Dynamic Model Learning: A Geometric Perspective

Lecture Notes in FEL3201/FEL3202

Håkan Hjalmarsson

*-marked sections and exercises are for FEL3202
# Contents

## 1 Background Theory

1.1 Vector Spaces .......................... 9
   1.1.1 Inner Product Spaces ................. 9
   1.1.2 Subspaces and Orthogonal Projections .. 11
   1.1.3 Hilbert Spaces .......................... 12
   1.1.4 Orthonormal bases ..................... 13

1.2 Probability Theory ..................... 14
   1.2.1 Information Contents in Events, Bayes Rule and Independence ..... 15
   1.2.2 Random Variables ....................... 15
   1.2.3 Expectation .......................... 16
   1.2.4 Information Contents in Random Variables and Conditional Probabilities . . 16
   1.2.5 Independent Random Variables ........... 19

1.3 References ................................ 19

## 2 Introduction to Estimation

2.1 Introduction .......................... 21

2.2 Minimum Mean Square Error Estimation of Random Variables ............. 21
   2.2.1 Introduction .......................... 21
   2.2.2 The Internal Structure of the Conditional Expectation .......... 22
   2.2.3 A Hilbert Space Interpretation ..................... 24
   2.2.4 Linear Estimators .......................... 26
   2.2.5 Maximum A Posteriori Estimation .................... 27
   2.2.6 Other Estimation Criteria* ..................... 27
   2.2.7 Exercises .......................... 27

2.3 Unbiased Parameter Estimation ........................................ 28
   2.3.1 Introduction .......................... 28
   2.3.2 The Cramér-Rao Lower Bound ...................... 29
   2.3.3 Efficient Estimators ....................... 30
      Efficient Estimators for Exponential Families ............. 30
      Efficient Estimators for General Distributions .......... 31
   2.3.4 The Maximum Likelihood Estimator .................. 32
   2.3.5 Data compression .......................... 32
      Maximal data compression ....................... 33
      Finite dimensional data compression ............. 33
   2.3.6 Uniform Minimum Variance Unbiased (UMVU) Estimators ............ 34
      UMVU for exponential families ..................... 36
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalizing the Rao-Blackwell theorem*</td>
<td>36</td>
</tr>
<tr>
<td>2.3.7 Efficient Estimators, Sufficient Statistics and Maximum Likelihood Estimators</td>
<td>37</td>
</tr>
<tr>
<td>2.3.8 Parameter Transformations</td>
<td>37</td>
</tr>
<tr>
<td>The Invariance Principle</td>
<td>37</td>
</tr>
<tr>
<td>Efficient Estimators</td>
<td>38</td>
</tr>
<tr>
<td>2.3.9 Best Linear Unbiased Estimators (BLUE)</td>
<td>39</td>
</tr>
<tr>
<td>2.3.10 Exercises</td>
<td>41</td>
</tr>
<tr>
<td>2.4 Estimation vs Parameter Estimation</td>
<td>43</td>
</tr>
<tr>
<td>2.5 Using Estimation for Parameter Estimation</td>
<td>45</td>
</tr>
<tr>
<td>2.6 References</td>
<td>46</td>
</tr>
<tr>
<td>2.A Block Matrix Inversion</td>
<td>47</td>
</tr>
<tr>
<td>2.A.1 Inverse of an Inner Product</td>
<td>47</td>
</tr>
<tr>
<td>2.B Schur Complement</td>
<td>47</td>
</tr>
<tr>
<td>3 Asymptotic Theory</td>
<td>49</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>49</td>
</tr>
<tr>
<td>3.2 Limits of Random Variables</td>
<td>50</td>
</tr>
<tr>
<td>3.2.1 Convergence in Mean</td>
<td>50</td>
</tr>
<tr>
<td>3.2.2 Convergence in Probability</td>
<td>51</td>
</tr>
<tr>
<td>3.2.3 Convergence with Probability 1</td>
<td>51</td>
</tr>
<tr>
<td>3.2.4 Convergence in Distribution</td>
<td>52</td>
</tr>
<tr>
<td>3.2.5 Relations Between Convergence Concepts</td>
<td>52</td>
</tr>
<tr>
<td>3.3 Large Sample Properties of Estimators</td>
<td>54</td>
</tr>
<tr>
<td>3.3.1 Asymptotics of Explicit Estimators</td>
<td>54</td>
</tr>
<tr>
<td>Consistency</td>
<td>54</td>
</tr>
<tr>
<td>Convergence in Distribution</td>
<td>55</td>
</tr>
<tr>
<td>3.3.2 Criterion Based Estimators</td>
<td>56</td>
</tr>
<tr>
<td>Consistency</td>
<td>56</td>
</tr>
<tr>
<td>Convergence in Distribution</td>
<td>57</td>
</tr>
<tr>
<td>3.3.3 Asymptotic Efficiency</td>
<td>59</td>
</tr>
<tr>
<td>3.3.4 Superefficient Estimators*</td>
<td>60</td>
</tr>
<tr>
<td>3.3.5 The Maximum Likelihood Estimator</td>
<td>61</td>
</tr>
<tr>
<td>Consistency</td>
<td>60</td>
</tr>
<tr>
<td>Asymptotic Efficiency</td>
<td>61</td>
</tr>
<tr>
<td>3.4 Parameter Transformations</td>
<td>63</td>
</tr>
<tr>
<td>3.4.1 Consistent Estimators</td>
<td>63</td>
</tr>
<tr>
<td>3.4.2 The Extended Invariance Principle (EXIP)*</td>
<td>64</td>
</tr>
<tr>
<td>3.4.3 Beyond the EXIP*</td>
<td>67</td>
</tr>
<tr>
<td>3.5 Using Estimation for Parameter Estimation</td>
<td>69</td>
</tr>
<tr>
<td>3.5.1 Exercises</td>
<td>71</td>
</tr>
<tr>
<td>3.6 References</td>
<td>71</td>
</tr>
<tr>
<td>3.A Asymptotic Dominance of the Variance Term</td>
<td>72</td>
</tr>
<tr>
<td>3.A.1 MSE Dominating the CRLB</td>
<td>72</td>
</tr>
<tr>
<td>3.A.2 Optimal MSE</td>
<td>74</td>
</tr>
</tbody>
</table>
5.3 Estimation Methods .................................................. 119
   5.3.1 Maximum Likelihood Estimation ............................... 119
      Direct calculation .............................................. 119
      Model Inversion ............................................... 121
      Marginalization ............................................... 122
      Filtering ...................................................... 123
      Filtering for Nonlinear State-Space Models .................. 123
      The Kalman Filter ............................................ 124
      The Innovations Model ....................................... 125
      The Stationary Kalman Filter ................................ 125
      Spectral Factorization ...................................... 127
      Relation to the Direct Method ................................ 128
      The Non-Gaussian Case ....................................... 129
      Unstable Systems .............................................. 130
   5.3.2 The Prediction Error Method ................................ 130
   5.3.3 Relation to Maximum Likelihood Estimation ............... 131
      Linear Models ................................................ 131
      Nonlinear Models ............................................ 132
   5.3.4 The Extended Invariance Principle ........................ 133
   5.3.5 Multi-Step Least-Squares Methods ......................... 133
   5.3.6 Instrumental Variable Methods ............................ 133
   5.3.7 Indirect Inference ........................................ 133
   5.3.8 Exercises .................................................. 133
   5.3.9 Computer exercises ...................................... 135
      Multi-Step Least-Squares Methods ......................... 141
   5.3.10 The Prediction Error Method .............................. 141
      Multi-Step Least-Squares Methods ......................... 141
      Subspace Identification ...................................... 141
      Instrumental Variable Methods .............................. 141
      Bayesian Methods ............................................ 141
      Time versus Frequency Domain Identification .............. 141
      Continuous Time Model Identification ....................... 141
   5.4 Nonlinear Models ............................................... 141
   5.5 Probabilistic Models .......................................... 141
   5.A Matrix Inversion Lemma ...................................... 141
   5.B Completing the Square ...................................... 141
      5.B.1 Expectation of a Normal PDF ......................... 143
   5.C Derivation of the Kalman Filter ............................ 143

6 Model Quality .......................................................... 147
   6.1 Variance Quantification ....................................... 147
      6.1.1 Fundamental Geometric Principles ....................... 147
      6.1.2 Fundamental Structural Results ......................... 147
      6.1.3 Variability of Estimated Frequency Response .......... 147
         A General Principle ....................................... 147
         Details for FIR Models .................................... 147
         Details for ARMAX Models ................................ 147
## CONTENTS

Details OE and BJ Models ........................................... 147

6.1.4 Variability of Nonlinear System Estimates .................. 147

6.1.5 Bootstrap Methods ........................................... 147

6.1.6 Exercises .................................................. 147

6.1.7 Computer exercises ......................................... 148

### 7 Experiment Design

7.1 Identifiability ................................................ 149

7.2 Informative Experiments ...................................... 149

7.3 Persistence of Excitation ..................................... 149

7.4 Input Signal Design .......................................... 149

7.4.1 Common Input Signals ..................................... 149

PRBS .......................................................... 149

Sums of Sine-Waves and Crest Factor Correction ............... 149

7.5 Application Oriented Experiment Design .................... 149

7.6 Adaptive Experiment Design ................................. 149

### 8 Model Validation

8.1 Residual whiteness Tests .................................... 151

8.2 Input to residual correlation tests .......................... 151

8.3 Model Error Modelling ....................................... 151

### 9 Applications

9.1 Closed Loop Identification .................................. 153

9.2 Network Models ............................................. 153

9.3 Errors-in-Variables Models .................................. 153

9.4 Block-structured Nonlinear Models ......................... 153
Chapter 1

Background Theory

1.1 Vector Spaces

There are other spaces of elements where the geometry of the Euclidean space holds. A vector space is a set \( V \) of elements \( v \), called vectors, on which two operations \( + \) and \( \cdot \), called vector addition and scalar multiplication, are defined such that for all vectors \( u, v, w \in V \) and scalars \( c, d \) it holds

1. Closure: \( u + v \in V \)
2. Commutativity: \( u + v = v + u \)
3. Associativity: \( (u + v) + w = u + (v + w) \)
4. Additive identity: \( V \) contains an element, denoted by \( 0 \), such that \( 0 + v = v \), \( \forall v \in V \)
5. Additive inverse: There exists a unique \( x(v) \in V \) such that \( v + x(v) = 0 \). \( x(v) \) is called \(-v\).
6. For any scalar \( c \), \( cv \in V \)
7. Distributivity: \( c \cdot (u + v) = c \cdot u + c \cdot v \)
8. Distributivity: \( (c + d) \cdot v = c \cdot v + d \cdot v \)
9. Associativity: \( c \cdot (d \cdot v) = (cd) \cdot v \)
10. Multiplicative identity: \( 1 \cdot v = v \)

If the above hold when the scalars belong to the field of reals, \( V \) is said to be a real vector space, and when \( c \in \mathbb{C} \), \( V \) is a complex vector space. We will consider complex vector spaces.

1.1.1 Inner Product Spaces

What gives the Euclidean space its geometry is the scalar product \( \circ \). For two vectors \( x \) and \( y \) we have that the angle between the two vectors can be determined from

\[
\cos(\alpha) = \frac{x \circ y}{\|x\| \|y\|}
\]

(1.1)

The corresponding operator in a complex vector space is the inner product \( \langle \cdot, \cdot \rangle : V \times V \to \mathbb{C} \), which, mimicking the properties of the scalar product, has to satisfy the following axioms for all \( u, v, w \in V \) and \( \lambda \in \mathbb{C} \):
CHAPTER 1. BACKGROUND THEORY

1. \( \langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle \)
2. \( \langle \lambda u, v \rangle = \lambda \langle u, v \rangle \)
3. \( \langle u, v \rangle = \langle v, u \rangle^* \)
4. \( \langle v, v \rangle \geq 0 \) with equality iff \( v = 0 \)

A vector space endowed with an inner product is called an inner product space. For such spaces we can introduce a topology via the norm

\[ \|v\| := \sqrt{\langle v, v \rangle} \]

It is easy to verify that this is a norm, i.e. that \( \| \cdot \| \) satisfies i) \( \|v\| \geq 0, \forall v \in V \) with equality iff \( v = 0 \), ii) \( \|\lambda v\| = |\lambda|\|v\| \), and iii) \( \|u + v\| \leq \|u\| + \|v\| \).

It can be noted, although we will not make use of this, that a normed vector space \( V \) is an inner product space iff the parallelogram law

\[ 2\|u\|^2 + 2\|v\|^2 = \|u + v\|^2 + \|u - v\|^2 \]

holds for all \( u, v \in V \). If this law holds then the inner product is given by

\[ \langle x, y \rangle = \frac{1}{4} \left( \|u + v\|^2 - \|u - v\|^2 \right) \]

The geometry in an inner product space becomes clear if we for a vector \( u \in V \) define its orthogonal projection on another vector \( v \in V \) as \( u_{\parallel v} := \alpha v \) where \( \alpha \) satisfies the normal equation

\[ \langle u - \alpha v, v \rangle = 0 \] (1.2)

i.e.

\[ u_{\parallel v} = \frac{\langle u, v \rangle}{\langle v, v \rangle} v \]

Then

\[ 0 \leq \|u - u_{\parallel v}\|^2 = \|u\|^2 - \frac{\langle u, v \rangle^2}{\|v\|^2} \]

with equality iff \( u = \lambda v \) for some \( \lambda \in \mathbb{C} \), so that

\[ 0 \leq \left| \frac{\langle u, v \rangle}{\|u\|\|v\|} \right| \leq 1 \] (1.3)

with the upper inequality true iff \( u = \lambda v \) for some \( \lambda \in \mathbb{C} \). As in (1.1) we can interpret the number in the middle above as the cosine of the angle between \( u \) and \( v \). Since

\[ \langle u - u_{\parallel v}, v \rangle = 0 \] (1.4)

we say that \( u - u_{\parallel v} \) is orthogonal to \( v \) (written \( u - u_{\parallel v} \perp v \)) and from this and the decomposition \( u = (u - u_{\parallel v}) + u_{\parallel v} \) we obtain Pythagoras theorem

\[ \|u\|^2 = \|u_{\parallel v}\|^2 + \|u - u_{\parallel v}\|^2 \] (1.5)

The upper inequality in (1.3) is known as the Cauchy-Schwarz inequality. Using this, the geometric interpretation is completed by considering

\[ \|u - \lambda v\|^2 = \|u - u_{\parallel v} + u_{\parallel v} - \lambda v\|^2 = \|u - u_{\parallel v}\|^2 + \|u_{\parallel v} - \lambda v\|^2 \geq \|u - u_{\parallel v}\|^2 \] (1.6)

with equality only if \( \lambda v = u_{\parallel v} \), which shows that \( u_{\parallel v} \) is the vector in the direction of \( v \) that is closest to \( u \), i.e. the notion of an orthogonal projection in an inner product space is consistent with the same notion in the Euclidean space.
1.1. VECTOR SPACES

1.1.2 Subspaces and Orthogonal Projections

A subspace \( S \) to a vector space \( V \) is a subset of \( V \) that is closed under addition and scalar multiplication, i.e. if \( u, v \in S \) then \( \lambda_1 u + \lambda_2 v \in S \) for any scalars \( \lambda_1 \) and \( \lambda_2 \).

Starting from a set of vectors \( \{ v_\alpha \}_{\alpha \in A} \), we can generate a subspace by all finite linear combinations of these vectors. We denote such a subspace \( \text{Span}\{ \{ v_\alpha \}_{\alpha \in A} \} \).

A finite set of vectors \( \{ v_k \}_{k=1}^n \) is said to be linearly independent if the only solution to \( \sum_{k=1}^n \alpha_k v_k = 0 \) is \( \alpha_1 = \ldots = \alpha_n = 0 \). More generally, a set \( M \) is said to be linearly independent if every finite collection of vectors from \( M \) is linearly independent.

A vector space \( V \) is said to be finite dimensional if there is an \( n < \infty \), the dimension of \( V \), such that \( V \) contains a linearly independent set of \( n \) vectors, whereas all sets of \( n + 1 \) vectors are linearly dependent.

A basis for a vector space is a linearly independent set such that all vectors in the space can be uniquely represented as a finite linear combination of elements in the set, the basis elements. A basis exists for every vector space but it is not unique.

For an inner product space \( V \) with an \( n \)-dimensional subspace \( S \) having basis \( \{ v_1, \ldots, v_n \} \), we can for a vector \( u \) define its orthogonal projection on \( S \) as \( u \parallel S := \sum_{k=1}^n \alpha_k v_k \) where, similarly to (1.2), the \( \{ \alpha_k \} \) are defined by the normal equations

\[
\langle u - \sum_{k=1}^n \alpha_k v_k, v_l \rangle = 0, \quad l = 1, \ldots, n
\]

which in matrix form becomes

\[
\begin{bmatrix}
\langle v_1, v_1 \rangle & \cdots & \langle v_1, v_n \rangle \\
\vdots & \ddots & \vdots \\
\langle v_n, v_1 \rangle & \cdots & \langle v_n, v_n \rangle
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\vdots \\
\alpha_n
\end{bmatrix}
= 
\begin{bmatrix}
\langle u, v_1 \rangle \\
\vdots \\
\langle u, v_n \rangle
\end{bmatrix}
\]  

(1.8)

If we for \( v = [v_1, \ldots, v_n]^T \) and \( x = [x_1, \ldots, x_m]^T \) define

\[
\langle v, x^T \rangle = 
\begin{bmatrix}
\langle v_1, x_1 \rangle & \cdots & \langle v_1, x_m \rangle \\
\vdots & \ddots & \vdots \\
\langle v_n, x_1 \rangle & \cdots & \langle v_n, x_m \rangle
\end{bmatrix}
\]

we can write (1.8) in compact form as

\[
\langle v, v^T \rangle \alpha = \langle v, x \rangle
\]

(1.9)

where \( \alpha = [\alpha_1 \ldots \alpha_n]^T \).

Considering an arbitrary point \( w \) in \( S \), we obtain the same inequality as in (1.6), i.e. \( u \parallel S \) is the unique point closest to \( u \) in \( S \). We call this the orthogonal projection of \( u \) on \( S \). We can also define the orthogonal complement \( S^\perp \) of \( S \) to be the set of all vectors in \( V \) orthogonal to all vectors in \( S \), i.e.

\[
S^\perp = \{ u \in V : \langle u, v_k \rangle = 0, \ k = 1, \ldots, n \}
\]

Clearly \( S^\perp \) is a subspace. Furthermore, we can uniquely decompose any vector \( u \in V \) into \( u = u \parallel S + u \perp S \), where \( u \parallel S \) is the orthogonal projection on \( S \) and where \( u \perp S := u - u \parallel S \). We write \( V = S \oplus S^\perp \).
1.1.3 Hilbert Spaces

It is easy to see that the limit point of convergent sequences in finite dimensional subspaces to normed spaces also belong to the subspace in question. Topologically, finite dimensional subspaces are always closed. A set $S$ in a normed space is said to be open if there to every point $v \in S$ exists a neighborhood $\{ u : \|u - v\| < \varepsilon \} \subset S, \varepsilon > 0$. A set is closed if its complement is open. Matters become somewhat more complicated when considering subspaces of infinite dimensions.

