

# 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

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

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

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

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

Now apply the Gaussian conditioning principle.

It can be shown using the matrix inversion lemma that

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

and

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

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

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

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  $\sigma^2 \rightarrow \infty$ . This can be thought of as a "non committal" prior. Then  $R_{\theta}^{-1} \rightarrow O_p$  and

$$\begin{aligned}\hat{\underline{\theta}}(\underline{x}) &= \underline{\mu}_{\theta} + (H^T R_w^{-1} H + R_{\theta}^{-1})^{-1} H^T R_w^{-1} (\underline{x} - H \underline{\mu}_{\theta}) \\ &= (H^T R_w^{-1} H)^{-1} H^T R_w^{-1} \underline{x} \\ &= \text{MLE / MVUE}\end{aligned}$$

## Exercise

Suppose we observe

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

where  $A$  is an unknown scalar and

$$A \sim N(\mu_A, \sigma_A^2)$$

$$W_i \stackrel{\text{iid}}{\sim} N(0, \sigma_w^2)$$

} 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

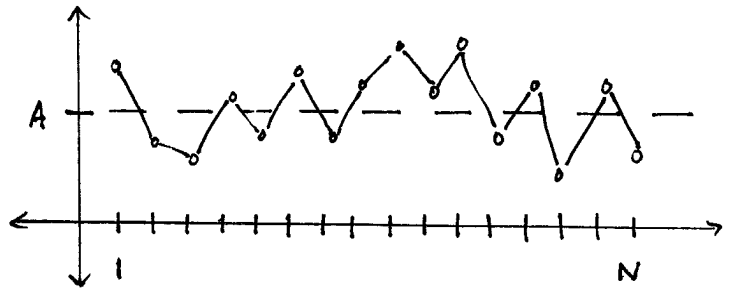
$$H = \underline{1} \quad (N \times 1)$$

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

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

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

$$R_w = \sigma^2 \mathbf{I}_N \quad (N \times N)$$



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

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

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

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

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

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)

$N \rightarrow \infty \Rightarrow \alpha \rightarrow$	$\Rightarrow \hat{A} \rightarrow$
$N = 0 \Rightarrow \alpha =$	$\Rightarrow \hat{A} =$
$\sigma_A^2 \rightarrow \infty \Rightarrow \alpha \rightarrow$	$\Rightarrow \hat{A} \rightarrow$
$\sigma_A^2 \rightarrow 0 \Rightarrow \alpha \rightarrow$	$\Rightarrow \hat{A} \rightarrow$



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

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

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

$$\underline{\theta} = \underline{\xi}$$

$$H = I_N$$

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

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

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

$$\hat{\underline{\xi}} =$$

Now suppose that  $R_{\xi\xi}$  and  $R_{ww}$  are simultaneously diagonalizable, meaning  $\exists$  an orthogonal matrix  $U$  such that

$$R_{\xi\xi} = U\Lambda_s U^T$$

and

$$R_{ww} = U\Lambda_w U^T$$

with  $\Lambda_s, \Lambda_w$  diagonal.

Example |  $R_{ww} = \sigma^2 I_N$  and  $R_{\xi\xi}$  is arbitrary

Then the estimator becomes

$$\begin{aligned}\hat{\underline{\xi}} &= R_{SS} (R_{SS} + R_{WW})^{-1} \underline{x} \\ &= U \Lambda_S U^T (U \Lambda_S U^T + U \Lambda_W U^T)^{-1} \underline{x} \\ &= U \Lambda_S U^T (U [\Lambda_S + \Lambda_W] U^T)^{-1} \underline{x} \\ &= U \cdot \underbrace{[\Lambda_S (\Lambda_S + \Lambda_W)^{-1}]}_{\Lambda} U^T \underline{x}\end{aligned}$$

where

$$\Lambda = \begin{bmatrix} \frac{\lambda_1^S}{\lambda_1^S + \lambda_1^W} & & & \\ & \frac{\lambda_2^S}{\lambda_2^S + \lambda_2^W} & & \\ & & \dots & \\ & & & \frac{\lambda_N^S}{\lambda_N^S + \lambda_N^W} \end{bmatrix}$$

Interpretation:

$U$  = change of basis matrix

$\underline{y} = U^T \underline{x}$  : coefficients of  $\underline{x}$  in new basis

$\underline{z} = \Lambda \underline{y}$  : coordinate-wise rescaling of  $\underline{y}$

$\hat{\underline{\xi}} = U \underline{z}$  : reconstruction of  $\underline{\xi}$  from  $\underline{z}$

How should we interpret the weights

$$\lambda_i = \frac{\lambda_i^s}{\lambda_i^s + \lambda_i^w} ?$$

Notice that  $\underline{u}^T \underline{x} = \underline{u}^T \underline{\varepsilon} + \underline{u}^T \underline{w}$  and

$$\underline{u}^T \underline{\varepsilon} \sim \mathcal{N}(\underline{0}, \underline{u}^T R_{ss} \underline{u}) = \mathcal{N}(\underline{0}, \Lambda_s)$$

