5. Estimation

Statistical Signal Processing

Version 1.7

Prof. Peter Schreier, Ph.D.
Signal and System Theory Group
Faculty of Electrical Engineering, Computer Science, and Mathematics
Universität Paderborn

Motivation

Based on the square footage y of a house in Paderborn, we would like to estimate its price x. Thus, we are looking for a function f such that $\hat{x} = f(y)$ is a good **estimate** of x.

- In order to build such an estimator we need to have statistical information about x and y. This allows us to model square footage and price as random variables Y and X.
- In general, f may be a **nonlinear** function. In many practical situations, however, we restrict our estimators to be **linear**, i.e., $\hat{X} = aY + b$.
- If we know the first- and second-order moments μ_X, μ_Y, σ_X^2, σ_Y^2, and σ_{XY}, we may build a linear estimator $\hat{X} = aY + b$ that minimizes the mean-squared error (MSE) $E\{|\hat{X} - X|^2\}$ (cf. Chapter 2).
- The required first- and second-order moments are generally unknown and need to be estimated themselves from observations of X and Y.

Frequentist vs. Bayesian estimation

There is a fundamental difference between the cases where we:

- estimate a random variable from another random variable (called **Bayesian estimation**)
- estimate an unknown, but deterministic, parameter (called **frequentist estimation**)

An example of a frequentist problem is the estimation of the mean μ_X of a random variable X from n observations $x_1, \dots, x_n$ as

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The estimate $\hat{\mu}_X$ is a particular value of the **estimator**

$$\hat{M}_X = \frac{1}{n} \sum_{i=1}^{n} X_i$$

where we assume that the X_i are n **independent identically distributed** (i.i.d.) random variables, modeling observations of X.

5.1 Measures of performance

Bias and consistency (frequentist)

Important definitions

Let $\hat{\Theta}$ denote an estimator of a deterministic parameter θ computed from n observations $X_1, X_2, \ldots, X_n$.

- The **bias** is $b(\theta) = E\{\hat{\Theta}|\theta\} - \theta$. If $b(\theta) = 0$, then $\hat{\Theta}$ is called **unbiased**.
- The estimator is called **consistent** if it **converges in probability**

$$\lim_{n\to\infty} P\{|\hat{\Theta} - \theta| > \epsilon\} = 0 \text{ for any } \epsilon > 0$$

- An unbiased and consistent estimator for the **mean vector**:

$$\hat{\boldsymbol{\Theta}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$$

- An unbiased and consistent estimator for the **covariance matrix**:

$$\hat{\boldsymbol{\Theta}} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_X)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_X)^H$$

Mean-squared error (MSE)

By far the most common performance metric of an estimator is the mean-squared error (MSE). It can be defined for the frequentist and Bayesian approach, but the definition given here is Bayesian.

- Let $\hat{\mathbf{X}}$ be an estimator of $\mathbf{X}$ and $\mathbf{E} = \hat{\mathbf{X}} - \mathbf{X}$ the error, then the MSE is

$$E\{\|\hat{\mathbf{X}} - \mathbf{X}\|^2\} = E\{\mathbf{E}^H\mathbf{E}\} = \operatorname{tr} E\{\mathbf{E}\mathbf{E}^H\} = \operatorname{tr} \mathbf{M}$$

  where $\mathbf{M} = E\{\mathbf{E}\mathbf{E}^H\}$ is the **mean-squared error matrix**.
- The **Bayes bias** is $\mathbf{b} = \boldsymbol{\mu}_{\mathrm{E}}$ and the **error covariance matrix** is

$$\mathbf{Q} = E\{(\mathbf{E} - \boldsymbol{\mu}_{\mathrm{E}})(\mathbf{E} - \boldsymbol{\mu}_{\mathrm{E}})^H\} = \mathbf{M} - \mathbf{b}\mathbf{b}^H$$

  Thus, $\mathbf{M} = \mathbf{Q} + \mathbf{b}\mathbf{b}^H$. Designing a minimum-MSE estimator hence requires the **right tradeoff between error covariance and bias**.
- If the estimator is **unbiased**, the error has zero mean, and then $\mathbf{Q} = \mathbf{M}$.

5.2 Frequentist approaches to estimation

Nota Bene

For the remaining sections in this chapter, we use **lower-case letters** to denote both random variables and their samples. The difference should be clear from the context.