**Example 1.1.1.** Let $V = C[0, 1]$, the space of continuous function on the interval $[0, 1]$. Clearly this is a vector space under standard definitions of addition and scalar multiplication. We take the norm to be $\|v\| = \max_{0 \leq x \leq 1} |v(x)|$ (this is not an inner product space).

Now, let $S$ be the subspace to $V$ consisting of all polynomials. Then the sequence of monomials $v_k : v_k(x) := x^k, k = 1, 2, \ldots$ converges to the discontinuous function

$$v^*(x) := \begin{cases} 
0, & 0 \leq x < 1 \\
1, & x = 1
\end{cases}$$

in the used norm, i.e. to a function not even belonging to $V$, and even less to $S$.

**Example 1.1.2.** Let the vector space $V$ consist of the real numbers over the field of rationals, i.e. the scalars we use are rational, equipped with the inner product $\langle u, v \rangle = uv$. Now consider the subset $S$ consisting of rational numbers. Clearly, $S$ is a subspace over the field of rationals. This subspace is not closed as there are rational sequences that converge to irrational numbers (this is in fact a way to extend rational numbers to reals).

When working with infinite dimensional subspaces we must therefore typically require that the subspace is closed as otherwise orthogonal projections in the spirit above may not even belong to the subspace onto which we project.

Apart from the geometrical properties discussed in the preceeding section, the Euclidean space possess another desirable property namely that convergence of a sequence $\{ x_k \}$ is equivalent to that the sequence is a Cauchy sequence, i.e. for every $\varepsilon > 0$ there exists an $N$ such that $\|x_k - x_l\| < \varepsilon$ when $k, l > N$. This holds also for subspaces to an Euclidean space. A metric space with the property that every Cauchy sequence converges to an element in the space is said to be complete. A Hilbert space is a complete inner product space. Perhaps not surprising, there is a strong connection between closedness and completeness: In a Hilbert space a subspace is closed iff it is complete. Furthermore, as in the Euclidean space, a finite dimensional subspace is complete.

Above we have seen that in a finite dimensional subspace $S$ to an inner product space $V$ there is a vector in the subspace that is closest to a given vector $u \in V$, c.f. with the Euclidean space. In a Hilbert space this generalizes to infinite dimensional subspaces:

**Theorem 1.1.1.** Let $S$ be a closed subspace to a Hilbert space $\mathcal{H}$ and let $u \in \mathcal{H}$ be given. Then there is a unique vector $v \in S$ such that $u - v \perp w$ for all $w \in S$. The vector $v$ solves

$$\min_{w \in S} \|u - v\|$$

For Hilbert spaces we can thus talk about the orthogonal projection on $S$ even when $S$ is infinite dimensional and any vector $u \in V$ can uniquely be split into $u = u_{\|S\|} + u_{\perp S}$ where $u_{\|S\|} \in S$ and where $u_{\perp S} \in S^\perp$, the orthogonal complement to $S$ defined as $S^\perp = \{ v : v \perp S \}$. There is dual formulation to the problem in Theorem 1.1.1 which has $u_{\perp S}$ as solution.
Corollary 1.1.1. Let $S$ be a closed subspace to a Hilbert space $H$ and let $u \in H$ be given. Consider the linear variety

$$L_u = \{ x = u + v, \ v \in S \}$$

Then the problem

$$\min_{x \in L_u} \| x \|$$

has a unique solution $u \perp S$. The solution is the unique $x \in L_u$ satisfying

$$\langle x, v \rangle = 0 \ \forall v \in S$$

1.1.4 Orthonormal bases

A subset $M$ in an inner product space is said to be an orthonormal set if all vectors in $M$ have norm 1 and for any pair $u, v \in M$, $\langle u, v \rangle = 0$ when $u \neq v$.

Given an orthonormal sequence $\{e_k\}_{k=1}^{\infty}$ in a Hilbert space $H$, we can take $S_k$ to be the span of $\{e_k\}_{k=1}^{n}$. Then the orthonormality gives

$$\|u\|_{S_k}^2 = \sum_{k=1}^{n} |\langle u, e_k \rangle|^2$$

and from Pythagoras theorem (1.5) it follows that

$$\|u\|_{S_k}^2 = \sum_{k=1}^{n} |\langle u, e_k \rangle|^2 \leq \|u\|^2, \ k = 1, 2, \ldots$$

Since this holds for every finite $k$, Bessel’s inequality

$$\sum_{k=1}^{\infty} |\langle u, e_k \rangle|^2 \leq \|u\|^2 \quad (1.10)$$

follows. A remarkable consequence of this inequality is that if one has an uncountable orthonormal set $M$ in an inner product space $V$, then for a given $u \in V$, at most a countable set of Fourier coefficients $\langle u, e \rangle, e \in M$ can be non-zero. In this case we can thus still associate $u$ to a series

$$\sum_{k=1}^{\infty} \langle u, e_k \rangle e_k$$

where $\{e_k\}$ is an enumeration of the elements in $M$ that have non-zero inner products with $u$.

Now, suppose that we have a series $\sum_{k=1}^{\infty} \alpha_k e_k$ in a Hilbert space $H$. Then this series is convergent, to $v$ say, iff $\sum_{k=1}^{\infty} |\alpha_k|^2$ is convergent. This follows since $\sum_{k=1}^{\infty} |\alpha_k|^2$ then is a Cauchy sequence.

Furthermore, in this case $\alpha_k = \langle v, e_k \rangle$.

In particular it holds that for any $u \in H$, $\sum_{k=1}^{\infty} \langle u, e_k \rangle e_k$ is convergent. However, the series may not correspond to $u$ despite that, as per the preceding paragraph, the Fourier coefficients $\langle u, e_k \rangle$ are the same as for $u$. Denoting the series by $v$, it is easy to see that $u - v$ is orthogonal to every $e_k$. Thus for $u - v$ to be non-zero must mean that the orthonormal sequence does not span the whole of $H$. 
This is guaranteed by requiring the orthonormal sequence \( \{e_k\}_{k=1}^{\infty} \) to be what is called a complete (or total) orthonormal basis in \( \mathcal{H} \). This means that the span of the sequence is dense in \( \mathcal{H} \), i.e. the closure of the span is \( \mathcal{H} \) itself.

For a complete orthonormal basis equality holds in (1.10) which is then known as Parseval’s relation which can be seen as a generalization of Pythagoras theorem. Conversely, if Parseval’s relation holds for every \( u \in \mathcal{H} \), then the orthonormal set is complete.

So for which Hilbert spaces does there exist a countable complete orthogonal set? It turns out that the space must be separable which means that it has a countable subset which is dense and for such spaces every orthonormal set is countable.

1.2 Probability Theory

In mathematical terms probability theory is concerned with functions that map sets to real numbers between 0 and 1. We start with a set \( \Omega \), called the sample space, whose elements \( \omega \), the sample points, can be thought of representing all possible outcomes that can occur. An event is simply a subset of \( \Omega \).

A probability measure \( P \) assigns a probability to every event, i.e. a number between 0 and 1. It has to hold that \( P(\Omega) = 1 \) and for a disjoint family of sets \( \{A_k\}_{k=1}^{\infty} \), that \( P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k) \).

We would of course like to assign a probability to all possible events, i.e. to all possible subsets of \( \Omega \). However, the sets of sets is humongous and leads to inconsistencies.

Example 1.2.1 (Banach and Tarski). Let \( \Omega \) be the unit sphere \( S^2 \) in \( \mathbb{R}^3 \) and define the probability measure for a set \( F \) as the area of \( F \) divided with the area of \( S^2 \). Then it can be shown that there exists an \( F \in S^2 \) and disjoint rotations \( \{F_{i,k}\}_{i=1}^{k} \) of \( F \) such that \( P(F_{i,j}) = P(F) \) and \( S^2 = \bigcup_{i=1}^{k} F_{i,k} \) for \( k \geq 3 \). However, since the sets are disjoint this implies that

\[
1 = P(S^2) = P(\bigcup_{i=1}^{k} F_{i,k}) = \sum_{i=1}^{k} P(F_{i,k}) = kP(F), \quad k = 3, 4, \ldots
\]

i.e. it would appear that the area of \( F \) is not unique.

Thus, if we want to have probabilities defined in a meaningful way the family of allowed sets must be delimited. We will denote such a family of sets by \( \mathcal{F} \). Natural requirements are that i) \( \Omega \in \mathcal{F} \), ii) if \( A \in \mathcal{F} \) then its complement \( A^c \in \mathcal{F} \), and iii) the union of two sets belonging to \( \mathcal{F} \) should be in \( \mathcal{F} \) as well. However, these requirements are not enough to be able to answer pertinent questions in estimation theory. Let us jump ahead and study a typical estimation problem.

Example 1.2.2. Let \( \hat{\theta}_N \) be an estimator of a scalar quantity \( \theta \) based on the random vector \( Y_N \in \mathbb{R}^N \) and let us assume that we would like to examine the properties of \( \hat{\theta}_N \) when \( N \) becomes large. For example we might be interested in the probability that \( \hat{\theta}_N \) eventually remains within a distance \( \varepsilon > 0 \) from \( \theta \). This event can be expressed as

\[
F = \{ \omega : \limsup_{N \to \infty} |\hat{\theta}_N - \theta| \leq \varepsilon \} = \{ \omega : |\hat{\theta}_N - \theta| \leq \varepsilon \text{ for } N \text{ sufficiently large} \}
\]

Defining \( F_N = \{ \omega : |\hat{\theta}_N - \theta| \leq \varepsilon \} \) we can write

\[
F = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} F_N
\]
For sets of the type $F$ in the example to belong to $\mathcal{F}$ given that the $F_N \in \mathcal{F}$, it turns out that we have to require

$$iv) \quad F_k \in \mathcal{F}, \quad k = 1, 2, \ldots \Rightarrow \bigcup_{k=1}^{\infty} F_k \in \mathcal{F}$$

A family $\mathcal{F}$ of sets satisfying i)--iv) is called a $\sigma$-algebra, the pair $(\Omega, \mathcal{F})$ a measurable space and the triplet $(\Omega, \mathcal{F}, \mathcal{P})$ a probability space.

Starting from a family $\mathcal{C}$ of subsets on $\Omega$, $\sigma(\mathcal{C})$ is the smallest $\sigma$-algebra that contains all sets in $\mathcal{C}$. An example is the Borel $\sigma$-algebra $\mathcal{B}$ which is the smallest $\sigma$-algebra containing the open sets on the real axis. The sets in this $\sigma$-algebra are called Borel sets. The concept of Borel algebra extends to $\mathbb{R}^N$, $N < \infty$.

### 1.2.1 Information Contents in Events, Bayes Rule and Independence

Consider two events $A$ and $B$. When the sample space is restricted to $B$, i.e. whenever an outcome occurs that is not in $A$ it is discarded, the probability of the event $B$ changes. The probability that both $A$ and $B$ occurs is $\mathcal{P}(A \cap B)$ and the probability that we will have $A$ is $\mathcal{P}(A)$ which gives then that the probability of $B$ occurring conditional on that $A$ occurs is

$$\mathcal{P}(B|A) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(A)} \quad (1.11)$$

provided $\mathcal{P}(B) > 0$.

When the probability of $B$ does not change if we are given that $A$ has occurred, i.e.

$$\mathcal{P}(B|A) = \mathcal{P}(B),$$

we say that the events are independent.

We can express the conditional $\mathcal{P}(B|A)$ using $\mathcal{P}(A|B)$

$$\mathcal{P}(B|A) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(A)} = \frac{\mathcal{P}(A|B)\mathcal{P}(B)}{\mathcal{P}(A)}$$

This is Bayes rule from which we see that if $A$ and $B$ are independent (according to the definition above)

$$\mathcal{P}(A|B) = \mathcal{P}(A)$$

From (1.11) we have see that see that when $\mathcal{P}(B) > 0$, independence is equivalent to

$$\mathcal{P}(A \cap B) = \mathcal{P}(A)\mathcal{P}(B|A) = \mathcal{P}(A)\mathcal{P}(B)$$

This relation relation is typically taken as the definition of independence.

### 1.2.2 Random Variables

Given a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, a real valued random variable is a function from the sample space to the real axis: $\Omega \rightarrow \mathbb{R}$. A typical event for a random variable $Y$ that we might be interested in is

$$\{\omega : Y(\omega) < c\} \quad (1.12)$$
for some constant $c$ (this leads to the distribution function). For us to be able to assign probabilities to such events, they have to be measurable, i.e. they have to belong to the measurable sets $\mathcal{F}$. Real valued functions $f$ from the sample space for which sets of the type $\{\omega : f(\omega) \in B\}$, where $B$ is any Borel set, are measurable are called measurable functions. Sets of the type (1.12) are of this type and by requiring random variables to be measurable functions, the probability for the event (1.12) is well defined.

A random variable $X$ is characterized by its probability distribution function $P_X(B) := P(X \in B)$ for all $B \in \mathcal{B}$, where $X \in B = \{\omega : X(\omega) \in B\}$. We can see $P_X(B)$ as a probability measure on the measurable space $(\mathbb{R}, \mathcal{B})$. The distribution function is defined as $F_X(x) := P_X((-\infty, x))$. If $P_X$ is absolutely continuous with respect to the Lebesgue measure, there is a measurable function $p_X : \mathbb{R} \to \mathbb{R}$ called the probability density function such that

$$P_X(B) = \int_B p_X(x) \, dx$$

Here absolutely continuous means that

$$\int_B dx = 0 \implies P_X(B) = 0$$

For example, there are no discrete events $X = x$ that occur with positive probability.

### 1.2.3 Expectation

A random variable $X$, defined on a probability space $(\Omega, \mathcal{F}, B)$, has expectation

$$\mathbb{E}[X] = \int_{\Omega} X \, dP := \int_{\Omega} X(\omega) \, P(d\omega)$$

when the integral is defined. We will not give a precise meaning of $P(d\omega)$ and it would take us too far into measure theory to do so. However, we can also express the expectation in terms of the probability distribution function $P_X$, or the distribution function

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x P_X(dx) = \int_{-\infty}^{\infty} x F_X(dx)$$

When the sample space is discrete it follows that

$$\mathbb{E}[X] = \sum_k x_k P_X(x_k)$$

and when the pdf exists

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x) \, dx$$

### 1.2.4 Information Contents in Random Variables and Conditional Probabilities

We will be interested in the information contained in an observation of a random variable $Y$, i.e. what information does the outcome $Y(\omega)$ carry? It is obvious that we can determine if the event (1.12) has occurred or not and in fact we can determine if events of the type $\{\omega : Y(\omega) \in B\}$, where $B$ is a
1.2. PROBABILITY THEORY

Borel set, has occurred or not. It turns out that we can determine if events belonging to the \( \sigma \)-algebra generated by \( Y \),

\[
\sigma(Y) := \sigma(\{\omega : Y(\omega) \in B, \ B \in B\}),
\]

have occurred or not. \( \sigma(Y) \) thus represents the events whose occurrence we can determine by observing \( Y \). \( \sigma(Y) \) is a sub \( \sigma \)-algebra to \( \mathcal{F} \), i.e. if \( A \in \sigma(Y) \), then \( A \in \mathcal{F} \). This means that there are typically events in \( \mathcal{F} \) which we cannot determined if they have occurred or not by observing \( Y \). For example, suppose that there is another real valued random variable \( X \). From the knowledge of \( Y \) we then only get partial information about \( X \), i.e. we typically cannot determine the occurrence of all events in \( \sigma(X) \). So what can be said in such a case? Knowing that \( Y \in B_Y \) changes the probability that \( X \in B_X \) to the conditional probability distribution function

\[
P_{X|Y}(B_X|B_Y) := \frac{P(X \in B_X \cap Y \in B_Y)}{P(Y \in B_Y)}
\] (1.13)

For each fix \( B_Y \) such that the denominator is positive, this is a probability measure over \((\Omega, \sigma(X))\). In the context of estimation we will call this the posterior distribution function of \( X \).

For example

\[
P_{X|Y}(B_X|y) = \frac{P(X \in B_X \cap Y = y)}{P(Y = y)}
\] (1.14)

would give the probability of different events related for \( X \) when \( Y = y \) is observed.

**Example 1.2.3.** Suppose that \( X \) is Bernouilli distributed with equal probabilities for taking the values 0 and 1 and suppose that \( Y \) takes the values \(-1, 0, 2\), and that the joint distribution of \( X, Y \) is given by

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( P(X \in B, Y = y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>1/8</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>2/8</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>3/8</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2/8</td>
</tr>
</tbody>
</table>

This gives

| \( x \) | \( y \) | \( P_{X|Y}(X = x | Y = y) \) |
|-------|-------|-----------------------------|
| -1    | -1    | \( \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2}} = \frac{1}{3} \) |
| 1     | -1    | \( \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2}} = \frac{1}{3} \) |
| -1    | 0     | \( \frac{\frac{1}{2} \cdot \frac{1}{4}}{\frac{1}{4}} = 1 \) |
| 1     | 2     | \( \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2}} = 1 \) |

However, there is a technical problem with (1.14), namely that it is not well defined when \( P(Y = y) = 0 \). For full generality we therefore need a more abstract definition of the conditional probability distribution function. What we would like is that \( P_{X|Y}(B_X|y) \) is a function such that when we
integrate it with respect to \( y \) over a region \( B_Y \), taking the probability measure for \( Y \) into account, we obtain
\[
\mathcal{P}(X \in B_X \cap Y \in B_Y) = \int_{Y \in B_Y} \mathcal{P}_{X|Y}(B_X|Y(\omega)) \mathcal{P}(d\omega) = \mathcal{P}(X \in B_X \cap Y \in B_Y) \tag{1.15}
\]
As noted above, in addition to (1.15) we also need that \( \mathcal{P}_{X|Y}(B_X|B_Y) \) is a probability measure on \( \sigma(X) \) for any fix \( B_Y \). While it can be shown that there exists a function satisfying these requirements it is beyond this exposition to prove this.