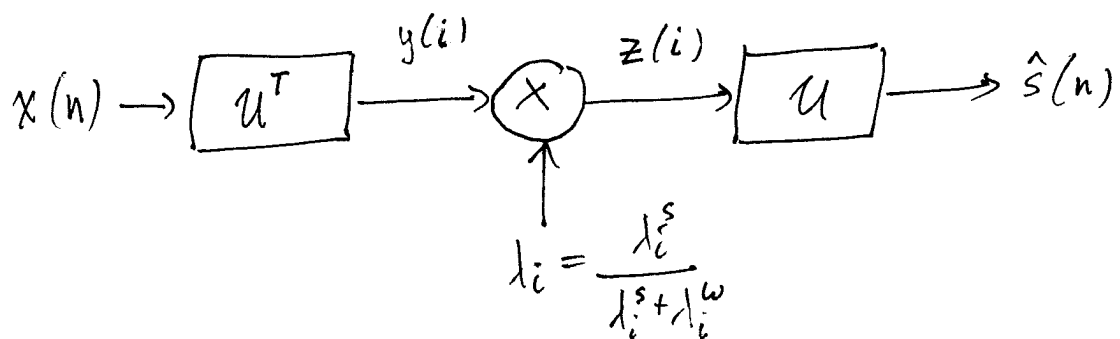
$$\underline{u}^T \underline{w} \sim \mathcal{N}(\underline{0}, \underline{u}^T R_{ww} \underline{u}) = \mathcal{N}(\underline{0}, \Lambda_w)$$

Writing  $\underline{u} = [\underline{u}_1, \dots, \underline{u}_N]$  we have

$$\underline{u}_i^T \underline{\varepsilon} \sim \mathcal{N}(0, \lambda_i^s)$$

$$\underline{u}_i^T \underline{w} \sim \mathcal{N}(0, \lambda_i^w)$$

Thus,  $\lambda_i$  reflects the proportion of the projection onto  $\underline{u}_i$  that is due to the signal.



“analysis → processing → synthesis”

## 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

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

the average signal energy at frequency  $k/N$ .

Let's also assume that signal content at different frequencies are independent.



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

$$\begin{aligned} E\{\underline{\Sigma}^T \underline{\Sigma}\} &= E\left\{(\underline{u}^H \underline{\Sigma})^H (\underline{u}^H \underline{\Sigma})\right\} \\ &= \sum_{k=0}^{N-1} \sigma_k^2 \end{aligned}$$

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

Assume the noise is IID:

$$R_{ww} = \sigma^2 \underline{I}_N,$$

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

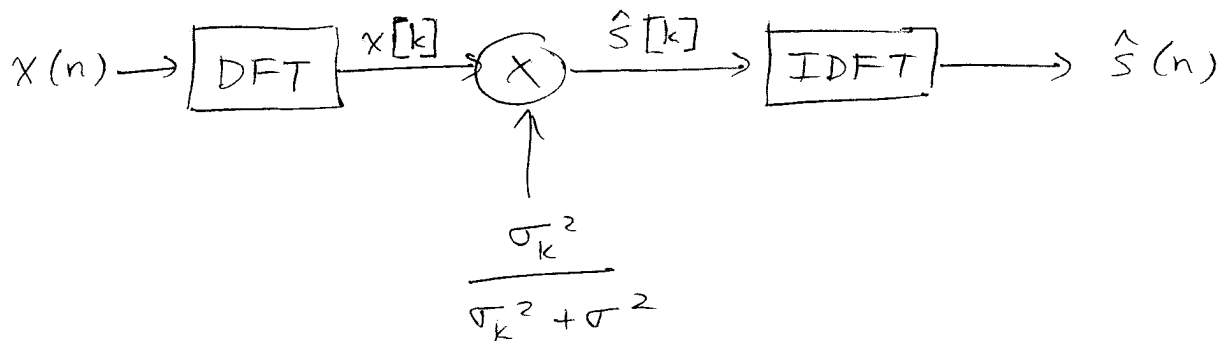
$$\begin{aligned} \hat{\underline{\Sigma}} &= R_{ss} (R_{ss} + R_{ww})^{-1} \underline{x} \\ &= \underline{u} \underline{\Sigma} \underline{u}^H (\underline{u} [\underline{\Sigma} + \sigma^2 \underline{I}] \underline{u}^H)^{-1} \underline{x} \\ &= \underline{u} [\underline{\Sigma} (\underline{\Sigma} + \sigma^2 \underline{I})^{-1}] \underline{u}^H \underline{x} \end{aligned}$$

Note that

$$\Sigma (\Sigma + \sigma^2 \mathbf{I})^{-1} =$$

$$\begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2 + \sigma^2} & & & \\ & \frac{\sigma_2^2}{\sigma_2^2 + \sigma^2} & & \\ & & \dots & \\ & & & \frac{\sigma_N^2}{\sigma_N^2 + \sigma^2} \end{bmatrix}$$

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



## 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.  $I, \bar{x}$

$0, \mu_A$

$I, \bar{x}$

$0, \mu_A$

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