Maximum Likelihood (ML) estimator

- Let $f_{x|\theta}(\mathbf{x}|\boldsymbol{\theta})$ be the **likelihood function**, i.e., the pdf of a random vector $\mathbf{x}$ parametrized by the unknown deterministic parameter $\boldsymbol{\theta}$.
- Let's say we observe the value $\mathbf{x}_0$. The ML estimator $\hat{\boldsymbol{\theta}}$ picks the value of $\boldsymbol{\theta}$ that maximizes the likelihood function or, equivalently, the **log-likelihood function**:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f_{x|\theta}(\mathbf{x}_0|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \log f_{x|\theta}(\mathbf{x}_0|\boldsymbol{\theta})$$

Important properties

The ML estimator is **consistent**, but it may be **biased** (even substantially so). If the data is Gaussian, then the ML estimator minimizes the variance $E\{\|\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\}\|^2\}$.

- The ML estimator requires knowledge of the pdf (often unrealistic).
- If the pdf is differentiable, the ML estimate may be computed as (one of) the root(s) of

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{x|\theta}(\mathbf{x}_0|\boldsymbol{\theta}) = \mathbf{0}$$

MVDR estimator

Consider the **linear model**

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}$$

where $\mathbf{H} \in \mathbb{C}^{m \times n}$ is known, $\boldsymbol{\theta} \in \mathbb{C}^n$ is an unknown parameter, and $\mathbf{n}$ is an m-dim. random vector with mean zero and covariance matrix $\mathbf{R}_{nn}$

We are looking for a **linear minimum variance unbiased estimator** = minimum variance distortionless response (MVDR estimator) = best linear unbiased estimator (BLUE) of $\boldsymbol{\theta}$ from $\mathbf{y}$:

$$\hat{\boldsymbol{\theta}} = \mathbf{W}^H \mathbf{y}$$

- **Unbiasedness constraint:**
  $E\{\hat{\boldsymbol{\theta}}\} = E\{\mathbf{W}^H(\mathbf{H}\boldsymbol{\theta} + \mathbf{n})\} = \mathbf{W}^H\mathbf{H}\boldsymbol{\theta} + \mathbf{W}^H E\{\mathbf{n}\} = \boldsymbol{\theta} \Rightarrow \mathbf{W}^H\mathbf{H} = \mathbf{I}$
- **Minimize the variance**

$$\begin{aligned} E\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2\} &= \operatorname{tr} E\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^H\} \\ &= \operatorname{tr} \mathbf{R}_{\hat{\theta}\hat{\theta}} = \operatorname{tr}\{\mathbf{W}^H \mathbf{R}_{yy} \mathbf{W}\} = \operatorname{tr}\{\mathbf{W}^H \mathbf{R}_{nn} \mathbf{W}\} \end{aligned}$$

Solution to the MVDR estimation problem

MVDR optimization problem

The solution to

$$\min \operatorname{tr}\{\mathbf{W}^H \mathbf{R}_{nn} \mathbf{W}\} \text{ under constraint } \mathbf{W}^H \mathbf{H} = \mathbf{I}$$

is

$$\mathbf{W}^H = (\mathbf{H}^H \mathbf{R}_{nn}^{-1} \mathbf{H})^{-1} \mathbf{H}^H \mathbf{R}_{nn}^{-1}$$

which results in $\mathbf{Q} = (\mathbf{H}^H \mathbf{R}_{nn}^{-1} \mathbf{H})^{-1}$.

- The MVDR estimator (= ML estimator for Gaussian data) is

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^H \mathbf{R}_{nn}^{-1} \mathbf{H})^{-1} \mathbf{H}^H \mathbf{R}_{nn}^{-1} \mathbf{y} = \boldsymbol{\theta} + (\mathbf{H}^H \mathbf{R}_{nn}^{-1} \mathbf{H})^{-1} \mathbf{H}^H \mathbf{R}_{nn}^{-1} \mathbf{n}$$

  (Notice that it is unbiased)

- If $\mathbf{R}_{nn} = \mathbf{I}$, then

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \mathbf{y} = \mathbf{H}^\dagger \mathbf{y}$$

  This is also the solution of the deterministic Least-Squares (LS) problem, where we minimize $\|\mathbf{y} - \mathbf{H}\boldsymbol{\theta}\|^2$ for given deterministic $\mathbf{y}$.