The expectation of the conditional distribution for \( X \) given \( Y = y \) is given by
\[
\mathbb{E} [X|Y] (\omega) = \int_{\Omega} X(\omega) \mathcal{P}_{X|Y}(X(\omega)|Y(\omega)) d\omega
\]
We can also define the conditional expectation of \( X \) given \( Y \) as
\[
\mathbb{E} [X|Y] (\omega) = \int_{\Omega} X(\omega) \mathcal{P}_{X|Y}(X(\omega)|Y(\omega)) d\omega \tag{1.16}
\]
This is a random variable which on the event \( Y = y \) has as outcome the mean of the conditional distribution for this observation of \( Y \).

This random variable has the same expectation as \( X \):
\[
\mathbb{E} [\mathbb{E} [X|Y]] = \int_{\Omega} \mathbb{E} [X|Y] (\omega) \mathcal{P}(d\omega) = \int_{\Omega} \int \mathcal{P}_{X|Y}(X(\omega)|Y(\omega)) \mathcal{P}(d\omega)
\]
\[
= \int_{\Omega} X(\omega) \mathcal{P}_{X|Y}(X(\omega)|Y(\omega)) d\omega = \int_{\Omega} X(\omega) \mathcal{P}(d\omega) = \mathbb{E} [X]
\]

**Example 1.2.4** (Example 1.2.3 continued). The conditional mean is given by
\[
\begin{align*}
\mathbb{E} [X|Y = -1] &= -1 \times \mathcal{P}_{X|Y}(X = -1|Y = 1) + 1 \times \mathcal{P}_{X|Y}(X = 1|Y = 1) = -1 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{1}{3} \\
\mathbb{E} [X|Y = 0] &= -1 \times \mathcal{P}_{X|Y}(X = -1|Y = 0) + 1 \times \mathcal{P}_{X|Y}(X = 1|Y = 0) = -1 \times 1 + 0 = -1 \\
\mathbb{E} [X|Y = 2] &= -1 \times \mathcal{P}_{X|Y}(X = -1|Y = 2) + 1 \times \mathcal{P}_{X|Y}(X = 1|Y = 2) = 0 + 1 \times 1 = 1
\end{align*}
\]

**Example 1.2.5.** In the case where the probability distributions of \( X \) and \( Y \) can be represented by the joint probability distribution function (pdf) \( p(x, y) \), the conditional distribution function of \( X \) given \( Y \) can be expressed in terms of the conditional pdf
\[
p_{X|Y}(x|y) = \begin{cases} 
\frac{p(x, y)}{p_Y(y)} & p(y) > 0 \\
0 & p_Y(y) = 0
\end{cases}
\]
where \( p_Y(y) \) is the marginal pdf
\[
p_Y(y) = \int p_Y(x, y) dx
\]
The conditional expectation is
\[
\mathbb{E} [X|Y] = \int_{-\infty}^{\infty} xp_{X|Y}(x|Y) dx
\]
1.2.5 Independent Random Variables

So when does one random variable $Y$ not contain any information about another random variable $X$? It is when the probability of any event in $\sigma(X)$ does not change when we observe an event in $\sigma(Y)$, i.e. when

$$P_{X|Y}(B_X|B_Y) = \frac{P(X \in B_X \cap Y \in B_Y)}{P(Y \in B_Y)} = P(X \in B_X)$$

(1.17)

i.e. when

$$P(X \in B_X \cap Y \in B_Y) = P(X \in B_X)P(Y \in B_Y) \quad \forall B_X \in \sigma(X), B_Y \in \sigma(Y)$$

(1.18)

Since the sub-$\sigma$-algebras $\sigma(X)$ and $\sigma(Y)$ are generated by sets of the type $\{\omega : X(\omega) < c\}$ it is sufficient to prove that (1.17) holds for all sets of this type. In this case the conditional probability is given by

$$P_{X|Y}(B_X|y) = P(X \in B_X)$$

We conclude that $Y$ does not carry any information about $X$. As a consequence the conditional expectation is

$$E[X|Y](\omega) = \int_{\Omega} X(\bar{\omega})dP_{X|Y}(\bar{\omega}|Y(\omega)) = \int_{\Omega} X(\bar{\omega})dP(\bar{\omega}) = E[X] \quad \text{a.e.}$$

1.3 References

Section 1.1

Chapters 2-3 in [9].

Section 1.2

Chapters 1-3 in [2].
Chapter 2

Introduction to Estimation

2.1 Introduction

This chapter introduces the problem of estimating the probability measure that governs an observed random variable $Y \in \mathbb{R}^N$. We will assume that $Y$ has a pdf and that it has the form $p(y; \theta)$ where $\theta \in \mathbb{R}^n$ is an unknown parameter vector. The problem is thus to estimate $\theta$. A function $\hat{\theta} : \mathbb{R}^N \to \mathbb{R}^n$ will be called an estimator of $\theta$.

Naturally, $\hat{\theta} = \hat{\theta}(Y)$, i.e. a function of $Y$, making the estimator a random variable. A quality measure for an estimator is called a risk function and here we will use the MSE

$$\text{MSE} \left[ \hat{\theta}(Y) \right] := \mathbb{E} \left[ \| \hat{\theta}(Y) - \theta \|^2 \right]$$

as risk function but one could also use other functions. Notice that the MSE is a matrix when $n > 1$ in which case the trace of the quantity above is what one would normally call the MSE. From the above we see that parameter estimation is about estimating a constant with a random variable. Thus we will start this chapter with general estimation theory on how to estimate one random variable from another.

2.2 Minimum Mean Square Error Estimation of Random Variables

2.2.1 Introduction

Let $X$ be a random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that we would like to estimate $X$, i.e. provide a guess $\hat{x}$. One possible quality measure is the mean square error (MSE)

$$\text{MSE} [\hat{x}] = \mathbb{E} \left[ \| X - \hat{x} \|^2 \right]$$

The MSE can be split in two terms

$$\text{MSE} [\hat{x}] = \mathbb{E} \left[ \| X - \mathbb{E} [X] \|^2 \right] + \| \mathbb{E} [X] - \hat{x} \|^2$$

(2.1)

The first term is called the variance error (it’s the variance of $X$) and the second term the bias (systematic) error. We see that here the variance error is something we cannot influence while we can eliminate the bias error by taking $\hat{x} = \hat{x}^* := \mathbb{E} [X]$. This gives the minimum mean square error (MMSE)

$$\text{MMSE} := \text{MSE} [\hat{x}^*] = \mathbb{E} \left[ \| X - \mathbb{E} [X] \|^2 \right] = \mathbb{E} [X^2] - \mathbb{E}^2 [X]$$
Suppose now that we are given the information that another random vector \( Y \), defined on the same probability space, has taken the value \( y \). Can we then improve our estimate of \( X \)? Well, now the probability distribution for \( X \) has changed to the conditional distribution \( P_{X|Y}(X \in B_X | Y = y) \) so the MSE is now given by

\[
\text{MSE} \left[ \hat{x} | Y = y \right] = E \left[ |X - \hat{x}|^2 | Y = y \right]
\]

However, we can expand this expression into a variance term and a bias term just as in (2.1)

\[
\text{MSE} \left[ \hat{x} | Y = y \right] = E \left[ |X - E[X|Y = y]|^2 | Y = y \right] + |E[X|Y = y] - \hat{x}|^2
\]

(2.2)

which again is minimized by taking \( \hat{x} \) to be the mean of \( X \), but now the conditional mean \( E[X|Y = y] \).

This means that the optimal \( \hat{x} \) now will depend on the observation \( y \): \( \hat{x} = \hat{x}^*(y) = E[X|Y = y] \).

The minimum MSE is now a function of \( y \) and given by

\[
\text{MMSE}(y) = \text{MSE} \left[ \hat{x}^*(y) | Y = y \right] = E \left[ |X - \hat{x}^*(y)|^2 | Y = y \right] = E[X^2|Y = y] - E^2[X|Y = y]
\]

(2.3)

We would expect that our observation should improve the quality of the estimate, i.e. that

\[
\text{MMSE}(y) < \text{MMSE}(E[X])
\]

but is this true, and, if so, how much better is \( \hat{x}(y) \) than \( \hat{x} \)? To examine this we re-write the MSE as

\[
\text{MSE} \left[ \hat{x} \right] = E \left[ |X - \hat{x}|^2 \right] = E_Y \left[ E \left[ |X - \hat{x}|^2 | Y \right] \right] = \int E \left[ |X - \hat{x}|^2 | Y = y \right] p_Y(y)dy
\]

(2.4)

\[
= \int \text{MSE} \left[ \hat{x} | Y = y \right] p_Y(y)dy
\]

(2.5)

Thus, while \( E[X] \) minimizes this integral expression, \( E[X|Y = y] \) minimizes the integrand at the point \( y \) so clearly

\[
\text{MSE} \left[ E[X|Y = y] | Y = y \right] \leq \text{MSE} \left[ E[X] | Y = y \right]
\]

Furthermore, we observe that when \( X \) and \( Y \) are independent we have equality since conditioning on \( Y \) then does not change the distribution of \( X \).

We now pose the question how to estimate \( X \) from an arbitrary observation of \( Y \). We will then allow \( \hat{x} \) to be a function of \( Y \), \( \hat{x} = \hat{x}(Y) \). We notice that the decomposition (2.4) is still valid in this case and that taking \( \hat{x}(Y) = \hat{x}^*(Y) = E[X|Y] \) will for each \( Y = y \) minimize the integrand, and hence minimize the MSE. Taking the expectation of (2.3) gives

\[
\text{MMSE} = \text{MSE} \left[ E[X|Y] \right] = E[X^2] - E\left[ E^2[X|Y] \right]
\]

(2.6)

2.2.2 The Internal Structure of the Conditional Expectation

Above, we have seen that the conditional expectation of \( X \) given the observed variable \( Y \) gives the MMSE estimator. In this section we will try to understand a little bit better why this is the case.

In general terms estimating a random variable \( X \) from another random variable \( Y \) means how well we can mimic the behaviour of \( X \) using \( Y \). Now \( X \) is defined by its probability distribution over arbitrary sets in \( \mathcal{F} \). In fact suppose that there is another random variable \( Z \) such that

\[
\int_A Z(\omega) = \int_A X(\omega)P(d\omega) \quad \forall A \in \sigma(X)
\]

(2.7)
Then it follows that $Z = X$ except possibly on a set of measure zero. To see this suppose that $Z \neq X$ on a set $A \in \sigma(X)$ for which $P(A) > 0$. Then we can take the subset on $A$ either for which $Z > X$ or $Z < X$, whichever has non-zero measure. This subset also belong to $\sigma(X)$ and clearly the two integrals above cannot be the same then. Hence we can take the set of pairs \( \{\{A \in \sigma(A), \int_A X(\omega)P(\omega)\}\} \) as definition of a random variable.

So a natural question is if we can use our observed variable $Y$ to construct a new random variable $Z = Z(Y)$ which has a probability distribution as close as possible to the above. Since $Z$ is a function of $Y$ we can only try to match (2.7) for events $A \in \sigma(Y)$, i.e. we can try to find $Z$ such that

$$\int_A Z((Y(\omega))P(d\omega) = \int_A X(\omega)P(d\omega) \quad \forall A \in \sigma(Y)$$

It is far from obvious that there is such a function and that, if so, it is a measurable function so that $Z$ is a random variable so let us study a couple of examples.

**Example 2.2.1** (Example 1.2.4 continued). In this example, the sets of $\sigma(Y)$ are $Y = -1$, $Y = 0$, $Y = 2$, and unions of these. For example

$$\int_{Y=-1} X(\omega)P(d\omega) = -1 \times P(X = -1, Y = -1) + 1 \times P(X = 1, Y = -1) = -1 \times \frac{1}{8} + 1 \times \frac{2}{8} = \frac{1}{8}$$

whereas

$$\int_{Y=-1} \mathbb{E}[X|Y](\omega)P(d\omega) = \mathbb{E}[X|Y = -1]P(Y = -1) = \frac{1}{3} \times \frac{3}{8} = \frac{1}{8}$$

showing that $\mathbb{E}[X|Y](\omega)$ has the same mean as $X$ over the event $Y = -1$. The same can be shown for the other sets of $\sigma(Y)$.

**Example 2.2.2.** Suppose that $X(\omega) = 1$ for $\omega \in A_X$ and zero otherwise. Similarly $Y(\omega) = 1$ for $\omega \in A_Y$ and zero otherwise. Let $B_X = \{X(\omega) : \omega \in A_X\}$ and $B_Y = \{Y(\omega) : \omega \in A_Y\}$.

Then $\sigma(Y) = \{0, A_Y, A_Y^c, \Omega\}$ and $Z$ has to be constant on $A_Y$ (and $A_Y^c$). Let us call this value $z$. In order to satisfy (2.8) we need

$$\int_{A_Y} ZdP = \int_{A_Y} XdP = P(A_X \cap A_Y)$$

but since $Z$ is constant, $Z = z$, on $A_Y$

$$\int_{A_Y} ZdP = zP(A_Y)$$

we have

$$z = \frac{P(A_X \cap A_Y)}{P(A_Y)} = \frac{P(X \in B_X \cap Y \in B_Y)}{P(Y \in B_Y)} = P_{X|Y}(B_X|B_Y)$$

Repeating these calculations for the event $B_Y^c$ gives that $Z$ should take the value $P(B_X|B_Y^c)$ on this set. We thus have

$$Z = \begin{cases} P_{X|Y}(B_X|B_Y) & Y \in B_Y \\ P_{X|Y}(B_X|B_Y^c) & Y \in B_Y^c \end{cases}$$
Consider now the conditional expectation (1.16)
\[ E[X|Y](\omega) = \int_\Omega X(\bar{\omega})P_{X|Y}(X(d\bar{\omega})|Y(\omega)) \]

For \( Y(\omega) \in B_Y \) this evaluates to \( P_{X|Y}(B_X|B_Y) \) and for \( Y(\omega) \in B_Y^c \) to \( P_{X|Y}(B_X|B_Y^c) \). Thus \( Z \) coincides with the conditional expectation \( E[X|Y] \).

The preceding two examples suggests that the conditional expectation \( E[X|Y] \) satisfies (2.8) and this is indeed true. We will not prove this in full generality but restrain ourselves to the simple case that \( X = I_{B_X} \), generalizing Example 2.2.2. Then
\[ \int_{Y \in B_Y} X(\omega)P(d\omega) = P(X \in B_X \cap Y \in B_Y) \]

while
\[ \int_{Y \in B_Y} E[X|Y](\omega)P(d\omega) = \int_{Y \in B_Y} \int_\Omega X(\bar{\omega})P_{X|Y}(d\bar{\omega}|Y(\omega))P(d\omega) = \int_{Y \in B_Y} P_{X|Y}(B_X|Y(\omega))P(d\omega) = P(X \in B_X \cap Y \in B_Y) \]

The general result is obtained by a constructive procedure where \( X \) is built up as a limit of indicator functions.

The condition (2.8) can be taken as a definition of the conditional expectation. From this we realize that the conditional expectation is not unique: If \( Z \) satisfies (2.8) we can change it on a subset of \( \Omega \) of measure zero, i.e. for which the probability measure \( P \) is zero, and still maintain (2.8). We say that the conditional expectation is uniquely defined a.e. (almost everywhere).

Suppose now that \( X = f(Y) \) where \( f \) is a measurable function. Then we see immediately that \( Z = X = f(Y) \) satisfies (2.8), i.e.
\[ E[f(Y)|Y] = f(Y) \]

More generally we see that when \( X = VW \) where \( V \in \sigma(Y) \) and \( W \) is another random variable,
\[ E[X|Y] = E[VW|Y] = VE[W|Y], \quad V \in \sigma(Y) \quad (2.9) \]

### 2.2.3 A Hilbert Space Interpretation

By embedding the estimation problem in a Hilbert space setting we will be able to handle a range of estimation problems.

Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \( \mathcal{H} \) be the vector space of random variables \( X \) defined on this probability space with
\[ E[X^2] < \infty \]

endowed with the inner product
\[ \langle X, Y \rangle = E[XY] \quad (2.10) \]
Then it can be shown that
\[ \lim_{n,m \to \infty} \|X_n - X_m\| = 0 \iff \exists X : \|X\| < \infty, \lim_{n \to \infty} \|X_n - X\| = 0 \]
i.e. the Cauchy criterion is equivalent to convergence. However, \( X \) is only unique a.e. Uniqueness is obtained by dividing \( \mathcal{H} \) into equivalence classes of random variables that are equal a.e. With this construct \( \mathcal{H} \) is a Hilbert space.

Consider now that we would like to use a random variable \( Y \) to estimate another random variable \( X \) in MMSE sense. We notice that
\[
\text{MSE}[g(Y)] = \|X - g(Y)\|^2
\]
To approach the problem of minimizing this quantity we form the subspace \( S \) consisting of all elements \( Z \) in \( \mathcal{H} \) for which the \( \sigma \)-algebra generated by the corresponding random variable, let us call it \( \sigma(Z) \), is a subset of \( \sigma(Y) \). This essentially means the subspace of all measurable functions of \( Y \) which have bounded second moment.

From the Hilbert space theory we then know that there is a unique element in \( S \) solving
\[
\min_{Z \in S} \|X - Z\| \quad (2.11)
\]
and that this element is uniquely determined by the orthogonality condition
\[
\langle X - Z, W \rangle = 0 \quad \forall W \in S \quad (2.12)
\]
in other words \( Z \) should be the orthogonal projection of \( X \) on \( S \). We observe that \( S \) is infinite dimensional so it does not seem trivial to find the projection. However, spurred by the results in Section 2.2.1 let us take \( \mathbb{E}[X|Y] \) as candidate. We have for any \( W \in S \)
\[
\langle X - \mathbb{E}[X|Y], W \rangle = \mathbb{E}[(X - \mathbb{E}[X|Y])W] = \mathbb{E}[(XW - W\mathbb{E}[X|Y])] = \mathbb{E}[WX] - \mathbb{E}[W\mathbb{E}[X|Y]] = \mathbb{E}[WX] - \mathbb{E}[WX] = 0 \quad (2.13)
\]
where the third equality follows from (2.9). We have thus proved that the conditional expectation is the orthogonal projection of \( X \) on \( S \)
\[
X_{||S} = \mathbb{E}[X|Y] \quad (2.14)
\]
A rather remarkable result considering that \( S \) is infinite dimensional.

