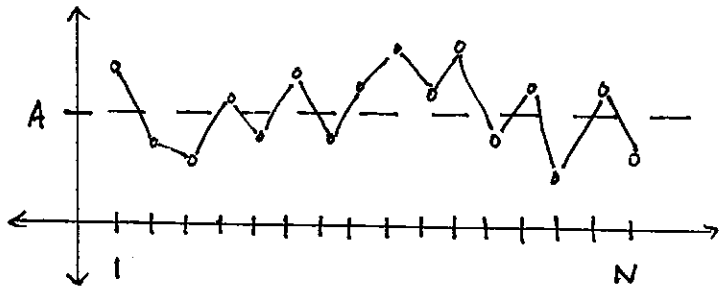
$$H = \underline{1} \quad (N \times 1)$$

$$\theta = A \quad (1 \times 1)$$

$$\mu_\theta = \mu_A \quad (1 \times 1)$$

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

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



Using the second formula for $\mu_{A/x}$ (the one that comes from the matrix inversion lemma) we obtain

$$\hat{A}(\underline{x}) = \mu_{A/x} = \mu_A + \left(\underline{1}^T \cdot \underline{1} \cdot \frac{1}{\sigma_w^2} + \frac{1}{\sigma_A^2}\right)^{-1} \underline{1}^T \cdot \frac{1}{\sigma_w^2} (\underline{x} - \underline{1}\mu_A)$$

$$= \mu_A + \left(\frac{N}{\sigma_w^2} + \frac{1}{\sigma_A^2}\right)^{-1} \frac{1}{\sigma_w^2} \left(\sum x_i - N\mu_A\right)$$

$$= \mu_A + \frac{1}{\frac{N}{\sigma_w^2} + \frac{1}{\sigma_A^2}} \cdot \frac{N}{\sigma_w^2} (\bar{x} - \mu_A)$$

$$= \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_w^2}{N}} (\bar{x} - \mu_A)$$

Thus

$$\hat{A}(\underline{x}) = (1-\alpha)\mu_A + \alpha \cdot \bar{x}$$

where

$$\alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_w^2}{N}}$$

controls the tradeoff between prior knowledge and data.

Limiting cases:

(a)

$N \to \infty \Rightarrow \alpha \to \Rightarrow \hat{A} \to$

$N = 0 \Rightarrow \alpha = \Rightarrow \hat{A} =$

$\sigma_A^2 \to \infty \Rightarrow \alpha \to \Rightarrow \hat{A} \to$

$\sigma_A^2 \to 0 \Rightarrow \alpha \to \Rightarrow \hat{A} \to$

It suffices to focus on the case $\underline{\mu}_\theta = \underline{0}$. Then the Bayesian estimator is

$$\hat{\underline{\theta}}(\underline{x}) = \underline{\mu}_{\theta|x} = R_\theta H^T \left(H R_\theta H^T + R_w\right)^{-1} \underline{x}$$

$$= \left(H^T R_w^{-1} H + R_\theta^{-1}\right)^{-1} H^T R_\theta^{-1} \underline{x}$$

If ever $\underline{\mu}_\theta \neq \underline{0}$, we may apply the above estimator to $\underline{x} - H\underline{\mu}_\theta$ and add $\underline{\mu}_\theta$ to the result.

Simultaneously Diagonalizable Covariance Matrices

Consider the problem of estimating a signal in additive Gaussian noise

$$\underline{x} = \underline{s} + \underline{w}$$

where

$\underline{x}$ = observed noisy signal

$\underline{s}$ = clean signal

$\underline{w}$ = noise

This can be modeled using the general linear model with

$$\underline{\theta} = \underline{s}$$

$$H = I_N$$

and adopting a Gaussian prior for $\underline{s}$:

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

The Bayesian estimate for $\underline{s}$ is

$$\hat{\underline{s}} =$$

Application: Bandpass Filtering

Suppose we observe

$$\underline{x} = \underline{s} + \underline{w}$$

and we know a priori that the signal of interest occupies a certain passband.

In other words, $|\underline{u}_k^H \underline{x}|$ is large on average for certain DFT basis vectors $\underline{u}_k$, and small for others.