5.3 Bayesian approaches to estimation

Conditional mean estimator

- We would like to construct an estimator $\hat{\mathbf{x}}$ to estimate a random vector $\mathbf{x}$ from another random vector $\mathbf{y}$.
- Its mean-squared error matrix is

$$
\begin{aligned}
\mathbf{Q} &= E[(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^H] \\
&= E[(\hat{\mathbf{x}}-E[\mathbf{x}|\mathbf{y}]+E[\mathbf{x}|\mathbf{y}]-\mathbf{x})(\hat{\mathbf{x}}-E[\mathbf{x}|\mathbf{y}]+E[\mathbf{x}|\mathbf{y}]-\mathbf{x})^H] \\
&= E[(E[\mathbf{x}|\mathbf{y}]-\mathbf{x})(E[\mathbf{x}|\mathbf{y}]-\mathbf{x})^H] + E[(\hat{\mathbf{x}}-E[\mathbf{x}|\mathbf{y}])(E[\mathbf{x}|\mathbf{y}]-\mathbf{x})^H] \\
&\quad + E[(E[\mathbf{x}|\mathbf{y}]-\mathbf{x})\underbrace{(\hat{\mathbf{x}}-E[\mathbf{x}|\mathbf{y}])^H}_{\mathrm{g}^H(\mathrm{y})}] + E[(\hat{\mathbf{x}}-E[\mathbf{x}|\mathbf{y}])(\hat{\mathbf{x}}-E[\mathbf{x}|\mathbf{y}])^H]
\end{aligned}
$$

- Law of total expectation:

$$
E\{E[(E[\mathbf{x}|\mathbf{y}]-\mathbf{x})\mathbf{g}^H(\mathbf{y})|\mathbf{y}]\} = \mathbf{0}
$$

makes second and third terms in $\mathbf{Q}$ zero

The optimum estimator is the **conditional mean estimator** $\hat{\mathbf{x}} = E[\mathbf{x}|\mathbf{y}]$.

Conditional mean estimator

Consider the error vector $\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x} = E[\mathbf{x}|\mathbf{y}] - \mathbf{x}$.

- $E[\mathbf{e}] = \mathbf{0}$, and thus $E[\hat{\mathbf{x}}] = E[\mathbf{x}]$. This says that $\hat{\mathbf{x}}$ is an **unbiased estimator** of $\mathbf{x}$.
- The covariance matrix of the error vector is

$$\mathbf{Q} = E[\mathbf{e}\mathbf{e}^H] = E[(E[\mathbf{x}|\mathbf{y}] - \mathbf{x})(E[\mathbf{x}|\mathbf{y}] - \mathbf{x})^H].$$

  Any competing estimator $\hat{\mathbf{x}}'$ with error covariance matrix $\mathbf{Q}' = E[(\hat{\mathbf{x}}' - \mathbf{x})(\hat{\mathbf{x}}' - \mathbf{x})^H]$ will have $\mathbf{Q}' \geq \mathbf{Q}$.
- As a consequence, the conditional mean estimator is a **minimum mean-squared error (MMSE)** estimator:

$$E\|\mathbf{e}\|^2 = \operatorname{tr}\mathbf{Q} \leq \operatorname{tr}\mathbf{Q}' = E\|\mathbf{e}'\|^2.$$

Orthogonality principle

The error vector $\mathbf{e}$ is orthogonal to every measurable function of $\mathbf{y}$, $\mathbf{g}(\mathbf{y})$. That is, $E[\mathbf{e}\mathbf{g}^H(\mathbf{y})] = E\{(E[\mathbf{x}|\mathbf{y}] - \mathbf{x})\mathbf{g}^H(\mathbf{y})\} = \mathbf{0}$.

Linear MMSE estimation

Let's consider the n-dim. **signal** (message) $\mathbf{x}$ and the m-dim. **observation** (measurement) $\mathbf{y}$, both zero mean. Their composite covariance matrix is the matrix

$$\mathbb{R}_{xy} = E\left\{ \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \begin{bmatrix} \mathbf{x}^H & \mathbf{y}^H \end{bmatrix} \right\} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{xy}^H & \mathbf{R}_{yy} \end{bmatrix}.$$

The error between the signal $\mathbf{x}$ and the linear estimator $\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}$ is $\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x}$ and the error covariance matrix is $\mathbf{Q} = E[(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^H]$:

$$\mathbf{Q} = E[(\mathbf{W}\mathbf{y} - \mathbf{x})(\mathbf{W}\mathbf{y} - \mathbf{x})^H] = \mathbf{W}\mathbf{R}_{yy}\mathbf{W}^H - \mathbf{R}_{xy}\mathbf{W}^H - \mathbf{W}\mathbf{R}_{xy}^H + \mathbf{R}_{xx}.$$