We recognize (2.6) as Pythagoras relation
\[
\|X - X_{||S}\|^2 = \|X\|^2 - \|X_{||S}\|^2 \quad (2.15)
\]
We can extend the setting to the case where both \( X \) and \( Y \) are random vectors. Then \( \mathbb{E}[X|Y] \) is a vector of the conditional expectations \( \mathbb{E}[X_i|Y] \) and each element of \( X - \mathbb{E}[X|Y] \) is orthogonal to all measurable functions of \( Y \) that have bounded second moments. We can then extend Pythagoras relation (2.15) to a matrix equality
\[
\langle X - X_{||S}, (X - X_{||S})^T \rangle = \langle X, X \rangle - \langle X_{||S}, X_{||S} \rangle \quad (2.16)
\]
Using that \( \mathbb{E}[X|S] = \mathbb{E}[X] \) (recall that the conditional mean has the same mean as the random variable itself), we can re-write this as

\[
\text{Cov}[X - X|S] = \text{Cov}[X] - \text{Cov}[X|S]
\]

(2.17)

It follows that

\[
\text{Cov}\left[X - \hat{X}(Y)\right] \geq \text{Cov}[X - X|S]
\]

(2.18)

for any (measurable) estimator \( \hat{X}(Y) \) with equality iff \( \hat{X}(Y) = X|S = \mathbb{E}[X|Y] \) a.e.

### 2.2.4 Linear Estimators

It is not always easy to compute the conditional mean, and also the entire joint distribution of \( X, Y \) may not be known. It may therefore be attractive to project on other subspaces than \( \sigma(Y) \). The derivations in the preceding section still hold. However, (2.17) holds only if the subspace is such that the projection has the same mean as the estimated variable.

Consider estimating the random variable \( X \in \mathbb{R} \) from the random vector \( Y \in \mathbb{R}^N \). Sacrificing accuracy, an estimator simpler to compute requiring only second order moments is the linear estimator

\[
\hat{X} = LY
\]

To solve this case we let \( S_L \) be the subspace \( S_L = \text{Span}\{Y_1, \ldots, Y_n\} \). Since this is a finite dimensional subspace spanned by \( Y_1, \ldots, Y_n \) it is sufficient that the orthogonality condition (2.12) holds for \( W = Y_1, \ldots, Y_n \)

\[
\langle X - LY, Y_k \rangle = 0 \quad k = 1, \ldots, n
\]

(2.19)

which, using (1.9), can be written

\[
L\langle Y, Y^T \rangle = \langle X, Y^T \rangle
\]

(2.20)

i.e. the orthogonal projection is given by

\[
X_{\|S_L} = \langle X, Y^T \rangle (Y, Y^T)^{-1}Y
\]

(2.21)

We call \( X_{\|S_L} \) the optimal linear estimator (OLE) of \( X \) given \( Y \). Since \( S_L \subset S \) it follows that

\[
\|X - X_{\|S}\|^2 \leq \|X - X_{\|S_L}\|^2
\]

Unless the means of \( X \) and \( Y \) are zero, the OLE may be biased. A simple way to improve the estimator is then to extend \( Y \) with a constant element, we denote the resulting subspace \( S_{Le} \). This is equivalent to adding a constant term to the OLE. Adjusting this term such that the mean of the OLE is the same as the mean of \( X \) gives the smallest MSE. The simplest way to do this is to construct the OLE for \( X - \mathbb{E}[X] \) given \( Y - \mathbb{E}[Y] \) and then to add \( \mathbb{E}[X] \) to the estimator:

\[
X_{\|S_{Le}} = \langle X - \mathbb{E}[X], (Y - \mathbb{E}[Y])^T \rangle (Y - \mathbb{E}[Y], (Y - \mathbb{E}[Y])^T)^{-1}Y + \mathbb{E}[X]
\]

(2.22)

The OLE can be expressed as

\[
X_{\|S_{Le}} = \text{Cov}[X, Y] \text{Cov}^{-1}[Y] (Y - \mathbb{E}[Y]) + \mathbb{E}[X]
\]
2.2. **MINIMUM MEAN SQUARE ERROR ESTIMATION OF RANDOM VARIABLES**

which has covariance

\[ \text{Cov} \left[ X_{\|S_Le} \right] = \text{Cov} \left[ X, Y \right] \text{Cov}^{-1} \left[ Y \right] \text{Cov} \left[ Y, X \right] \]

and Pythagoras relation reads like (2.17).

We can also estimate a random vector \( X = [X_1 \ldots X_m]^T \) using linear combinations of \( Y \). With now \( L \) being a matrix, each element of \( X \) can be estimated as above rendering exactly the same expression as above for \( X_{\|S_Le} \) and its covariance, and the matrix expression (2.17) for Pythagoras relation remain intact.

We also remark that whenever we encounter an expression that can be written on the form

\[ \langle x, y \rangle \langle y, y \rangle^{-1} \langle y, x \rangle \]  \hspace{1cm} (2.22)

for vectors \( x \) and \( y \) whose elements are in a Hilbert space, then we can interpret this expression as the norm of the projection of \( x \) on the span of \( y \).

### 2.2.5 Maximum A Posteriori Estimation

The Maximum A Posteriori (MAP) estimate given \( Y = y \) is defined as the mode of the posterior density

\[ \hat{X}_{MAP} = \text{arg max} \ p_X|Y(x|y) \]

### 2.2.6 Other Estimation Criteria*

A generalization to the MSE is to include a weighting function \( W \) which may depend on \( X \), the variable we want to estimate

\[ \mathbb{E} \left[ W(X) || X - \hat{X}(Y) ||^2 \right] \]

The optimal estimator for this criterion is

\[ \hat{X}(Y) = \frac{\mathbb{E} \left[ W(X) X | Y \right]}{\mathbb{E} \left[ W(X) | Y \right]} \]

while the criterion

\[ \mathbb{E} \left[ W(X) || X - \hat{X}(Y) || \right] \]

leads to the optimal estimator being the median of the conditional distribution.

### 2.2.7 Exercises

2.2.1. Let \( X \in \mathbb{R}^m \) and \( Y \in \mathbb{R}^n \) be two random vectors that are jointly Gaussian:

\[
\begin{bmatrix}
X \\
Y
\end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix}
m_X \\
m_Y
\end{bmatrix}, \begin{bmatrix}
\Sigma_{XX} & \Sigma_{XY} \\
\Sigma_{YX} & \Sigma_{YY}
\end{bmatrix} \right)
\]

Derive the conditional probability density function for \( X \) given \( Y \). What is the conditional mean and the conditional covariance matrix?

Compare with the OLE. What conclusions can you draw?
2.2.2. A typical situation is that the distribution of the observed variable $Y$ is known when the variable $X$ to be estimated is given, i.e. $p(y|x)$. Suppose that $X$ is Bernoulli distributed with probability $p$ that $X = 0$ and suppose that $Y|X = 0$ is $N(2, 1)$ and $Y|X = 1$ is $\chi^2(3)$. What is the conditional distribution of $X$ given $Y$?

2.2.3. The Monty Hall problem. Suppose that you are in a game show where a car is hidden behind one of three closed doors. Initially you choose one of the doors and then the game host, which knows where the car is hidden, opens one of the other doors which is empty. You are now given the option of keeping the door that you selected in the first place or to change to the other closed door. Compute the posterior probabilities for which door the car hides behind, given your initial choice and the game host’s choice. What is the posterior mean? Which door has the maximum a posteriori probability? What is the optimal strategy and what are the winning chances?

This problem caused a big media ruckus in 1990 with an ensuing torrent of mails from the public in regards to the correct solution. Interestingly, 62% of the answers coming from PhDs were incorrect. Google after you have solved the problem!

2.3 Unbiased Parameter Estimation

2.3.1 Introduction

In this section we will study estimators that are unbiased. That is

$$
\mathbb{E} \left[ \hat{\theta}(Y) \right] = \theta
$$

in which case the MSE becomes the covariance matrix of $\hat{\theta}$. We are of course interested in minimizing this quantity. We will take the approach that we are given an arbitrary unbiased estimator $\hat{\theta}(Y)$ and then see if we can improve its accuracy in some way. We know from Pythagoras relation (2.17) that the orthogonal projection $X \| S$ of $X := \hat{\theta}(Y)$ on a subspace $S$ generated by some random variables, including the constant one, will also be unbiased and have a smaller covariance matrix than $X$ (unless the estimator already belongs to the subspace in question). $X \| S$ will thus give a new unbiased estimator which has smaller MSE (in matrix sense) than the original $\hat{\theta}$. However, if $X \| S$ depends on $\theta$ this is not an estimator so some care has to be exercised when determining a suitable subspace to project on.

Before continuing we remark that unbiased estimators do not always exist.

**Example 2.3.1** (Example 1.2 in [10]). Let $Y_i, i = 1, \ldots, N$ be binomially distributed, with probability $1/\theta$ of taking the value 0 (and $1 - 1/\theta$ of taking the value 1). Let $n = \sum_{i=0}^N Y_i$. Suppose that $\hat{\theta}$ is unbiased. Then

$$
\sum_{n=0}^N \hat{\theta}(n) \binom{N}{n} \left( \frac{1}{\theta} \right)^n \left( 1 - \frac{1}{\theta} \right)^{N-n} = \theta, \quad \forall 1 < \theta < \infty
$$

Consider now that $\theta \to \infty$ (i.e. the probability of obtaining 0 tends to 1). Then the left hand side tends to $\hat{\theta}(0)$, while the right hand side tends to $\infty$. 


2.3. UNBIASED PARAMETER ESTIMATION

2.3.2 The Cramér-Rao Lower Bound

In this section we will consider the case where the support of \( p(t; \theta) \) does not depend on \( \theta \) in which case the projection on the span of the elements of the score function

\[
S(Y; \theta) = \frac{\partial}{\partial \theta} \log p(Y; \theta) = \frac{\partial}{\partial \theta} p(Y; \theta)
\]

will turn out to be rewarding. We notice that the score function has zero mean

\[
\mathbb{E} [S(Y; \theta)] = \int \frac{\partial}{\partial \theta} p(y; \theta) p(y; \theta) dy = \frac{\partial}{\partial \theta} \int p(y; \theta) dy = \frac{\partial}{\partial \theta} 1 = 0
\]

Let \( \hat{\theta}(Y) \) be any unbiased estimator of \( \theta \) and let \( \hat{\theta}(Y)_{\mid S_{Le}} \) be the OLE estimator of \( \hat{\theta}(Y) \) given \( S(Y; \theta) \) and a constant, i.e.

\[
\hat{\theta}(Y)_{\mid S_{Le}} = \langle \hat{\theta}(Y) - \theta, S^T(Y; \theta) \rangle \langle S(Y; \theta), S(Y; \theta)^T \rangle^{-1} S + \theta
\]

\[
= \langle \hat{\theta}(Y), S^T(Y; \theta) \rangle \langle S(Y; \theta), S^T(Y; \theta) \rangle^{-1} S(Y; \theta) + \theta
\]

Let us introduce the Fisher Information Matrix (FIM)

\[
\langle I_F(\theta) := S(Y; \theta), S^T(Y; \theta) \rangle = \mathbb{E} [S(Y; \theta)S^T(Y; \theta)]
\]

Before proceeding we notice that

\[
\mathbb{E} \left[ \frac{\partial}{\partial \theta} S(Y; \theta) \right] = \mathbb{E} \left[ \frac{\partial^2 p(Y; \theta)}{\partial \theta^2} p(Y; \theta) \right] - \mathbb{E} [S(Y; \theta)S^T(Y; \theta)] = \frac{\partial^2}{\partial \theta \partial \theta^T} 1 - I_F(\theta) = -I_F(\theta)
\]

Notice now that

\[
\langle \hat{\theta}(Y), S^T(Y; \theta) \rangle = \mathbb{E} \left[ \hat{\theta}(Y)S^T(Y; \theta) \right] = \int \hat{\theta}(y) \frac{\partial}{\partial \theta} p(y; \theta) p(y; \theta) dy
\]

\[
= \int \hat{\theta}(y) \frac{\partial}{\partial \theta} p(y; \theta) dy = \frac{\partial}{\partial \theta} \int \hat{\theta}(y)p(y; \theta) dy = \frac{\partial}{\partial \theta} 1 = I
\]

The score function thus has the same cross-correlation with all unbiased estimators. This is a quite remarkable result as it means that the best estimator of any unbiased estimator based on the score function is independent of which unbiased estimator we want to estimate. This “estimator” is given by

\[
\hat{\theta}(Y)_{\mid S_{Le}} = I_F^{-1} S(Y; \theta) + \theta
\]

Here Pythagoras relation (2.17) becomes

\[
\text{Cov} \left[ \hat{\theta}(Y) - \hat{\theta}(Y)_{\mid S_{Le}} \right] = \text{Cov} \left[ \hat{\theta}(Y) \right] - \text{Cov} \left[ \hat{\theta}(Y)_{\mid S_{Le}} \right]
\]

where

\[
\text{Cov} \left[ \hat{\theta}(Y)_{\mid S_{Le}} \right] = \mathbb{E} \left[ I_F^{-1}(\theta)S(Y; \theta)S^T(Y; \theta)I_F^{-1}(\theta) \right] = I_F^{-1}(\theta)
\]
and since a covariance matrix is positive semi-definite we have that
\[
\text{Cov} \left[ \hat{\theta}(Y) \right] \geq I_F^{-1}(\theta) \tag{2.28}
\]
holds for any unbiased estimator \( \hat{\theta}(Y) \). This is the Cramér-Rao lower bound (CRLB).

Suppose now a general function \( f(Y) : \mathbb{R}^N \rightarrow \mathbb{R}^m \) for some \( m \). Repeating the derivations above gives that
\[
\text{Cov} \left[ f(Y) \right] \geq \frac{\partial}{\partial \theta} \mathbb{E} [ f(Y) ] \frac{\partial}{\partial \theta} \mathbb{E} [ f(Y) ]^T \tag{2.29}
\]
We can use this result to derive a CRLB for biased estimators. Suppose that \( \hat{\theta}(Y) \) has bias \( b(\theta) \), i.e.
\[
\mathbb{E} \left[ \hat{\theta}(Y) \right] = \theta + b(\theta)
\]
then (2.29) gives directly that
\[
\text{Cov} \left[ \hat{\theta}(Y) \right] \geq \left( I + b'(\theta) \right) I_F^{-1}(\theta) \left( I + b'(\theta) \right)^T \tag{2.30}
\]

### 2.3.3 Efficient Estimators

An unbiased estimator is efficient if it attains the CRLB. Equality in (2.28) will hold only if
\[
\hat{\theta}(Y) = \hat{\theta}(Y)_{\|S_L = I} = I_F^{-1}(\theta) S(Y; \theta) + \theta \quad \text{a.e.} \tag{2.31}
\]
but for this to be an estimator it cannot depend on \( \theta \).

For a given pdf \( p(Y; \theta) \) one can always check if (2.31) is independent of \( \theta \), in which case it is an estimator which is efficient.

Existence of efficient estimators has been studied extensively. Here we will consider exponential families.

**Efficient Estimators for Exponential Families**

Exponential families have pdfs that can be written as
\[
p(y; \theta) = e^{\varphi^T(\theta) T(y) - B(\theta) h(y)} \tag{2.32}
\]
where \( \varphi(\cdot) : \mathbb{R}^s \rightarrow \mathbb{R}^s \) is 1-1 and twice differentiable. Introducing \( \eta = \varphi(\theta) \) we can write the pdf on the canonical form
\[
e^{\eta^T T(y) - A(\eta) h(y)} \tag{2.33}
\]
where \( A(\eta) = B(\varphi^{-1}(\eta)) \). For such a family it holds that \( T(Y) \) has distribution
\[
p(t; \eta) = e^{\eta^T t - A(\eta) k(t)} \tag{2.34}
\]
for some positive function \( k \). This distribution has mean and covariance matrix
\[
\mathbb{E} \left[ T(Y) \right] = A'(\eta) = A'(\varphi(\theta)) = \left( \varphi'(\theta) \right)^T B'(\theta), \quad \text{Cov} \left[ T(Y) \right] = A''(\eta) = A''(\varphi(\theta)) \tag{2.35}
\]
2.3. UNBIASED PARAMETER ESTIMATION

For such a family, the score function is given by

\[ S(Y; \theta) = \frac{\partial}{\partial \theta} \log p(Y; \theta) = (\varphi'(\theta))^T T(Y) - B'(\theta) \]

which, using (2.35), can be written as

\[ S(Y; \theta) = (\varphi'(\theta))^T (T(Y) - \mathbb{E}[T(Y)]) \]

giving the Fisher Information Matrix

\[ I_F(\theta) = \mathbb{E} [S(Y; \theta)S^T(Y; \theta)] = (\varphi'(\theta))^T \text{Cov}[T(Y)] \varphi'(\theta) \]

Hence an efficient estimator has to have the form

\[ \hat{\theta}(Y)_{S_{Le}} = I_F^{-1}(\theta)S(Y; \theta) + \theta \]

from which we see that a second necessary condition is that

\[ \mathbb{E}[T(Y)] = \theta \] (2.39)

However if (2.39) holds, differentiating the first equation in (2.35) gives

\[ I = A''(\varphi(\theta))\varphi'(\theta) = \text{Cov}[T(Y)] \varphi'(\theta) \]

which is (2.38), and hence (2.39) is sufficient (and necessary) for (2.36) to be independent of \( \theta \). Under (2.39), the FIM thus simplifies to

\[ I_F(\theta) = \text{Cov}^{-1}[T(Y)] \]

We have thus established that for the exponential family (2.32) satisfying (2.39), the estimator

\[ \hat{\theta}_{ML}(Y) := T(Y) \] (2.40)

is efficient and the CRLB is \( \text{Cov}[T(Y)] = A''(\varphi(\theta)) \).

Efficient Estimators for General Distributions

A perhaps very disappointing result is that if an estimator is efficient, then the pdf \( p(Y; \theta) \) belongs to an exponential family (2.32) and (2.39) holds. Thus, the class of problems for which there exist an efficient estimator is very limited.

We can also reverse the reasoning above. Suppose that we have the pdf (2.33) where \( \eta = \varphi(\theta) \). Then there exists an efficient estimator for \( \theta \) iff

\[ \mathbb{E}[T(Y)] = A'(\varphi(\theta)) = \theta \]

i.e. it has to hold that \( A'(\eta) = \varphi^{-1}(\eta) \) for all \( \eta \) under consideration. Since the split into \( T \) and \( \varphi \) is not unique any linear combination of \( \theta \) can of course also be estimated efficiently.
2.3.4 The Maximum Likelihood Estimator

For an exponential family (2.32) satisfying (2.38) and (2.39), the score function is given by

$$ S(Y; \theta) = \text{Cov}^{-1}[T(Y)](T(Y) - \theta) $$

and hence the efficient estimator (2.40) solves the equation

$$ S(Y; \hat{\theta}_{ML}(Y)) = 0 $$

i.e. $\hat{\theta}_{ML}(Y)$ is a stationary point of $p(Y; \theta)$. However,

$$ \frac{\partial}{\partial \theta^T} S(Y; \theta) = -\text{Cov}^{-1}[T(Y)] + L(T(Y) - \theta, \theta) $$

for some matrix $L$ linear in $T(Y) - \theta$. Thus

$$ \frac{\partial}{\partial \theta^T} S(Y; \hat{\theta}_{ML}(Y)) = -\text{Cov}^{-1}[T(Y)] = -I_F(\theta) < 0 $$

if the FIM is positive definite.