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

Let us assume we can specify

$$\sigma_k^2 = E\left\{ |\underline{u}_k^H \underline{s}|^2 \right\},$$

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

$\mathcal{U}$ = DFT matrix

$$\mathcal{U}^H \underline{S} \sim N(\underline{0}, \Sigma)$$

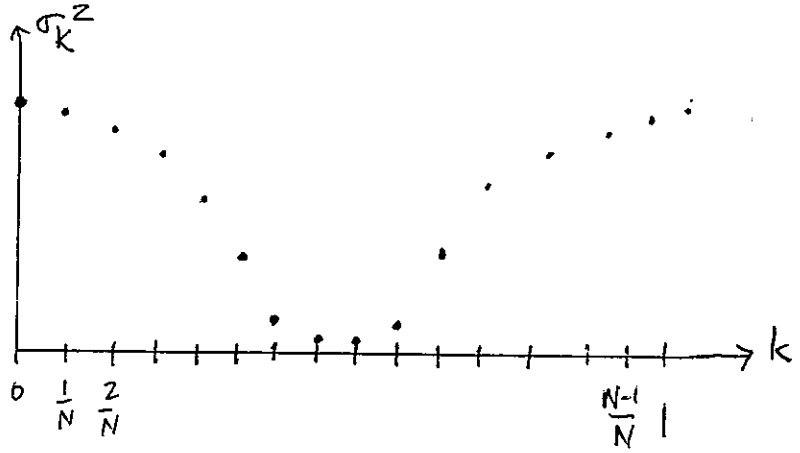
as a prior, where

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_N^2 \end{pmatrix}.$$

Equivalently, the prior on $\underline{S}$ is

$$\underline{S} \sim N(\underline{0}, \underbrace{\mathcal{U} \Sigma \mathcal{U}^H}_{R_{SS}})$$

For example



Note: $\sigma_k^2 = \sigma_{N-k}^2$ by conjugate symmetry

is a <u>lowpass</u> model.

Notice that the energy of $\underline{s}$ is

$$E\{\underline{s}^T\underline{s}\} = E\{(\mathcal{U}^H\underline{s})^H(\mathcal{U}^H\underline{s})\}$$

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

So to specify the σ_k^2 it suffices to know the signal energy and the <u>shape</u> of the frequency response.

Assume the noise is IID:

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

σ^2 known. Then the MMSE estimator is

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

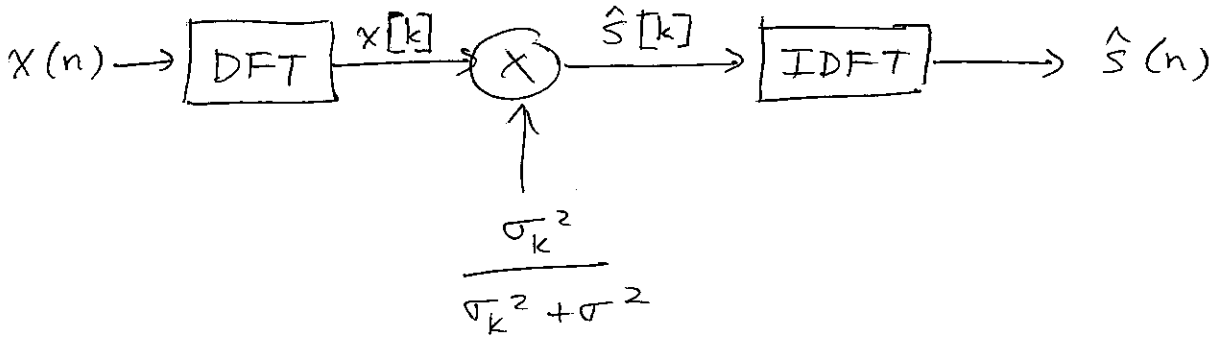
$$= \mathcal{U}\Sigma\mathcal{U}^H(\mathcal{U}[\Sigma + \sigma^2 I]\mathcal{U}^H)^{-1}\underline{x}$$

$$= \mathcal{U}[\Sigma(\Sigma + \sigma^2 I)^{-1}]\mathcal{U}^H\underline{x}$$

Note that

$$\Sigma\left(\Sigma+\sigma^2 I\right)^{-1} = \begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\sigma_0^2} & & & \\ & \frac{\sigma_2^2}{\sigma_2^2+\sigma_0^2} & & \\ & & \ddots & \\ & & & \frac{\sigma_N^2}{\sigma_N^2+\sigma^2} \end{bmatrix}$$

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

Extension

The preceding example generalizes easily to any situation where R_{ss} and R_{ww} are simultaneously diagonalizable.

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

$$\hat{\underline{s}} = U \underbrace{\left[\Lambda_s (\Lambda_s + \Lambda_w)^{-1} \right]}_{\Lambda} U^T \underline{x}$$

where

$$\Lambda = \begin{bmatrix} \frac{\lambda_1^s}{\lambda_1^s + \lambda_1^w} & & 0 \\ & \ddots & \\ 0 & & \frac{\lambda_N^s}{\lambda_N^s + \lambda_N^w} \end{bmatrix}$$

$\Longrightarrow$ "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_θ, R_w are simultaneously diagonalizable
  $\Rightarrow$ transform domain "shrinkage"
  e.g., bandpass filtering.

Key

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

$0, \mu_A$

$1, \bar{x}$

$0, \mu_A$