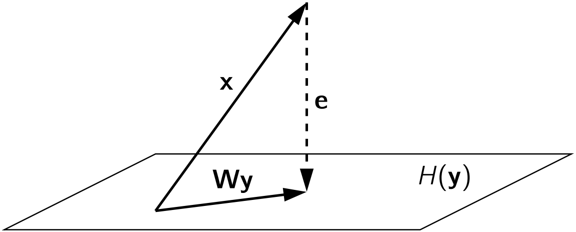
After completing the square, this may be written

$$\mathbf{Q} = \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^H + (\mathbf{W} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})\mathbf{R}_{yy}(\mathbf{W} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})^H.$$

This quadratic form in $\mathbf{W}$ is positive semidefinite, so $\mathbf{Q} \geq \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^H$ with equality for

$$\mathbf{W} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1} \quad \text{and} \quad \mathbf{Q} = \mathbf{R}_{xx} - \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{R}_{xy}^H.$$

Orthogonality principle



The solution for $\mathbf{W}$ may be written as the solution to the **normal** equations $\mathbf{W}\mathbf{R}_{yy} - \mathbf{R}_{xy} = \mathbf{0}$, or

$$E[(\mathbf{W}\mathbf{y} - \mathbf{x})\mathbf{y}^H] = \mathbf{0}.$$

The estimator error $\mathbf{e} = \mathbf{W}\mathbf{y} - \mathbf{x}$ is **orthogonal** to the measurement $\mathbf{y}$.

Nonzero means

- What if the signal has known mean $\boldsymbol{\mu}_x$ and the measurement has known mean $\boldsymbol{\mu}_y$?
- The centered signal and measurement $\mathbf{x} - \boldsymbol{\mu}_x$ and $\mathbf{y} - \boldsymbol{\mu}_y$ then share the composite covariance matrix $\mathbb{R}_{xy}$.
- The LMMSE estimator of $\mathbf{x} - \boldsymbol{\mu}_x$ from $\mathbf{y} - \boldsymbol{\mu}_y$ obeys all of the equations already derived:

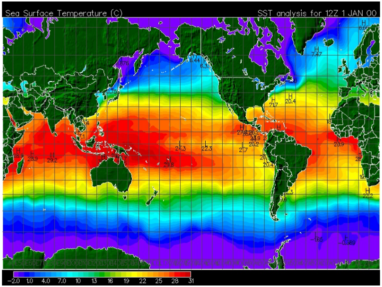
$$\hat{\mathbf{x}} - \boldsymbol{\mu}_x = \mathbf{W}(\mathbf{y} - \boldsymbol{\mu}_y) \Leftrightarrow \hat{\mathbf{x}} = \mathbf{W}(\mathbf{y} - \boldsymbol{\mu}_y) + \boldsymbol{\mu}_x$$

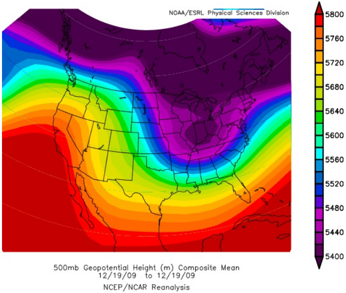
- The **orthogonality principle** says the error between the estimator and the signal is orthogonal to the measurement:

$$\begin{aligned} E[(\hat{\mathbf{x}} - \mathbf{x})\mathbf{y}^H] &= E\{[\hat{\mathbf{x}} - \boldsymbol{\mu}_x - (\mathbf{x} - \boldsymbol{\mu}_x)](\mathbf{y} - \boldsymbol{\mu}_y)^H\} \\ &\quad + E\{[\hat{\mathbf{x}} - \boldsymbol{\mu}_x - (\mathbf{x} - \boldsymbol{\mu}_x)]\boldsymbol{\mu}_y^H\} = \mathbf{0} \end{aligned}$$

  The first term on the right is zero due to the orthogonality principle already established for zero mean LMMSE estimators. The second term is zero because $\hat{\mathbf{x}}$ is an unbiased estimator of $\mathbf{x}$, i.e., $E[\hat{\mathbf{x}}] = \boldsymbol{\mu}_x$.

We would like to estimate the 500 mb height ($\approx$ air pressure) over the USA (right figure) from the sea surface temperature (left figure).