Thus $\hat{\theta}_{ML}(Y)$ is a local maximum of $\log p(Y; \theta)$ and therefore, since log is a monotone increasing function, a local maximum of $p(Y; \theta)$:

$$ \hat{\theta}_{ML} = \arg \max_{\theta \in D_{\theta}} p(Y; \theta) $$

Above $D_{\theta}$ is the domain of definition of the pdf. We call such an estimator a Maximum Likelihood (ML) estimator.

2.3.5 Data compression

Let as before $Y \in \mathbb{R}^N$ be a random vector with pdf $p(Y; \theta)$. A statistic is any function of the observation $Z = T(Y) \in \mathbb{R}^M$. The question is now if we loose information about $\theta$ if we instead of using an observation $y$ of $Y$ to estimate $\theta$, use the statistic $z = T(y)$. When $T$ is one-to-one, there is of course no loss since we then can re-create $y$ from $z$. Now we can use Bayes rule to express the pdf of $Y$ as

$$ p(y; \theta) = p(y|z; \theta)p(z; \theta) $$

Suppose now that $p(y|z; \theta) = p(y|z)$, i.e. the conditional distribution of $Y$ given $Z$ is independent of $\theta$. Then given $z = T(y)$ we can draw a sample $\bar{y}$ from the conditional distribution $p(y|z)$ which comes from exactly the same distribution as the original sample $y$. Thus we can equally well use $\bar{y}$ instead of $y$ to estimate $\theta$, suggesting that there is no loss of information in using $z$.

A statistic $Z = T(Y)$ for which $p(y|z; \theta) = p(y|z)$ holds is called a sufficient statistic. The next result tells us how to find sufficient statistics given the pdf.

**Theorem 2.3.1** (Neyman-Fisher Factorization Theorem). A sufficient statistic for $Y$ exists iff the pdf has the factorization

$$ p(y; \theta) = g(f(y); \theta)h(y) $$

Furthermore, $f(Y)$ is a sufficient statistic.
2.3. UNBIASED PARAMETER ESTIMATION

Proof. Suppose there exists a sufficient statistic \( f(Y) \). Then we can take \( h(y) = p(y|f(y)) \) and \( g(f(y); \theta) = p(f(y); \theta) \).

Conversely, let \( z = f(y) \). Then

\[
p(y|z; \theta) = \frac{p(y; \theta)}{p(z; \theta)} = \frac{p(y; \theta)}{\int_{f(y)=z} p(y; \theta) dy} = \frac{g(f(y); \theta) h(y)}{\int_{f(y)=z} h(y) dy}
\]

which does not depend on \( \theta \).

We would suspect that the CRLB should not depend on if we use the original data or a sufficient statistic and indeed this is the case. This follows directly from that it is the score function that determines the CRLB and for a sufficient statistic \( Z \) we have

\[
S(Y; \theta) = \frac{\partial}{\partial \theta} \log p(Y; \theta) = \frac{\partial}{\partial \theta} \log p(Y|Z)p(Z; \theta) = \frac{\partial}{\partial \theta} \log p(Z; \theta) = S(Z; \theta)
\]

Maximal data compression

A sufficient statistic \( T \) is said to be minimal if for every other sufficient statistic \( U \) there is a function \( H_U \) such that \( T = H_U(U) \).

We can refine the notion of data reduction by introducing the concept of ancillarity. A statistic \( V(Y) \) is said to be ancillary if its distribution does not depend on \( \theta \). Clearly such a statistic is useless for estimating \( \theta \). \( V(Y) \) is first order ancillary if its mean is independent of \( \theta \).

If a sufficient statistic \( T \) is such that there is some function \( f \) such that the distribution of \( f \) does not depend on \( \theta \), we may suspect that it should be possible to simplify it without losing information about \( \theta \). A statistic \( T \) is said to be complete if there is no non-constant function \( f \) such that \( f(T) \) is first order ancillary.

A complete sufficient statistic is always minimal.

Finite dimensional data compression

Example 2.3.2. Suppose that \( Y_1, \ldots, Y_N \) are iid \( N(m, \sigma^2) \). Then the joint pdf is given by

\[
p(y^N; \theta) = \prod_{k=1}^{N} p(y; \theta) = \frac{1}{(2\pi)^{N/2}} \frac{1}{(\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^{N} (y_t - m)^2} = \frac{1}{(2\pi)^{N/2}} \frac{1}{(\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2} (\sum_{k=1}^{N} y_t^2 - 2m \sum_{k=1}^{N} y_t + m^2)}
\]

from which we see that

\[
T(Y) = \begin{bmatrix} \sum_{k=1}^{N} y_k \\ \sum_{k=1}^{N} y_k^2 \end{bmatrix}
\]

is a complete sufficient statistic.

For iid normal distributed random variables, no matter how many samples we have, the data can always be compressed into a two-dimensional vector without loss of information. A natural question is if this attractive feature holds in general. Unfortunately, this is not true in general.
Theorem 2.3.2 (Theorem 6.18 in [10]). Suppose that \( Y_1, \ldots, Y_N \) are iid with pdf \( p(y_i; \theta) \), which is continuous in \( y \) and whose support for all \( \theta \) is an interval \( I \). Suppose that for the joint density of \( Y_1, \ldots, Y_N \)

\[
p(y; \theta) = \prod_{k=1}^{N} p(y_k; \theta)
\]

there exists a continuous \( r \)-dimensional sufficient statistic. Then

(i) if \( r = 1 \), \( p(y; \theta) \) is an exponential family (2.32) with \( s = 1 \).

(ii) if \( r > 1 \), and if the densities \( p(y_i; \theta) \) have continuous partial derivatives with respect to \( y_i \), then \( p(y; \theta) \) is an exponential family (2.32) with \( s \leq r \).

Example 2.3.3. If \( Y_1, \ldots, Y_N \) are iid with uniform distribution \([0, \theta]\) it can be shown that

\[
\max_{1 \leq k \leq N} Y_k
\]

is a sufficient statistic despite that the joint distribution of \( Y_1, \ldots, Y_N \) does not belong to the exponential family (2.32). The reason why the theorem above does not apply is that the support depends on \( \theta \).

Remark 2.3.1. While it may be regarded as unfortunate that no other distributions than the exponential families allow finite dimensional compression of data, it should be compared to the fact that exponential distributions in a sense contain the least information. Recall that for exponential families, the conditional mean is linear in the observed variable while for other distributions it is a non-linear function which gives a better estimate (in terms of the MSE) of the variable of interest.

2.3.6 Uniform Minimum Variance Unbiased (UMVU) Estimators

At the beginning of Section 2.3.1 we discussed the possibility to project an unbiased estimator on a subspace to reduce its variance and hence improve its accuracy. We also realized that the main issue with this approach is that the projection may depend on \( \theta \), rendering the projection useless as an estimator. We proceeded with the ambitious aim to characterize estimators that are efficient, i.e. that are unbiased and reach the CRLB. This analysis was based on the restrictive assumption that the support of \( p(y; \theta) \) does not depend on \( \theta \), precluding, e.g., uniform distributions from the analysis. From this analysis we also saw that the classes of distributions for which efficient estimators exist is limited to exponential families, and then only certain parameters can be estimated efficiently.

We will now lower our objective somewhat. Rather than aiming for estimators reaching the CRLB, we aim to derive estimation methods that generate uniform minimum variance unbiased (UMVU) estimators. Such estimators have covariance not larger than any other estimator for all \( \theta \). For this, we will return to the idea of reducing the variance of an estimator by projecting it on some subspace. To avoid the problem that such a projection will be parameter dependent we will now project on the \( \sigma \)-algebra generated by a sufficient statistic \( T(Y) \). Recall that in that case \( p(y|T(y)) \) does not depend on \( \theta \) and hence the orthogonal projection \( \mathbb{E} \left[ \hat{\theta}(Y)|T(Y) \right] \) does not depend on \( \theta \):

\[
\mathbb{E} \left[ \hat{\theta}(Y)|T(Y) = z \right] = \int \hat{\theta}(y)p_{Y|T(Y)}(y|z)dy
\]

We also realize that this projection will reduce the MSE also if the \( \hat{\theta}(Y) \) is not unbiased since the projection has the same mean as \( \hat{\theta}(Y) \) but smaller covariance. We summarize this result.
Theorem 2.3.3 (Rao-Blackwell). Let \( T(Y) \) be a sufficient statistic for the distribution \( p(y; \theta) \), \( \theta \in D_\theta \) of the random variable \( Y \). Let \( \hat{\theta}(Y) \) be any estimator of \( \theta \) with finite expectation and MSE. Then

\[
E \left[ \hat{\theta}(Y) | T(Y) \right]
\]

has strictly smaller MSE than \( \hat{\theta}(Y) \) unless the two estimators coincide a.e.

Suppose that \( \hat{\theta}_i, i = 1, 2 \) are two unbiased estimators. While the Rao-Blackwell theorem gives that \( E \left[ \hat{\theta}_1 | T(Y) \right] \) has lower MSE than \( \hat{\theta}_1 \), it may not necessarily be better than \( \hat{\theta}_2 \). Recall, that when we projected on the span of the score function, the orthogonal projection was independent of which unbiased estimator we started with. We need this to be the case here as well in order for \( E \left[ \hat{\theta}_1 | T(Y) \right] \) to be UMVU. The requirement is that \( T \) is complete.

Corollary 2.3.1. If in addition to the assumptions in Theorem 2.3.3, \( \hat{\theta}(Y) \) is unbiased and \( T(Y) \) is complete, \( E \left[ \hat{\theta}(Y) | T(Y) \right] \) is the UMVU estimator.

Proof. From Theorem 2.3.3 it follows that an UMVU estimator must be a function of \( T(Y) \). Suppose now that there are two unbiased estimators \( \hat{\theta}_1(T(Y)) \) and \( \hat{\theta}_2(T(Y)) \) of \( \theta \) that are functions of \( T(Y) \). Then

\[
E \left[ \hat{\theta}_1(T(Y)) - \hat{\theta}_2(T(Y)) \right] = 0 \quad \forall \theta \in D_\theta
\]

but since \( T \) is complete this gives that \( \hat{\theta}_1(T(Y)) = \hat{\theta}_2(T(Y)) \) a.e. with respect to \( p \). \( \square \)

We can give the corollary a Hilbert space interpretation. For simplicity consider the scalar setting that \( Y \) is a random variable. Consider the setting in Section 2.2.3. Let \( S \) be the subspace consisting of functions of \( Y \) having zero mean. Then the set of unbiased estimators of \( \theta \), i.e. the set of functions of \( Y \) that have expectation \( \theta \) form a linear variety: they are the sum of any unbiased estimator \( \hat{\theta}(Y) \) and \( S \). The squared norm for an element in this variety is its variance + the squared norm of \( \theta \), where the latter can be seen as a constant. Thus, the UMVU estimator is the point in this linear variety with minimum norm. But then Corollary 1.1.1 gives that there is a unique such point \( \hat{\theta}_{UMVU} \) in this variety and that it is uniquely characterized as being orthogonal to all elements of \( S \):

\[
\langle \hat{\theta}_{UMVU}, \Delta(Y) \rangle = 0, \quad \forall \Delta(Y) \text{ s.t. } E[\Delta(Y)] = 0 \quad (2.41)
\]

For \( \hat{\theta}_{UMVU} \) to be UMVU (2.41) has to hold for all \( \theta \in D_\theta \). If there exists a complete sufficient statistic \( T(Y) \), this point can be computed from \( T(Y) \) as follows. Recall that complete means that if the mean of a function of the sufficient statistic is zero for all \( \theta \), then this function must be zero a.e. Let \( \Delta(Y) \) have zero mean regardless of \( \theta \), then \( E_{Y|T(Y)} [\Delta(Y)|T(Y)] \) is a function of \( T \) and satisfies

\[
E \left[ E_{Y|T(Y)} [\Delta(Y)|T(Y)] \right] = E[\Delta(Y)] = 0 \quad \forall \theta \in D_\theta
\]

Thus completeness implies that \( E_{Y|T(Y)} [\Delta(Y)|T(Y)] = 0 \) a.e for any zero mean \( \Delta(Y) \).

Now take any unbiased \( \hat{\theta}(Y) \), then for any zero mean \( \Delta(Y) \)

\[
E \left[ \hat{\theta}(Y)|T(Y) \right] = E \left[ \Delta(Y) E \left[ \hat{\theta}(Y)|T(Y) \right] \right]
\]

\[
= E \left[ E \left[ \Delta(Y) E \left[ \hat{\theta}(Y)|T(Y) \right] | T(Y) \right] \right]
\]

\[
= E \left[ E \left[ \hat{\theta}(Y)|T(Y) \right] E[\Delta(Y)|T(Y)] \right]
\]

\[
= E \left[ E \left[ \hat{\theta}(Y)|T(Y) \right] 0 \right] = 0
\]
CHAPTER 2. INTRODUCTION TO ESTIMATION

Thus $E_{Y|T(Y)} \left[ \hat{\theta}(Y)|T(Y) \right]$ is UMVU when $T$ is a complete sufficient statistic.

Given a complete sufficient statistic $T(Y)$, rather than explicitly computing the conditional mean with respect to this statistic for some arbitrary unbiased estimator, often it can be easier to explicitly try to find a function $g(T)$ which is unbiased. If such a function exists it has to be the conditional mean since the $T$ is complete.

UMVU for exponential families

For an exponential family

$$p_Y(y; \theta) = e^{\theta^T T(y) - A(\theta) h(y)} \quad (2.42)$$

where $T$ is complete, it follows trivially that $T(Y)$ is UMVU for estimating $E[Y]$, or using (2.35), $A'(\theta)$.

When it comes the natural parameter $\theta$, using Stein’s identity Lemma 4.A.1 with $g(y) = 1$ gives that for (2.42)

$$E \left[ -\frac{d}{dY} \log h(Y) \right] = E \left[ T'(Y) \right]^T \theta$$

which, in case that $T(Y) = Y$, gives that

$$-\frac{d}{dY} \log h(Y)$$

is an unbiased estimate of $\theta$. Since $T$ is sufficient, it follows that when $T$ is complete the UMVU estimator of $\theta$ is given by

$$\hat{\theta}_{UMVU} = -\frac{d}{dY} \log h(Y)$$

which, again using Stein’s identity, can be shown to have covariance

$$\text{Cov} \left[ \hat{\theta}_{UMVU} \right] = -E \left[ \frac{d}{dY} \log h(Y) \frac{d}{dY} \log h(Y) \right]$$

Generalizing the Rao-Blackwell theorem

We can strengthen the Rao-Blackwell theorem using Jensen’s inequality

$$\Phi \left( E[Y] \right) \leq E \left[ \Phi(Y) \right] \quad (2.43)$$

which applies if $\Phi$ is a convex function defined over an open interval and $Y$ is a random variable with range being the domain of $\Phi$ having finite expectation. The inequality in (2.43) is strict if $\Phi$ is strictly convex unless $Y$ is constant a.e. Jensen’s inequality applies also to conditional expectations.

**Lemma 2.3.1.** If $\Phi$ is a convex function defined over an open interval $I$, and $X$ is a random variable with $P(X \in I) = 1$ and finite expectation, then

$$\Phi (E[X]) \leq E \left[ \Phi(X) \right]$$

If $\Phi$ is strictly convex then the inequality is strict unless $X$ is constant a.e.
Let $L(\theta, \hat{\theta})$ be strictly convex in $\hat{\theta}$. Then using Jensen’s inequality on the conditional mean gives

$$L(\theta, \mathbb{E}[\hat{\theta}(Y)|T(Y)]) < \mathbb{E}[L(\theta, \hat{\theta}(Y))|T(Y)]$$

unless $\mathbb{E}[\hat{\theta}(Y)|T(Y)] = \hat{\theta}(Y)$ a.e. Taking expectation on both sides of the inequality above gives

$$\mathbb{E}[L(\theta, \mathbb{E}[\hat{\theta}(Y)|T(Y)])] < \mathbb{E}[L(\theta, \hat{\theta}(Y))]}$$

Thus conditioning on a sufficient statistic strictly reduces the risk when $L$ is strictly convex.

### 2.3.7 Efficient Estimators, Sufficient Statistics and Maximum Likelihood Estimators

It is known that for an exponential family (2.32), $T(Y)$ is a complete sufficient statistic provided that there is no $L \in \mathbb{R}^s$ such that $L^T T(y) = 0$ and $\theta$ is not subject to any linear constraint, and the parameter space contains an $s$-dimensional rectangle.

In Section 2.3.3 we saw that the projection of an unbiased estimator on the span of the score function resulted in the estimator $T(Y)$ for an exponential family subject to conditions (2.38) and (2.39). Since this is the only family of distributions for which an efficient estimator exists it follows that an efficient estimator is also a sufficient statistic.

Finally, we saw in Section 2.3.4 that an efficient estimator is a ML estimator.

$$\text{Efficient} \Rightarrow \text{Sufficient statistic} \Rightarrow \text{ML estimator} \quad (2.44)$$

### 2.3.8 Parameter Transformations

#### The Invariance Principle

Suppose that $p(y; \theta)$, $\theta \in D_\theta \subset \mathbb{R}^n$, is a parametrization of the pdf of the random variable $Y$ and let $\hat{\theta}_{ML}$ be the ML estimator of $\theta$:

$$\hat{\theta}_{ML} = \hat{\theta}_{ML}(Y) = \arg \max_{\theta \in D_\theta} p(Y; \theta)$$

Suppose now that $\tilde{p}(y; \alpha)$, $\alpha \in D_\alpha \subset \mathbb{R}^m$, is another parametrization of the pdf and suppose that there is a function $q : \mathbb{R}^n \to \mathbb{R}^m$ that is 1-1 and maps $D_\theta$ to $D_\alpha$. Then $p(y; \theta) = \tilde{p}(y; q(\theta))$ and a ML estimator of $\alpha$ is given by

$$\hat{\alpha}_{ML} = q(\hat{\theta}_{ML}) \quad (2.45)$$

since for any $\alpha \in D_\alpha$ there exists a $\theta \in D_\theta$ such that $\alpha = q(\theta)$ and hence

$$\tilde{p}(y; \hat{\alpha}_{ML}) = \tilde{p}(y; q(\hat{\theta}_{ML})) = p(y; \hat{\theta}_{ML}) \geq p(y; \theta) = \tilde{p}(y; q(\theta)) = \tilde{p}(y; \alpha)$$

for any $\alpha \in D_\alpha$. The relation (2.45) is known as the invariance principle. We notice that this principle holds even if the inverse of $q$ is not unique, as long as the range of $q$ is $D_\alpha$. We may phrase the invariance principle in words as

The ML estimate of any quantity is obtained by using the ML-estimate of the unknown parameter in the underlying pdf in place of its true value.
Efficient Estimators

Suppose now that $\hat{\alpha}$ is an efficient estimator for $\alpha$ in a set $\tilde{D}_\alpha$ larger than $D_\alpha$: $\tilde{D}_\alpha \supset D_\alpha$. Can we then recover a ML estimator of $\theta$ from $\hat{\alpha}$? Since $\hat{\alpha}$ is efficient, the distribution of $Y$ must be given by an exponential family. For simplicity we will assume that the pdf has the form

$$p(y; \alpha) = e^{\alpha^T \Sigma^{-1} \hat{\alpha}(y) - \frac{1}{2} \alpha^T \Sigma^{-1} \alpha} h(y)$$

for which $\hat{\alpha}(Y)$ is a complete sufficient statistic. This is a pdf of the form (2.32), with $\theta$ replaced by $\alpha$ and with

$$\varphi(\alpha) = \Sigma^{-1} \alpha, \quad B(\alpha) = \frac{1}{2} \alpha^T \Sigma^{-1} \alpha$$

From the properties of exponential families we know that $\hat{\alpha}(Y)$ has distribution

$$p(x; \alpha) = e^{\alpha^T \Sigma^{-1} x - \frac{1}{2} \alpha^T \Sigma^{-1} \alpha} k(x)$$

Using that $\varphi^{-1}(\eta) = \Sigma \eta$ we can express this as (2.33) with

$$A(\eta) = B(\varphi^{-1}(\eta)) = \frac{1}{2} \eta^T \Sigma \Sigma^{-1} \eta = \frac{1}{2} \eta^T \Sigma \eta$$

and thus the mean of $\hat{\alpha}(Y)$ is $A'(\eta) = \Sigma \eta = \alpha$ and covariance matrix

$$\text{Cov} [\hat{\alpha}(Y)] = A''(\eta) = \Sigma$$

As in the preceding section, let $q(\theta)$ be the map from $D_\theta$ to $D_\alpha$. Then the pdf of $\hat{\alpha}(Y)$ is given by

$$p(x; \theta) = p(x; q(\theta)) = e^{\theta^T (\Sigma^{-1} x - \frac{1}{2} \alpha^T \Sigma^{-1} \alpha)} k(x)$$

When $q$ is 1-1 not much has happened. We know from previous that for an exponential family that there is only a certain parametrization that can be estimated efficiently and those parameters are given by $A'(\alpha)$. The more interesting case is when $\theta$ lies in a lower dimensional space than $\alpha$. For simplicity suppose that

$$q(\theta) = \Phi \theta, \quad \Phi \in \mathbb{R}^{m \times n}, \quad m > n$$

where $\Phi$ has full column rank. We then have

$$p(x; \theta) = e^{\theta^T \Phi^T \Sigma^{-1} \Phi \theta} k(x)$$

Introducing

$$\hat{\theta}(x) = (\Phi^T \Sigma^{-1} \Phi)^{-1} \Phi^T \Sigma^{-1} x \in \mathbb{R}^n$$

we can write this as

$$p(x; \theta) = e^{\theta^T \Phi^T \Sigma^{-1} \Phi \hat{\theta}(x) - \frac{1}{2} \theta^T \Phi^T \Sigma^{-1} \Phi \theta} k(x)$$

This pdf has the same form as (2.46) and hence $\hat{\theta}(\hat{\alpha}(Y))$ is a complete sufficient statistic and

$$\mathbb{E} [\hat{\theta}(\hat{\alpha}(Y))] = \theta$$

$$\text{Cov} [\hat{\theta}(\hat{\alpha}(Y))] = (\Phi^T \Sigma^{-1} \Phi)^{-1}$$
2.3. UNBIASED PARAMETER ESTIMATION

We can also obtain the mean explicitly from

$$
E\left[\hat{\theta}(\hat{\alpha})\right] = (\Phi^T\Sigma^{-1}\Phi)^{-1}\Phi^T\Sigma^{-1}E[\hat{\alpha}] = (\Phi^T\Sigma^{-1}\Phi)^{-1}\Phi^T\Sigma^{-1}\Phi \theta = \theta
$$

In conclusion,

$$
\hat{\theta}(\hat{\alpha}(Y)) = (\Phi^T\Sigma^{-1}\Phi)^{-1}\Phi^T\Sigma^{-1}\hat{\alpha}(Y)
$$

is an efficient estimator and $(\Phi^T\Sigma^{-1}\Phi)^{-1}$ is the smallest covariance matrix of an unbiased estimator of $\theta$.

The case when $q$ is a nonlinear map leads to curved exponential families and is considerably more difficult. We will not pursue this here.

2.3.9 Best Linear Unbiased Estimators (BLUE)

Let $Y \in \mathbb{R}^N$ be a random vector with mean and covariance

$$
E[Y] = \Phi \theta, \quad \theta \in \mathbb{R}^n, \quad n \leq N, \quad \text{Cov}[Y] = \Sigma > 0
$$

The class of linear unbiased estimators of $\theta$ is given by

$$
\{LY : L \in \mathbb{R}^{n \times N}, L\Phi = I\}
$$

The BLUE is the estimator in this class that has the smallest covariance. To find this estimator we will follow the same path as in Section 2.3.1 when we derived the CRLB. Recall that, there the score function was instrumental and we used two properties of this function: that it has zero mean and that it has the same covariance, the identity, with any unbiased estimator. From these properties it followed directly that the projection of any unbiased estimator on the span of the score function and a constant had lower variance than the original estimator. If the projection did not depend on $\theta$, it was an estimator and it followed that it reached the CRLB.

In this section the aim is less ambitious as we only consider the class of linear estimators (2.49), and not all estimators. If we can find a “score” function which has zero mean and the identity as covariance with all estimators in (2.49), our derivations in Section 2.3.1 provide a lower bound on the covariance of any linear unbiased estimator of $\theta$. If in addition the projection in question is independent of $\theta$, it is the BLUE.

We will see if we can find a “score” function that is affine in $Y$:

$$
S(Y) = SY + m, \quad S \in \mathbb{R}^{n \times N}
$$

For $S(Y)$ to have zero mean we see that $m = -S\mathbb{E}[Y]$ is necessary giving

$$
S(Y) = S(Y - \mathbb{E}[Y])
$$

Now, let $\hat{\theta}$ be a linear unbiased estimator of $\theta$, i.e. $\hat{\theta} = LY$, where $L\Phi = I$ for some $L \in \mathbb{R}^{n \times N}$. Then

$$
\langle \hat{\theta}, S(Y)^T \rangle = E[LY(Y - \mathbb{E}[Y])S^T] = L\Sigma \tilde{S}^T
$$

Taking $\tilde{S} = \Phi^T\Sigma^{-1}$ gives

$$
\langle \hat{\theta}, S(Y)^T \rangle = I
$$
For this $S$

\[
\langle S(Y), S(Y)^T \rangle = \Phi^T \Sigma^{-1} \Sigma \Sigma^{-1} \Phi = \Phi^T \Sigma^{-1} \Phi
\]  

(2.50)

giving the projection on the subspace generated by $S(Y)$ and a constant (2.26) as

\[
\hat{\theta}_{\parallel S Le} = (\Phi^T \Sigma^{-1} \Phi)^{-1} \Phi \Sigma^{-1} (Y - \mathbb{E}[Y]) + \theta
\]

\[
= (\Phi^T \Sigma^{-1} \Phi)^{-1} \Phi \Sigma^{-1} (Y - \Phi \theta) + \theta
\]

\[
= (\Phi^T \Sigma^{-1} \Phi)^{-1} \Phi \Sigma^{-1} Y
\]

(2.51)

from which we see that this is a bona fide estimator, and hence the BLUE. The quantity (2.50) is the equivalent to the Fisher Information Matrix in this setting. We also observe that $S(Y)$ can be seen as the projection of $Y$ on the subspace spanned by the columns of $\Phi$, i.e. the subspace where $\alpha$ lives. More precisely,

\[
Y = Y_{\parallel} + Y_{\perp}
\]

where $Y_{\parallel} = \Phi (\Phi^T \Phi)^{-1} \Phi^T Y$ and $Y_{\perp} = Y - Y_{\parallel}$. Hence $\Phi^T Y = \Phi^T Y_{\parallel}$.

We summarize, the BLUE under the assumptions (2.48) is given by

\[
\hat{\theta}_{\text{BLUE}} = (\Phi^T \Sigma^{-1} \Phi)^{-1} \Phi \Sigma^{-1} Y
\]

(2.52)

**Remark 2.3.2.** Returning to the Section 2.3.8 and the paragraph on efficient estimators, we see that $\hat{\theta}(\hat{\alpha}(Y))$, defined in (2.47), is nothing but the BLUE of $\theta$ given $\hat{\alpha}$ which has mean $\alpha = \Phi \theta$. This means that $\hat{\theta}(\hat{\alpha}(Y))$ is the projection of $\hat{\alpha}(Y)$.

In the setting of this section we can interpret $\Phi \theta$ as $\alpha$ and $Y$ as our estimate $\hat{\alpha}$: $\hat{\alpha} = Y$. Thus we conclude that the BLUE is efficient when, in addition to the assumptions (2.48), the distribution of $Y$ is given by an exponential family. In particular when $Y$ is normal distributed $N(\Phi \theta, \Sigma)$, this result applies.

**Remark 2.3.3.** The BLUE is also the solution to

\[
\hat{\theta}_N = \arg \min_{\theta} V_N(\theta), \text{ where }
\]

\[
V_N(\theta) = \frac{1}{N} (Y - \bar{Y}(\theta))^T \Sigma^{-1} (Y - \bar{Y}(\theta))
\]

(2.53)

where the estimator $\bar{Y}(\theta)$ is the mean $\bar{Y}(\theta) = \Phi \theta$.

Using this observation we can say a little bit more. Let $W \in \mathbb{R}^{N \times N}$, $W > 0$ and consider the weighted least-squares estimate

\[
\hat{\theta}(W) = \arg \min_{\theta} V_N(\theta, W), \text{ where }
\]

\[
V_N(\theta, W) = \frac{1}{N} (Y - \bar{Y}(\theta))^T W (Y - \bar{Y}(\theta))
\]

(2.54)

This estimate is unbiased regardless of $W$ and belongs therefore to the class (2.49). Since $W = \Sigma^{-1}$ gives the BLUE, we conclude that the optimal weighting in a least-squares criterion is the inverse covariance matrix of the observations.
Suppose for example that $\Sigma$ is diagonal, i.e. that the observations are uncorrelated. Then the optimal criterion is

$$V_N(\theta) = \frac{1}{N} \sum_{k=1}^{N} \frac{(Y_k - \varphi_k^T \theta)^2}{\Sigma_{kk}}$$

where $\varphi_k^T$ is the $k$th row of $\Phi$. Thus the optimal weighting is the inverse of the variance of each measurement, i.e. more weight should be given to more accurate measurements -a quite intuitive result.

**Remark 2.3.4.** There are a number of other ways to derive the BLUE. We refer to the proof of Lemma 4.3 in [13]. In particular, knowing the solution, one can use Schur complement (Section 2.B). We have

$$0 \leq \text{Cov}\left[ \begin{bmatrix} \hat{\theta} \\ S(Y) \end{bmatrix} \right] = \begin{bmatrix} \text{Cov}(\hat{\theta}) & I \\ I & \Phi^T \Sigma^{-1} \Phi \end{bmatrix}$$

(2.55)

Since $\Phi$ is assumed to be full column rank and $\Sigma > 0$, $\Phi^T \Sigma^{-1} \Phi > 0$ and hence Schur complement gives that (2.55) is equivalent to

$$\text{Cov}\left[ \hat{\theta} \right] \geq I (\Phi^T \Sigma^{-1} \Phi)^{-1} I = (\Phi^T \Sigma^{-1} \Phi)^{-1}$$

and the lower bound is reached for the BLUE (2.52).

### 2.3.10 Exercises

2.3.1. Let $Y_1, \ldots, Y_N$ be independent Gaussian distributed random variables with mean $\theta$ and variance 1.

(a) Show that $Y_1$ is an unbiased estimator of $\theta$.

(b) Derive a minimal sufficient statistic for $\theta$.

(c) Use the Rao-Blackwell theorem to improve the estimator in (a). Is this estimator UMVU?

(d) What is the score function and the CRLB? Do the estimators in a) and c) reach the CRLB?

(e) Project the estimator in a) on the span of the score function and a constant. Is the result an estimator? Compare with the one in c).

2.3.2. * Let $Y_1, \ldots, Y_N$ be independent Gaussian distributed random variables with mean $\mu$ and variance $\sigma^2$ and let $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}^T$.

(a) Compute the CRLB for $\theta$.

(b) Derive a minimal sufficient statistic for $\theta$.

(c) Use that your statistic in (b) is complete to compute the UMVU estimator for $\theta$. Does it reach the CRLB?

(d) Use the reasoning at the end of Section 2.3.3 to interpret your findings in (a) and (c). Which parametrization of the pdf should be used for the CRLB to be reachable?

2.3.3. Let $Y_1, \ldots, Y_N$ be independent identically distributed (iid) random variables having pdf $p(y; \theta)$. 

(a) Show that the FIM is $O(N^a)$ for some $a$. Determine $a$. Express the CRLB in Ordo of $N$.

Below, assume that the $Y_i$:s are uniformly distributed on the interval $[0, \theta]$, i.e. $p(y) = 1/\theta$ for $y \in [0, \theta]$.

(b) What is the CRLB for $\theta$?

(c) Consider the order statistics estimator $\hat{\theta}(Y) = \frac{N+1}{N} \max_{1 \leq k \leq N} Y_k$. Show that this estimator is unbiased.

Hint: You can look up the distribution of $\max_{1 \leq k \leq N} Y_k$.

(d) What is the variance of the estimator in c)? Compare with b). Comment?

2.3.4. Suppose that $E$ is a random variable with zero mean and variance 1 and pdf $p$ with support $(-\infty, \infty)$. Suppose that

$Y = \theta + E$

is observed.

(a) What is the Fisher Information Matrix when $E$ is normal distributed?

(b) Provide a lower bound for the Fisher Information Matrix that is valid for any twice differentiable $p$, with continuous second derivative.

(c) Comment on the results in (a) and (b). Compare with Problem 2.2.1.

2.3.5. Suppose that $Y_t$, $t = 1, \ldots, N$ are independent random variables with the distribution

$\frac{1}{2} \mathcal{N}(0, 1) + \frac{1}{2} \mathcal{N}(\theta, (e^{-\theta^2})^2)$

Plot the density function for $\theta = 0.6$. Plot the likelihood function for $N = 10, N = 30$ and $N = 100$ samples. What do you observe? How does ML work in this case?

This example is taken from Radford Neils blogg:

2.3.6. Let $Y$ be Poisson distributed $\text{Po}(\theta)$, i.e.

$\mathcal{P}(Y = y) = \frac{1}{y!} \theta^y e^{-\theta}$, $y = 0, 1, \ldots$

Use that $Y$ is a complete sufficient statistic to determine the UMVU estimator for $e^{-2\theta}$.

2.3.7. Suppose that $\hat{\theta}$ is UMVU for $\theta$. Show that $a\hat{\theta} + b$ is UMVU for $a\theta + b$.

2.3.8. * Suppose that $\hat{\alpha}$ is an efficient estimator for $\alpha \in \mathbb{R}^m$ and let the CRLB be $\Sigma$. Now suppose that we in addition know that

$Q\alpha = 0$

where $Q \in \mathbb{R}^{n \times m}$, $n < m$, is a known full rank matrix. Determine an efficient estimator for $\alpha$ in this case. What is the CRLB in this case?

Your solutions should be explicit expressions in $\Sigma$ and $Q$. 

2.4 Estimation vs Parameter Estimation

As we have seen, estimation and parameter estimation are closely linked. At this point it may be useful to review our findings. In Section 2.2, we studied the problem of estimating a random variable $X$ using another random variable $Y$, i.e., the estimator being of the form $\hat{X} = \tilde{X}(Y)$. Here, we found that the conditional mean $E[X|Y]$ is the optimal estimator (with the MSE as criterion). In comparison, in Section 2.3.6, we studied the problem of estimating a constant $\theta$ using a random variable $Y$. Here, we found that $E[\hat{\theta}(Y)|T(Y)]$ is the UMVU estimator of $\theta$ if $T(Y)$ is a complete sufficient statistic and $\hat{\theta}(Y)$ any unbiased estimator of $\theta$. The two problems are essentially the same with $X$ and $\theta$ playing the same role. In fact, we can see the parameter $\theta$ as a random variable with a point distribution. So why are the solutions different? This has to do with which information is available. In the parameter estimation problem, we could not condition on $Y$ but had to use a sufficient statistic to ensure that the result did not depend on the unknown $\theta$. To ensure uniqueness of the result, we further had to require $T$ to be complete. Furthermore, we could not use the conditional mean of $\theta$ since this would give the estimator $E[\theta|T(Y)] = \theta$ which is indeed exact but useless. Thus $\theta$ has to be replaced in the conditioning by the second best choice, namely an unbiased estimator of $\theta$.

In Section 2.3.2, we derived the Cramér-Rao Lower Bound based on projecting an unbiased estimate on the score function. The derivation was entirely based on estimation theory, so we may suspect that a similar bound should apply in estimation. As we have argued above, $X$ and $\theta$ should play the same role so let us analogously to (2.23) introduce the score function

$$S(X|Y) = \frac{\partial}{\partial X} \log p_{X,Y}(X,Y)$$

The reason for the notation is that

$$S(X|Y) = \frac{\partial}{\partial X} \log p_{X,Y}(X,Y) = \frac{\partial}{\partial X} \log p_{X|Y}(X|Y)p_Y(Y) = \frac{\partial}{\partial X} \log p_{X|Y}(X|Y)$$

Following the derivations in Section 2.3.2, it is easy to show that the score function has zero mean and that

$$\mathbb{E}[\hat{X}(Y)S(X|Y)^T] = I$$

for any unbiased estimator $\hat{X}(Y)$ of $X$. Thus, the projection of $X$ on the linear span of the elements of the score function and a constant is given by

$$I^{-1}_F(X|Y)S(X|Y) + \mathbb{E}[X]$$

where the Fisher Information Matrix is given by

$$I_F(X|Y) := \mathbb{E} \left[ S(X|Y)S^T(X|Y) \right] = -\mathbb{E} \left[ \frac{\partial}{\partial X} S(X|Y) \right]$$

with the expectation being over both $X$ and $Y$, and the CRLB (2.28) becomes

$$\text{Cov} \left[ X - \hat{X}(Y) \right] \geq I^{-1}_F(X|Y)$$

for any unbiased estimator $\hat{X}(Y)$ of $X$. This is known as the Bayesian Cramér-Rao Lower (BCRLB) bound.
For the parameter estimation problem, when an UMVU estimator $\hat{\theta}_{UMVU}$ exists, it must hold that
\[
\text{Cov} \left[ \hat{\theta} \right] \geq \text{Cov} \left[ \hat{\theta}_{UMVU} \right] \geq I^{-1}_F(\theta) \tag{2.58}
\]
for any unbiased estimator $\hat{\theta}$. The second inequality must hold since the CRLB holds for all unbiased parameter estimators. We thus see that when an UMVU estimator exists, the CRLB is valid and is reached by some estimator, the UMVU estimator must reach the CRLB.