- Arrange measurements of the 500 mb height in a vector $\mathbf{x}$ and measurements of the sea surface temperature in a vector $\mathbf{y}$
- We need to estimate the covariance matrices of $\mathbf{x}$ and $\mathbf{y}$ from sample (observation) pairs $(\mathbf{x}_i, \mathbf{y}_i)$. Which assumptions are necessary?

Tradeoff between bias and variance

Let's apply the LMMSE estimator to the linear model $\mathbf{y} = \mathbf{Hx} + \mathbf{n}$ (with $\mathbf{x}$ and $\mathbf{n}$ zero mean):

$$\hat{\mathbf{x}} = \mathbf{Wy} = \mathbf{W}(\mathbf{Hx} + \mathbf{n}) = \mathbf{WHx} + \mathbf{Wn}$$

- The LMMSE filter $\mathbf{W}$ does not equalize $\mathbf{H}$ to produce $\mathbf{WH} = \mathbf{I}$.
- Rather, it approximates $\mathbf{I}$ so that the error

$$\mathbf{e} = \mathbf{Wy} - \mathbf{x} = (\mathbf{WH} - \mathbf{I})\mathbf{x} + \mathbf{Wn}$$

with covariance matrix

$$\mathbf{Q} = \underbrace{(\mathbf{WH} - \mathbf{I})\mathbf{R}_{xx}(\mathbf{WH} - \mathbf{I})^H}_{\text{model-bias-squared}} + \underbrace{\mathbf{WR}_{nn}\mathbf{W}^H}_{\text{filtered noise variance}}$$

provides the best tradeoff between model-bias-squared and filtered noise variance to minimize the error covariance matrix $\mathbf{Q}$.

Comparison with MVDR estimator

We can write the LMMSE estimator for the linear model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ as

$$\begin{aligned}\hat{\mathbf{x}} &= \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{y} = \mathbf{R}_{xx}\mathbf{H}^H(\mathbf{R}_{nn} + \mathbf{H}\mathbf{R}_{xx}\mathbf{H}^H)^{-1}\mathbf{y} \\ &= (\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H} + \mathbf{R}_{xx}^{-1})^{-1}\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{y} \\ &= (\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H} + \mathbf{R}_{xx}^{-1})^{-1}\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H}\mathbf{x} + (\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H} + \mathbf{R}_{xx}^{-1})^{-1}\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{n}.\end{aligned}$$

Let's compare that with the MVDR estimator for the linear model $\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n}$ with deterministic parameter $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H})^{-1}\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{y} = \boldsymbol{\theta} + (\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H})^{-1}\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{n}$$

- The **frequentist bias** of the LMMSE estimator is $\mathbf{b}(\mathbf{x}) = (\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H} + \mathbf{R}_{xx}^{-1})^{-1}\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H}\mathbf{x} - \mathbf{x}$ but its **Bayes bias** is $\mathbf{b} = \mathbf{0}$. On the other hand, the MVDR estimator has zero frequentist and Bayes bias.
- The error covariances are related as

$$\mathbf{Q}_{\mathrm{LMMSE}} = (\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H} + \mathbf{R}_{xx}^{-1})^{-1} \leq (\mathbf{H}^H\mathbf{R}_{nn}^{-1}\mathbf{H})^{-1} = \mathbf{Q}_{\mathrm{MVDR}}$$

Gaussian case

If the composite vector $[\mathbf{x}^T, \mathbf{y}^T]^T$ is multivariate Gaussian with zero mean and composite covariance matrix $\mathbb{R}_{xy}$, then the conditional pdf for $\mathbf{x}$, given $\mathbf{y}$, is

$$f_{x|y}(\mathbf{x}|\mathbf{y}) = \frac{f_{xy}(\mathbf{x}, \mathbf{y})}{f_y(\mathbf{y})} = \frac{\det \mathbf{R}_{yy}}{\pi^n \det \mathbb{R}_{xy}} \exp\left\{ - \begin{bmatrix} \mathbf{x}^H & \mathbf{y}^H \end{bmatrix} \mathbb{R}_{xy}^{-1} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} + \mathbf{y}^H \mathbf{R}_{yy}^{-1} \mathbf{y} \right\}.$$

Using the identity $\det \mathbb{R}_{xy} = \det \mathbf{R}_{yy} \det \mathbf{Q}$ and one of the factorizations for $\mathbb{R}_{xy}^{-1}$ given in Chapter 3, this pdf may be written

$$f_{x|y}(\mathbf{x}|\mathbf{y}) = \frac{1}{\pi^n \det \mathbf{Q}} \exp\left\{ -(\mathbf{x} - \mathbf{W}\mathbf{y})^H \mathbf{Q}^{-1} (\mathbf{x} - \mathbf{W}\mathbf{y}) \right\}.$$

Thus the posterior pdf for $\mathbf{x}$, given $\mathbf{y}$, is Gaussian with conditional mean $\mathbf{W}\mathbf{y}$ and conditional covariance $\mathbf{Q}$.