We can re-express the inequalities (2.58) as (why?)
\[
\text{Cov} \left[ \theta - \hat{\theta} \right] \geq \text{Cov} \left[ \theta - \hat{\theta}_{UMVU} \right] \geq I^{-1}_F(\theta)
\]
for which the equivalent in estimation is given by
\[
\text{Cov} \left[ X - \hat{X} \right] \geq \text{Cov} \left[ X - \mathbb{E}[X|Y] \right] \geq I^{-1}_F(X|Y) \tag{2.59}
\]
Here we see that whenever the CRLB is reached by some estimator, this estimator must be the conditional mean. Now, we saw in Section 2.3.3 that the settings where a parameter estimator can reach the CRLB is very limited. The distribution has to be an exponential family (2.32) and even then an efficient estimator can only be obtained for particular parametrizations. It has to hold that the mean of the sufficient statistic $T(Y)$ is $\theta$, see (2.39), and then the estimator is the sufficient statistic itself, see (2.40). So how is the corresponding situation in the pure estimation case? Well, any estimator reaching the CRLB is given by the estimator corresponding to (2.26), i.e.
\[
I^{-1}_F(X|Y)S(X|Y) + \mathbb{E}[X] = I^{-1}_F \frac{\partial}{\partial X} \log p_{X|Y}(X|Y) + \mathbb{E}[X]
\]
But for this to be an estimator it cannot depend on $X$ and thus $I^{-1}_F(X|Y) \frac{\partial}{\partial X} \log p_{X|Y}(X|Y)$ must be independent of $X$, i.e. $\log p_{X|Y}(X|Y)$ has to be affine in $X$ and hence in the scalar case
\[
\log p_{X|Y}(x|y) = a(y) + b(y)x \quad \Rightarrow \quad p_{X|Y}(x|y) = e^{a(y)}e^{b(y)x}
\]
but this cannot be a density if $X$ has infinite support. Thus we have that the BCRLB cannot be reached when $X$ has infinite support. So why would we be interested in the BCRLB rather than the achievable bound given by the first inequality in (2.59)? One reason is that the BCRLB can often be easier to compute.

Before closing this section we remark that the FIM in the estimation can be expressed using $p_{X|Y}(x|y) = p_Y(y|x)p_X(x)/p_Y(y)$ as
\[
I_F(X|Y) = I_F(Y|X) + I_F(X|\emptyset)
\]
where
\[
S(X|\emptyset) := \frac{\partial}{\partial X} \log p_X(X)
\]
Notice that
\[
I_F(Y|X) = \mathbb{E}[I_F(X)|X = x]
\]
where $I_F(x)$ is the FIM in the parametric case (2.24). The FIM in the Bayesian case is thus the sum of the averages (over $X$) of the FIMs for estimating $X = x$ from $Y$ and from no information at all, respectively.

Finally, in the case of parameter estimation we saw in Section 2.3.4 that an estimator (2.56) solved $S(Y, \hat{\theta}(Y)) = 0$ and that it could be interpreted as a ML-estimator. In the pure estimation problem, an estimator of $X$ setting the score $S(X|Y)$ to zero can be given the interpretation as a Maximum A Posteriori (MAP) estimator since

$$S(\hat{X}(Y)|Y) = 0 \iff \frac{\partial}{\partial X} \log p_{X|Y}(X|Y)|_{X=\hat{\theta}(Y)} = 0$$

### 2.5 Using Estimation for Parameter Estimation

Clearly, parameter estimation means matching the model to the observations in some way. Many approaches can be seen as selecting the model that gives the best estimator of the observations, i.e. not only can we use estimation theory to analyze parameter estimators but we can also use it to construct such estimators.

Suppose that $Y^N \in \mathbb{R}^N$ is a random variable with the model represented by the pdf $p(y; \theta)$, $\theta \in \mathbb{R}^n$. The best estimator of $Y^N$ using constants is obtained by projecting the elements of $Y$ to the subspace spanned by the constant 1. This gives the estimator

$$\hat{Y}^N = \hat{Y}^N(\theta) = \int yp(y; \theta)dy$$

We now know that the optimal estimator is the true mean of $Y^N$, i.e.

$$E[\|Y^N - E[Y^N]\|^2] \leq E[\|Y^N - \hat{Y}^N(\theta)\|^2]$$

with equality only if $\hat{Y}^N(\theta) = E[Y^N]$, i.e. we have selected a model with the correct mean. Unfortunately, we cannot compute expectations, we only have an observation $Y^N$ at our disposal. However, we can form an unbiased estimate of $\frac{1}{N}E\|Y^N - \hat{Y}(\theta)\|^2$:

$$V_N(\theta) = \frac{1}{N}\|Y^N - \hat{Y}^N(\theta)\|^2$$

and we can now select as model estimate the model that minimizes this MSE estimate

$$\hat{\theta}_N := \arg \max_{\theta} V_N(\theta)$$

The formulation (2.53) of the BLUE problem is exactly of this type when the mean is linear in $\theta$. Notice that $\hat{\theta}_N$ is a function of the random variable $Y^N$ and therefore a random variable. The realization of $\hat{\theta}_N$ that corresponds to an observation $y^N$ of $Y^N$ is obtained by replacing $Y^N$ by $y^N$ in $V_N$ and solving the minimization problem above.

The mean seen as an estimator does not seem particularly good in this perspective, as the same constant is the optimal estimator for any distribution that has that constant as its expectation. We would expect more elaborate estimators, which are tailored to the particular distribution in question, to be more sensitive, and therefore more accurate. We can for example decide to estimate the means of different functions of $Y^N$, e.g. monomials of the elements of $Y^N$. This leads to a class of estimators called moment methods.
But we can also be more elaborate. In our setting above we could for example split $Y^N$ into two vectors $Y^N_1 \in \mathbb{R}^M$ and $Y^N_2 \in \mathbb{R}^{N-M}$ and then consider the OLE of $Y^N_1$ given $Y^N_2 \in \mathbb{R}^{N-M}$ and then pursue the approach above. Here a myriad of possibilities opens up, e.g., we can do different splits and consider the sum of estimated MSEs for the different splits. For example we can consider the OLE of $Y_k$ given all other observations, and do this for $k = 1, \ldots, N$. Being even more refined, we can consider the conditional mean estimator for the different splits.

Notice that the ML method corresponds to finding parameters that sets the sample estimate of the mean of the score function to zero, here the sample estimate is simply the score function itself. Thus the approach above does not have to use a cost function but can also be formulated as solving some equation. The theory of *estimating functions*, is a simplification of the ML approach where the score function is replaced by some other quantity that is simpler to compute.

Both with the approach outlined above and ML, the computation of the probability distributions in question can be computationally prohibitive (more on this later). Often, however, the used models are constructive in the sense that it is simple to generate samples from a given model. In this situation, *indirect inference* can be used. Here, the idea is to use a proxy model which is simpler than the model we are interested in, in the sense that the required probability distributions can be computed so that model estimates are simple to compute from data. Using simulated data from the original model having parameter $\theta$ then results in a proxy model which depends on $\theta$. $\theta$ is then tuned so that this proxy model is as close as possible in some sense to the proxy model obtained from the observations.

### 2.6 References

**Section 2.3.1**

Chapter 1, Sections 2.1, 2.5-2.6 in [10].

Section 2.3.9: Section 4.3 in [13] (Chapter 6 in [8]).
2.A Block Matrix Inversion

Let $\Delta_A = D - CA^{-1}B$ and $\Delta_D = A - BD^{-1}C$. Then

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B\Delta_A^{-1}CA^{-1} & -A^{-1}B\Delta_A^{-1} \\ -\Delta_A^{-1}CA^{-1} & \Delta_A^{-1} \end{bmatrix} = \begin{bmatrix} \Delta_D^{-1} & -\Delta_D^{-1}D^{-1}B \\ -D^{-1}C\Delta_D^{-1} & D^{-1} + D^{-1}C\Delta_D^{-1}BD^{-1} \end{bmatrix}$$

whenever the inverses exist.

2.A.1 Inverse of an Inner Product

As an application of the formulae above, let

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

and let $\theta_{Y_i|Y_j} = \langle Y_i, Y_j \rangle \langle Y_j, Y_j \rangle^{-1}$ be the coordinates for the projection of the elements of $Y_i$ on the linear span of $Y_j$. Then

$$\langle Y,Y \rangle^{-1} = \begin{bmatrix} -\theta_{Y_1|Y_2} \langle Y_{1\perp Y_2}, Y_{1\perp Y_2} \rangle^{-1} & -\theta_{Y_2|Y_1} \langle Y_{2\perp Y_1}, Y_{2\perp Y_1} \rangle^{-1} \\ \theta_{Y_1|Y_2} \langle Y_{1\perp Y_2}, Y_{1\perp Y_2} \rangle^{-1} & \theta_{Y_2|Y_1} \langle Y_{2\perp Y_1}, Y_{2\perp Y_1} \rangle^{-1} \end{bmatrix} = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

Thus $Y_{1\parallel Y_2} = -A^{-1}BY_2$ and the “norm” of the error $Y_1 - Y_{1\parallel Y_2}$ is $A^{-1}$. Similarly $Y_{2\parallel Y_1} = -BD^{-1}Y_1$ with “norm” of the error being $D^{-1}$. Notice that this holds regardless of how $Y$ is divided into $Y_1$ and $Y_2$.

2.B Schur Complement

Consider the symmetric block-matrix

$$Z = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

If $A > 0$ then

$$Z > 0 \iff C - B^TA^{-1}B > 0$$
$$Z \geq 0 \iff C - B^TA^{-1}B \geq 0$$

If $C > 0$ then

$$Z > 0 \iff A - BC^{-1}B^T > 0$$
$$Z \geq 0 \iff A - BC^{-1}B^T \geq 0$$

For

$$Z = \left\langle \begin{bmatrix} X \\ Y \end{bmatrix}, \begin{bmatrix} X \\ Y \end{bmatrix} \right\rangle$$

we get the following special case: When $\langle X, X \rangle > 0$, $Z$ is positive definite iff no linear combination of $X$ can be linearly predicted exactly by $Y$. Symmetrically, when $\langle Y, Y \rangle > 0$, $Z$ is positive definite iff no linear combination of $Y$ can be linearly predicted exactly by $X$. 

Chapter 3

Asymptotic Theory

3.1 Introduction

In the next chapter we will discuss how to reduce the risk using biased estimators. We give a simple example of this.

Example 3.1.1. Suppose that $Y^N \in \mathbb{R}^N$ consists of iid $\mathcal{N}(\theta, 1)$ random variables. The sample mean

$$\hat{\theta}_N := \frac{1}{N} \sum_{k=1}^{N} Y_k$$

is unbiased and attains the CRLB

$$\text{Cov} \left[ \hat{\theta}_N \right] = \frac{1}{N}$$

Now take $\lambda \in \mathbb{R}$ and define

$$\hat{\theta}_N(\lambda) := \lambda \hat{\theta}_N$$

which has mean and variance

$$\mathbb{E} \left[ \hat{\theta}_N(\lambda) \right] = \lambda \theta, \quad \text{Cov} \left[ \hat{\theta}_N(\lambda) \right] = \frac{\lambda^2}{N}$$

so that the MSE is given by

$$\text{MSE} \left[ \hat{\theta}_N(\lambda) \right] = (1 - \lambda)^2 \theta^2 + \frac{\lambda^2}{N}$$

which is minimized by

$$\lambda = \lambda_* := \frac{\theta^2}{\theta^2 + \frac{1}{N}}$$

We notice that we can not use this scaling in practice since it depends on $\theta$. However, we are here only interested in the properties of biased estimators when the MSE is minimized so we will ignore this practicality. Using $\lambda_*$ gives the MSE

$$\text{MSE} \left[ \hat{\theta}_N(\lambda_*) \right] = \frac{1}{N^2} \left( \frac{1}{\theta^2 + \frac{1}{N}} \right)^2 + \frac{1}{N} \left( \frac{\theta^2}{\theta^2 + \frac{1}{N}} \right)^2$$
which for large $N$ is well approximated by

$$\text{MSE} \left[ \hat{\theta}_N(\lambda^*_N) \right] \approx \frac{1}{N^2} \frac{1}{\theta^4} + \frac{1}{N} \left( 1 - \frac{1}{N\theta^2} \right)^2 \approx \frac{1}{N^2} \frac{1}{\theta^4} + \frac{1}{N} - \frac{2}{N^2 \theta^2} \approx \frac{1}{N}$$

We see that for large sample size $N$ it is the variance term that dominates even when we have chosen the estimator such that the MSE is minimized.

In Appendix 3.1 we show that the observation in Example 3.1.1, that for large sample size $N$ the variance term dominates over the bias term when the MSE is optimized, holds in general. This means that for large $N$, estimators, including biased ones, that achieve the CRLB are “optimal”. We make this statement more precise in the following definition.

**Definition 3.1.1.** Let $Y_1, Y_2, \ldots$ be a sequence of independent random variables with pdfs $p_k(y; \theta)$, $\theta \in \mathbb{R}^n$, having common support independent of $\theta$. Let $I_{F,k}(\theta)$ be the Fisher Information Matrix for $p_k(y; \theta)$ and suppose that

$$\frac{1}{N} \sum_{k=1}^{N} I_{F,k}(\theta) \to I_F(\theta)$$

for some $I_F(\theta)$. Let $\hat{\theta}_N$ be an estimator of $\theta$ based on $Y_1, \ldots, Y_N$. Then $\hat{\theta}_N$ is efficient in the limit if

$$N\text{MSE} \left[ \hat{\theta}_N \right] \to I_F^{-1}(\theta) \text{ as } N \to \infty$$

Efficiency in the limit is a weaker concept than efficiency (for every $N$). Recall that an efficient estimator is essentially unique, but for any sequence $\delta_N$ of random variables such that $N\mathbb{E} \left[ ||\delta_N||^2 \right] \to 0$ we will have that $\hat{\theta}_N + \delta_N$ is efficient in the limit if $\hat{\theta}_N$ is efficient in the limit.

There are a number of questions that arise when we now start to look at limits of estimators when the number of samples grows to infinity. Suppose that $\hat{\theta}_N$ is efficient in the limit. Can such an estimator be biased in the limit? Can we say anything about the behaviour of $\hat{\theta}_N$ for different sample points? To answer such questions we need some background theory.

### 3.2 Limits of Random Variables

#### 3.2.1 Convergence in Mean

We say that $\hat{\theta}_N$ converges to $\theta$ in mean square sense if

$$\mathbb{E} \left[ ||\hat{\theta}_N - \theta||^2 \right] \to 0 \quad \text{as } N \to \infty$$

This is the same as

$$\text{MSE} \left[ \hat{\theta}_N \right] \to 0 \quad \text{as } N \to \infty$$

In view of the bias-variance decomposition (2.1) where both terms are positive we see that mean square convergence implies that

$$\mathbb{E} \left[ \hat{\theta}_N \right] \to \theta \quad \text{as } N \to \infty$$

An estimator with this property is said to be unbiased in the limit.
3.2. LIMITS OF RANDOM VARIABLES

3.2.2 Convergence in Probability

Explicitly, convergence in mean square sense means that
\[ \int_{\Omega} \| \hat{\theta}_N(\omega) - \theta \|^2 dP(\omega) \to 0, \quad as \ N \to \infty \]

For a given \( \varepsilon > 0 \), we can split the integral into the set \( A(\varepsilon) := \{ \omega : \| \hat{\theta}_N(\omega) - \theta \|^2 \leq \varepsilon \} \) and its complement giving
\[
0 \leftarrow \int_A \| \hat{\theta}_N(\omega) - \theta \|^2 dP(\omega) + \int_{A^c} \| \hat{\theta}_N(\omega) - \theta \|^2 dP(\omega) \geq \varepsilon \int_{A^c} dP(\omega) = \varepsilon P(\| \hat{\theta}_N - \theta \|^2 > \varepsilon)
\]
Thus we see that for any \( \varepsilon > 0 \) it holds that
\[ P(\| \hat{\theta}_N - \theta \|^2 \geq \varepsilon) \to 0, \quad as \ N \to \infty \]

An estimator with this property is said to converge in probability to \( \theta \) and we write
\[ \hat{\theta}_N \overset{p}{\to} \theta \]

Above we have proved that convergence in mean square sense implies convergence in probability.

3.2.3 Convergence with Probability 1

As all random variables \( Y_1, Y_2, \ldots \) are defined on a common probability space \((\Omega, F, P)\) we can think of \( \hat{\theta}_N(\omega) \) as a sequence in \( N \) for every sample point \( \omega \in \Omega \). Since we will only draw one \( \omega \) it would be highly desirable to have estimators such that
\[ \hat{\theta}_N(\omega) \to \theta, \quad as \ N \to \infty \quad (3.1) \]

for every \( \omega \in \Omega \). However, our measure of the “size” of events in \( \Omega \) is given by the probability measure \( P \) and there are non-empty sets \( A \) in \( \Omega \) for which \( P(A) = 0 \). This means that we can at best prove that point-wise convergence (3.1) occurs on a set that has probability 1.

**Definition 3.2.1.** If there exists a set \( A \) with \( P(A) = 1 \) and (3.1) holds for all \( \omega \in A \) then we say that \( \hat{\theta}_N \) converges to \( \theta \) with probability one (w.p.1).

We write
\[ \hat{\theta}_N \overset{w.p.1}{\to} \theta \]

Alternative terminology is that \( \hat{\theta}_N \) converges to \( \theta \) almost everywhere (a.e.) or almost surely (a.s.).