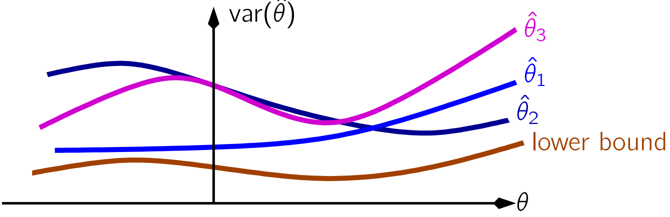
Conditional mean estimator for Gaussian data

For jointly Gaussian data, the **conditional mean estimator** $E\{\mathbf{x}|\mathbf{y}\}$ is the **linear MMSE estimator $\mathbf{W}\mathbf{y}$**.

5.4 Frequentist performance bounds for parameter estimation

Estimator accuracy considerations

- We would like to find a **lower bound** on the **error variance** of an unbiased estimator.



  - There may or may not be an estimator that achieves the bound.
- The **estimation accuracy** depends directly on the likelihood function, i.e., the pdf $f_\theta(y) = f(y|\theta)$ of the measurement y conditioned on the parameter θ we would like to estimate.
- The more the likelihood function $f_\theta(y)$ is influenced by θ, the more accurately we should be able to estimate it.
  - **Example:** Consider $y = \theta + n$ with $n \sim N(0, \sigma^2)$
  - If σ^2 is small, then we should be able to estimate θ more accurately.

Cramér-Rao bound for scalar parameters

A special case of the **Cramér-Rao bound** (which we will prove later) says that the variance of any unbiased estimator $\hat{\theta}$ must satisfy

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{-E\left[\frac{\partial^2 \log f_\theta(y)}{\partial \theta^2}\right]}$$

- The "sharpness" of the likelihood function $f_\theta(y)$ should determine how accurately we can determine θ.
- This sharpness is measured by the average curvature of the log-likelihood function.

Example:

- Estimating the phase θ of a sinusoid with known amplitude A and frequency f_0 in AWGN with variance σ^2:

$$y[n] = A\cos(2\pi f_0 n + \theta) + w[n], \quad n = 0, 1, \ldots, N-1.$$

- The Cramér-Rao bound is $\mathrm{var}(\hat{\theta}) \geq \frac{2\sigma^2}{NA^2}$.

General quadratic frequentist bounds

Now let's talk about the general case.

Notation: From the measurement $\mathbf{y}$ we compute the estimator $\hat{\boldsymbol{\theta}}(\mathbf{y})$, which estimates the parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. The likelihood of the measurement is $f_{\boldsymbol{\theta}}(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})$. The estimation error is $\mathbf{e}(\mathbf{y}) = \hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}$ and the **centered error** is $\boldsymbol{\epsilon}(\mathbf{y}) = \mathbf{e}(\mathbf{y}) - \mathbf{b}(\boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}(\mathbf{y}) - E[\hat{\boldsymbol{\theta}}(\mathbf{y})]$.

- Let's define the function $\mathbf{s}(\mathbf{y}) = [s_1(\mathbf{y}), s_2(\mathbf{y}), \cdots, s_m(\mathbf{y})]^T$ to be an m-dimensional vector of **measurement scores**.
- The mean of the score is $E[\mathbf{s}(\mathbf{y})]$ and the **centered measurement score** is $\boldsymbol{\sigma}(\mathbf{y}) = \mathbf{s}(\mathbf{y}) - E[\mathbf{s}(\mathbf{y})]$.
- The centered measurement score is a judiciously-chosen function of the measurement that brings information about the centered error $\boldsymbol{\epsilon}(\mathbf{y})$.

We would like to approximate this centered error by using a linear function of the centered measurement score $\boldsymbol{\sigma}(\mathbf{y})$. This will lead to a bound on the error covariance matrix $\mathbf{Q}(\boldsymbol{\theta}) = E[\boldsymbol{\epsilon}(\mathbf{y})\boldsymbol{\epsilon}^H(\mathbf{y})]$.

Quadratic frequentist bound

Let's **linearly** estimate the centered error $\boldsymbol{\epsilon}(\mathbf{y})$ from the centered measurement score $\boldsymbol{\sigma}(\mathbf{y})$ using the LMMSE estimator

$$\hat{\epsilon}(\mathbf{y}) = \mathbf{T}^H(\theta)\mathbf{J}^{-1}(\theta)\boldsymbol{\sigma}(\mathbf{y})$$

with

- the **sensitivity matrix** $\mathbf{T}(\theta) = E[\boldsymbol{\sigma}(\mathbf{y})\boldsymbol{\epsilon}^H(\mathbf{y})]$
- the **information matrix** $\mathbf{J}(\theta) = E[\boldsymbol{\sigma}(\mathbf{y})\boldsymbol{\sigma}^H(\mathbf{y})]$

This estimator has error covariance matrix $\mathbf{Q}(\theta) - \mathbf{T}^H(\theta)\mathbf{J}^{-1}(\theta)\mathbf{T}(\theta)$, which must be positive semidefinite. Hence:

Quadratic frequentist bound

The error covariance matrix is bounded by the general quadratic frequentist bound

$$\mathbf{Q}(\theta) \geq \mathbf{T}^H(\theta)\mathbf{J}^{-1}(\theta)\mathbf{T}(\theta),$$

and the mean-squared error matrix is bounded as

$$\mathbf{M}(\theta) \geq \mathbf{T}^H(\theta)\mathbf{J}^{-1}(\theta)\mathbf{T}(\theta) + \mathbf{b}(\theta)\mathbf{b}^H(\theta).$$

Fisher score

Our results for quadratic frequentist bounds so far are general. To make them applicable we need to consider a concrete score for which $\mathbf{T}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ can be computed. For this we choose the Fisher score:

Fisher score

The **Fisher score** is defined as

$$\mathbf{s}(\mathbf{y}) = \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\mathbf{y})\right]^T = \left[\frac{\partial}{\partial \theta_1} \log f_{\boldsymbol{\theta}}(\mathbf{y}), \cdots, \frac{\partial}{\partial \theta_p} \log f_{\boldsymbol{\theta}}(\mathbf{y})\right]^T,$$

where the partial derivatives are evaluated at the true value of $\boldsymbol{\theta}$.

The Fisher score has a number of important properties:

1. We may write the partial derivative as

$$\frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{\theta}}(\mathbf{y}) = \frac{1}{f_{\boldsymbol{\theta}}(\mathbf{y})} \frac{\partial}{\partial \theta_i} f_{\boldsymbol{\theta}}(\mathbf{y}),$$

   which is a normalized measure of the sensitivity of the pdf $f_{\boldsymbol{\theta}}(\mathbf{y})$ to variations in the parameter θ_i. Large sensitivity is valued and this will be measured by the variance of the score.

Fisher score

2. The Fisher score is a zero-mean random variable, hence $\boldsymbol{\sigma}(\mathbf{y}) = \mathbf{s}(\mathbf{y})$.
3. The cross-correlation between the centered Fisher score and the centered error score is the sensitivity matrix

$$\mathbf{T}(\boldsymbol{\theta}) = \mathbf{I} + \left[\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{b}(\boldsymbol{\theta})\right]^H .$$

   The (i,j)th element of the matrix $(\partial/\partial\boldsymbol{\theta})\mathbf{b}(\boldsymbol{\theta})$ is $(\partial/\partial\theta_j)b_i(\boldsymbol{\theta})$. When the estimator is unbiased, then $\mathbf{T} = \mathbf{I}$.
4. The **Fisher information matrix** is the expected Hessian of the score function (up to a minus sign):

$$\mathbf{J}_{\mathrm{F}}(\boldsymbol{\theta}) = E\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\mathbf{y})\right]^T \frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\mathbf{y})\right\} = -E\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\mathbf{y})\right]^T\right\}$$

Cramér-Rao bound

Cramér-Rao bound

The quadratic frequentist bound for Fisher score is the **Cramér-Rao bound (CRB)**:

$$\mathbf{Q}(\boldsymbol{\theta}) \geq \left[\mathbf{I} + \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{b}(\boldsymbol{\theta})\right] \mathbf{J}_{\mathrm{F}}^{-1}(\boldsymbol{\theta}) \left[\mathbf{I} + \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{b}(\boldsymbol{\theta})\right]^{H}$$

For an **unbiased estimator**, the Cramér-Rao bound is

$$\mathbf{Q}(\boldsymbol{\theta}) \geq \mathbf{J}_{\mathrm{F}}^{-1}(\boldsymbol{\theta})$$

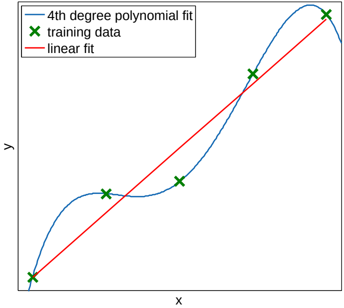
where $\mathbf{J}_{\mathrm{F}}(\boldsymbol{\theta})$ is the **Fisher information matrix**.

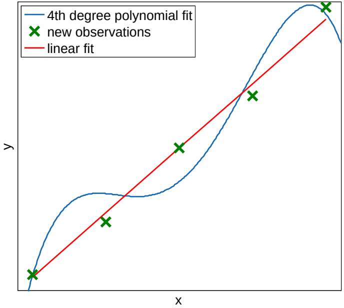
- The CRB is only a **bound**. An estimator that achieves it (which does not always exist) is called **efficient**.
- If repeated measurements carry information about a fixed $\boldsymbol{\theta}$ through the product pdf $\prod_{i=1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i)$, then $\mathbf{T}(\theta)$ remains fixed and $\mathbf{J}_{\mathrm{F}}(\theta)$ scales with N, hence the CRB decreases as N^{-1}.
- An ML estimator is **asymptotically** (for $N \to \infty$) **efficient**.

5.5 Reduced-rank estimation

Principle of parsimony

Example:



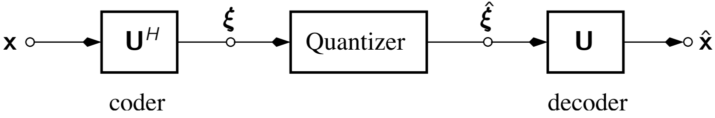


Rank reduction follows the principle of parsimony:

One should seek the **simplest possible model** to describe the phenomenon under study in order to **avoid overfitting to random noise** fluctuations. Rank reduction is a matter of finding the **right bias-variance trade-off**.

Example: Transform coder

Consider the following simplified model of a **transform coder**:



The coder and decoder are assumed to be unitary. Let's model the quantizer as an additive zero-mean white noise source: $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi} + \mathbf{n}$, with $\mathbf{R}_{nn} = \sigma_n^2 \mathbf{I}$, uncorrelated with the data.

One can show that in order to minimize the MSE $E\{\|\hat{\mathbf{x}} - \mathbf{x}\|^2\}$, we

- choose the coder and decoder as the matrix of eigenvectors of $\mathbf{R}_{xx}$, i.e., $\mathbf{R}_{xx} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H$. The components of $\boldsymbol{\xi} = [\xi_1, ..., \xi_n]$ are the **principal components**.
- keep only those principal components whose variance exceeds the noise level: $\sigma_{\xi_i}^2 > \sigma_n^2$. The other components are replaced with zeros. This is a **bias-variance trade-off**.

LMMSE filtering with PCA preprocessing step

What happens in the LMMSE estimator

$$\hat{\mathbf{x}} = \mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}\mathbf{y}$$

when $\mathbf{R}_{yy}$ is singular or close to singular? This can easily happen when $\mathbf{R}_{yy}$ is **estimated** from samples.

- We can replace $\mathbf{R}_{yy}^{-1}$ with its pseudo-inverse $\mathbf{R}_{yy}^{\dagger}$. However, we might still run into numerical problems if $\mathbf{R}_{yy}$ is close to singular (ill-conditioned).
- A more stable solution is to compute a rank r pseudo-inverse by considering only the largest r principal components:

$$\mathbf{R}_{yy}^{\dagger_r} = \mathbf{U}^H \mathbf{\Lambda}_r^{-1} \mathbf{U} \qquad (\mathbf{U} : \text{eigenvectors of } \mathbf{R}_{yy})$$

  where $\mathbf{\Lambda}_r^{-1} = \mathbf{Diag}\left(\lambda_1^{-1}, \lambda_2^{-1}, ..., \lambda_r^{-1}, 0, ..., 0\right)$. This is also suitable as a dimension-reduction step if the dimension of $\mathbf{y}$ is very large.
- Problem: The components that contribute most to the variance of $\mathbf{y}$ are not necessarily those that are most strongly correlated with $\mathbf{x}$.