The strong law of large numbers deal with almost sure convergence of series. A simple version is the following

**Lemma 3.2.1** (Corollary to Theorem 5.4.1 in [2]). Let \( \{Y_k\}_{k=1}^{\infty} \) be a sequence of independent zero mean random variables. Suppose that for some \( p, 1 \leq p \leq 2 \),
\[ \sum_{k=1}^{\infty} \frac{1}{k^p} \mathbb{E}[|Y_k|^p] < \infty \]

Then
\[ \frac{1}{N} \sum_{k=1}^{N} Y_k \overset{w.p.1}{\to} 0 \]
A simple case of when this lemma holds is when \( \mathbb{E}[|Y_k|^{1+\delta}] \) is bounded for some \( 0 < \delta \leq 1 \). The strong law of large numbers can also hold for dependent sequences.

**Lemma 3.2.2 ([12]).** Let \( \{Y_k\}_{k=1}^\infty \) be a sequence of zero mean random variables with finite second moments. Suppose that there exists constants \( \sigma > 0 \) and \( C < \infty \) such that for all integers \( N > M > 0 \)

\[
\mathbb{E}\left[ \left( \sum_{k=M+1}^{N} Y_k \right) \right] \leq C(N^\sigma - M^\sigma)
\]

Then for any \( \delta > 0 \)

\[
\left| \frac{1}{N} \sum_{k=1}^{N} Y_k \right| \leq CN^{\frac{1}{2}\sigma - 1} \log^{2+\delta} N \text{ w.p.} 1
\]

This lemma also gives a convergence rate from which we see that the smaller \( \sigma \) the faster the convergence. For an independent sequence we can use \( \sigma = 1 \) giving the rate \( \frac{\log^{2+\delta} N}{\sqrt{N}} \).

### 3.2.4 Convergence in Distribution

We say that \( X_N \to X \) in distribution (or in law) if

\[
\mathcal{P}(\{\omega : X_N(\omega) \in B\}) \to \mathcal{P}(\{\omega : X(\omega) \in B\}) \quad \text{as} \quad N \to \infty \quad \forall B \in \mathcal{B}
\]

Whenever \( X_N \) and \( X \) have distribution functions \( F_N(x) = \mathcal{P}(\{\omega : X_N(\omega) < x\}) \) convergence in distribution is equivalent to

\[
F_N(x) \to F(x) \quad \text{as} \quad N \to \infty \quad \forall x : F \text{ is continuous}
\]

We will write both

\[
X_N \overset{\text{dist}}{\to} X, \quad \text{and} \quad X_N \overset{\text{dist}}{\to} F
\]

where \( F \) denotes the distribution of \( X \).

A typical situation is that \( \sqrt{N}(\hat{\theta}_N - \theta) \) becomes asymptotically normal distributed with zero mean and some covariance matrix \( P \)

\[
\sqrt{N}(\hat{\theta}_N - \theta) \overset{\text{dist}}{\to} N(0, P)
\]

### 3.2.5 Relations Between Convergence Concepts

As we have seen convergence in mean square sense implies convergence in probability and the same is true for convergence w.p.1.

**Example 3.2.1.** Let the sample space \( \Omega \) be the unit circle and let the probability measure \( \mathcal{P} \) be the Lesbesgue measure on the circle normalized by \( 1/(2\pi) \). Suppose now that the estimator \( \hat{\theta}_N \) is such that

\[
A_N = \left\{ \omega : \|\hat{\theta}_N - \theta\| < \frac{1}{N} \right\}
\]
corresponds to the angular segment $[0, 2\pi(1 - 1/N)]$ counting in the standard counter clock-wise direction. Suppose that on $A_N$ it holds that 
\[ \|\hat{\theta}_N - \theta\| = N \]
Then clearly for every $\omega$, $\hat{\theta}_N \rightarrow \theta$ and convergence w.p.1 holds. However 
\[ \mathbb{E}\left[\|\hat{\theta}_N - \theta\|^2\right] \geq N^2\mathcal{P}(A_N) = N^2 \frac{1}{N} \rightarrow \infty \text{ as } N \rightarrow \infty \]
so convergence in mean does not hold.

The problem in the example with convergence in mean is that $\hat{\theta}_N$ becomes excessively large.

**Lemma 3.2.3.** Suppose that $X_N \rightarrow 0$ in probability and that there exists a random variable $Y$, $\mathbb{E}[|Y|^2] < \infty$, such that $|X_N| \leq Y$ a.e. Then $X_N \rightarrow 0$ in mean square sense.

**Example 3.2.2.** Consider the same probability space as in Example 3.2.1 but now assume that $A_N$ corresponds to the angular segment $[\alpha_N, \alpha_N + 2\pi(1 - 1/N)]$ where 
\[ \alpha_N = \sum_{k=1}^{N-1} \frac{1}{k} \]
The circle segment $A_N$ thus rotates around the unit circle growing simultaneously to cover the entire circle. However, since $\alpha_N \rightarrow \infty$ the rotation never stops which means that any given point $\omega$ on the circle will belong to $A_N$ for an infinite number of $N$'s, precluding point-wise convergence for any $\omega$. Thus while convergence in probability still holds, convergence w.p.1 does not hold in this case.

Convergence in probability implies convergence in distribution but the converse is not true. However

**Lemma 3.2.4.** Suppose that 
\[ X_N \overset{\text{dist}}{\rightarrow} X, \ Y_N \overset{\text{dist}}{\rightarrow} y, \ Z_N \overset{\text{dist}}{\rightarrow} z \]
where $y$ and $z$ are constants. Then 
\[ X_N Y_N + Z_N \overset{\text{dist}}{\rightarrow} Xy + z \]
Furthermore, $Z_N \overset{p}{\rightarrow} z$.

**Corollary 3.2.1.** Suppose that $Y_N \overset{p}{\rightarrow} Y$ and $y_N \rightarrow y$. Then 
\[ \mathcal{P}(Y_N \leq y_N) \rightarrow F_Y(y) \]
if $y$ is a continuity point of the cumulative distribution function $F_Y$ of $Y$.

**Theorem 3.2.1** ($\delta$-method, Gauss approximation method). Suppose that $\{X_N\}$ is sequence of random vectors such that
\[ \sqrt{N}(X_N - \theta) \overset{\text{dist}}{\rightarrow} N(0, P) \]
Let $q(x)$ be differentiable with $q'(\theta) \neq 0$. Then 
\[ \sqrt{N}(q(X_N) - q(\theta)) \overset{\text{dist}}{\rightarrow} N\left(0, q'(\theta)P(q'(\theta))^T\right) \]
Lemma 3.2.5 (Convergence of moments). Suppose that
\[ X_N \xrightarrow{d} X \]
and that
\[ \sup N \mathbb{E}[\|X_N\|^p] < \infty \]
Then for all \(0 < r < p\)
\[ \lim_{N \to \infty} \mathbb{E}[\|X_N\|^p] = \mathbb{E}[\|X\|^p] \]
and for all integers \(r, 0 < r < p\)
\[ \lim_{N \to \infty} \mathbb{E}[X_N^r] = \mathbb{E}[X^r] \]

3.3 Large Sample Properties of Estimators

3.3.1 Asymptotics of Explicit Estimators

Consistency

We will from now denote the parameter of the true underlying distribution by \(\theta_0\). If
\[ \hat{\theta}_N \xrightarrow{w.p.1} \theta_0 \]
we will say that \(\hat{\theta}_N\) is a consistent estimate of \(\theta_0\). In statistics it is common to use consistency for the weaker condition \(\hat{\theta}_N \xrightarrow{p} \theta_0\).

Sometimes there is an explicit expression for how an estimator depends on the observations \(Y^N = [Y_1 \ldots Y_N]^T\)
\[ \hat{\theta}_N = \hat{\theta}_N(Y^N) = f_N(Y^N) \]
Establishing consistency of such an estimator thus means verifying convergence w.p.1 for the sequence \(\{f(Y^N_N)\}\), i.e.
\[ f_N(Y^N) \xrightarrow{w.p.1} \theta_0 \]
which typically means using some strong law of large numbers
Example 3.3.1. Consider the model
\[ Y_k = \varphi_k^T \theta + E_k \]
where \{\varphi_k\} is a deterministic sequence where \{\|\varphi_k\|\} is bounded and
\[ R_N := \frac{1}{N} \sum_{k=1}^{N} \varphi_k \varphi_k^T \rightarrow R > 0 \quad (3.2) \]
and where \{E_k\} is iid with zero mean and variance \(\lambda < \infty\). The least squares estimator is given by
\[ \hat{\theta}_N = R_N^{-1} \frac{1}{N} \sum_{k=1}^{N} \varphi_k Y_k = \theta_o + R_N^{-1} \frac{1}{N} \sum_{k=1}^{N} \varphi_k E_k \quad (3.3) \]
and Lemma 3.2.1 gives
\[ \frac{1}{N} \sum_{k=1}^{N} \varphi_k E_k \xrightarrow{w.p.1} 0 \]
From this and (3.2) it follows that
\[ \hat{\theta}_N \xrightarrow{w.p.1} \theta_o \]

Convergence in Distribution

Example 3.3.2 (Continuation of Example 3.3.1). In Example 3.3.1 we have an explicit expression for the estimation error \(\hat{\theta}_N - \theta_o\). Then the standard central limit theorem (Theorem 6.4.4 in [2]) for iid sequences gives that
\[ \frac{1}{\sqrt{N}} \sum_{k=1}^{N} E_k \xrightarrow{\text{dist}} N(0, \lambda) \]
However in (3.3) we have instead the factor
\[ \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \varphi_k E_k \]
where, while being independent, \{\varphi_k E_k\} is not identically distributed. Therefore a stronger result is required to prove convergence in distribution. In the case of independent terms, as we have here,
\[ \mathbb{E} \left[ \left( \sum_{k=1}^{N} \varphi_k E_k \right) \left( \sum_{k=1}^{N} \varphi_k E_k \right)^T \right] = \lambda \sum_{k=1}^{N} \varphi_k \varphi_k^T = \lambda N R_N \]
and one can use the Lindberg-Feller theorem (Theorem 7.2.1 in [2]) to prove that
\[ \frac{1}{\sqrt{\lambda N}} R_N^{-1/2} \sum_{k=1}^{N} \varphi_k E_k \xrightarrow{\text{dist}} N(0, I) \]
when \(\mathbb{E}[|E_k|^{2+\delta}] \leq M < \infty\) for all \(k\) (and under the assumption in Example 3.3.1 that \(\|\varphi_k\|\) is bounded). Lemma 3.2.4 then gives
\[ \sqrt{N} (\hat{\theta}_N - \theta_o) \xrightarrow{\text{dist}} N(0, R^{-1}) \]
The path followed in the preceding example is typical for proving convergence in distribution when there is an explicit expression for \(\hat{\theta}_N\).
3.3.2 Criterion Based Estimators

We saw in Section 2.5 that estimators can be solutions to optimization problems

$$\hat{\theta}_N = \arg \min_{\theta \in D_\theta} V_N(\theta)$$  \hspace{1cm} (3.4)

where the cost function $V_N(\theta)$ is constructed from the observations, $Y^N = [Y_1 \ldots Y_N]^T$ say.

Consistency

Notice that for a fix $\theta$, $\{V_N(\theta)\}$ is a sequence of random variables. Under suitable regularity conditions on the data generation mechanism, i.e. the sequence of pdfs $p(y^N; \theta)$, it may hold that this sequence will converge

$$V_N(\theta) \xrightarrow{w.p.1} V(\theta)$$

It may seem intuitive that $\{\hat{\theta}_N\}$ will converge to the minimizer of $V(\theta)$. For this, however, additional conditions are required.

**Theorem 3.3.1.** Suppose that

$$\sup_{\theta \in D_\theta} |V_N(\theta) - V(\theta)| \xrightarrow{w.p.1} 0$$  \hspace{1cm} (3.5)

holds uniformly on the compact set $\Theta$. Suppose that the limit function in (3.5), $V(\theta)$, is continuous in an open set containing $\Theta$ and let $\hat{\theta}_N$ be any solution to (3.4). Then

$$\hat{\theta}_N \xrightarrow{w.p.1} D_c = \{ \theta : V(\theta) \leq V(\xi) \forall \xi \in D_\theta \}$$  \hspace{1cm} (3.6)

Under the conditions of Theorem 3.3.1, consistency hinges on if $D_c$ is a singleton containing the true parameter $\theta_0$. Typically, the estimator is constructed such that the limit $V(\theta)$ is minimized by the true parameter $\theta_0$. However, $D_c$ may contain other parameters as well depending on the criterion that is chosen. $D_c$ also typically depend on the experimental conditions (which in fact determine the functional form of the pdf $p(y; \theta)$).

**Example 3.3.3.** Suppose $Y_k, k = 1, 2, \ldots$ is a sequence of independent random variables with mean $\theta$ and variance $\theta^2$.

To estimate $\theta$ we could use the best constant predictor of $Y^2_k$ which is the mean $\theta^2 + \theta^4$. This leads to the criterion

$$V_N(\theta) = \frac{1}{N} \sum_{k=1}^N |Y_k^2 - (\theta^2 + \theta^4)|^2$$

where $\{y_k\}$ are our observations of $\{Y_k\}$. Clearly if $\theta_s$ is a solution to (3.4) then so is $-\theta_s$ and it follow that the same will be true for minima of the limit $V(\theta)$.

Alternatively, we can use the best constant predictor of $Y_k$ itself

$$V_N(\theta) = \frac{1}{N} \sum_{k=1}^N |Y_k - \theta|^2$$  \hspace{1cm} (3.7)

By the law of large numbers, $V_N(\theta) \to V(\theta) := \mathbb{E}[|Y_k - \theta|^2]$ which has the true parameter $\theta_0$ as unique minimum.
### 3.3. LARGE SAMPLE PROPERTIES OF ESTIMATORS

#### Convergence in Distribution

Suppose that $\hat{\theta}_N$ is uniquely defined by (3.4) and convergence to $\theta_*$ (typically $\theta_*=\theta_o$, the true parameter). Then

$$0 = V_N'(\hat{\theta}_N) = V_N'(\theta_*) + V_N''(\xi_N)(\hat{\theta}_N - \theta_*)$$

for some $\xi_N$ between $\hat{\theta}_N$ and $\theta_*$, giving that

$$\hat{\theta}_N - \theta_* = - (V_N''(\xi_N))^{-1} V_N'(\theta_*)$$

Since $\hat{\theta}_N$ is consistent it follows that $\xi_N \xrightarrow{w.p.1} \theta_*$. Furthermore, assuming that $V_N''(\theta) \xrightarrow{w.p.1} V''(\theta)$ we will have

$$\hat{\theta}_N - \theta_* \approx - (V''(\theta_*))^{-1} V_N'(\theta_*) \tag{3.8}$$

A typical situation is that $V_N(\theta)$ is an unbiased estimate of $V(\theta)$ which implies that $V_N'(\theta)$ is an unbiased estimate of $V'(\theta)$. Thus since $V'(\theta_*) = 0$, $E[V_N'(\theta_*)] = 0$. However, often even more can be said about $V_N'(\theta_*)$.

**Example 3.3.4.** $V_N(\theta)$ given by (3.7) is an unbiased estimate of $V(\theta) = E[|Y_k - \theta|^2]$ and

$$V_N'(\theta) = \frac{2}{N} \sum_{k=1}^{N} (Y_k - \theta)$$

which clearly has zero mean for $\theta = \theta_o$. Furthermore

$$\sqrt{N}V_N'(\theta_o) = 2\theta \frac{1}{\sqrt{N}} \sum_{k=1}^{N} (Y_k - \theta) \tag{3.9}$$

has variance

$$N E \left[ V_N'(\theta_o)^2 \right] = \frac{4 \theta_o^2}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} E[(Y_k - \theta_o)(Y_l - \theta_o)]$$

$$= \frac{4 \theta_o^2}{N} \sum_{k=1}^{N} E[(Y_k - \theta_o)^2]$$

$$= 4 \theta_o^6$$

Since the sum in (3.9) consists of iid random variables with zero mean, the central limit theorem applies (provided some regularity conditions on the pdf $p(y; \theta)$ applies) giving

$$\sqrt{N}V_N'(\theta_o) \xrightarrow{\text{dist}} N(0, 4 \theta_o^6)$$

As in the example, often

$$N E \left[ V'_N(\theta_*) (V_N'(\theta_*))^T \right] \rightarrow P(\theta_*) \text{ as } N \rightarrow \infty \tag{3.10}$$
and
\[ \sqrt{N} V_N'(\theta_o) \overset{\text{dist}}{\to} N(0, P(\theta_*)) \]
for some matrices \( \tilde{P}(\theta_o) \) and \( P(\theta_*). \) Assuming this and returning to (3.8) we obtain
\[ N \mathbb{E}\left[ (\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T \right] \to (V''(\theta_*))^{-1} \tilde{P}(\theta_*) (V''(\theta_*))^{-1} \] (3.11)
and
\[ \sqrt{N}(\hat{\theta}_N - \theta_0) \overset{\text{dist}}{\to} N \left( 0, (V''(\theta_*))^{-1} P(\theta_*) (V''(\theta_*))^{-1} \right) \] (3.12)
In Example 3.3.4, \( \tilde{P}(\theta_0) = P(\theta_0). \) However, this is not always the case.

**Lemma 3.3.1** (Lemma 6.1.14 in [10]). Suppose that \( Y_k \overset{\text{dist}}{\to} Y \) where \( \mathbb{E}[Y] = 0 \) and \( \mathbb{E}[Y^2] < \infty. \) Then
\[ \mathbb{E}[Y^2] \leq \liminf_k \mathbb{E}[Y_k^2] \]
Convergence in distribution of the parameter estimator to some distribution \( H, \)
\[ \sqrt{N}(\hat{\theta}_N - \theta_0) \overset{\text{dist}}{\to} H \] (3.13)
can also hold for estimators constructed in other ways than as minimizers of some cost function. We make the following remarks:

- The result (3.13) implies that
  \[ \hat{\theta}_N \overset{p}{\to} \theta_0 \]
  This follows from the last part of Lemma 3.2.4.

- \( \hat{\theta}_N \) does not have to be unbiased in the limit, i.e. \( \mathbb{E}[\hat{\theta}_N] \) does not have to converge to \( \theta_0. \)
  However, if \( \sup_N N^{1/2(1+\delta)} \mathbb{E}\left[ |\hat{\theta}_N - \theta_0|^{1+\delta} \right] < \infty \) for some \( \delta > 0, \) Lemma 3.2.5 gives the stronger
  \[ \lim_{N \to \infty} \sqrt{N} \mathbb{E}[\hat{\theta}_N - \theta_0] = 0 \]
  As we noted earlier, for an estimator efficient in the limit, \( \lim_{N \to \infty} \mathbb{E}[\hat{\theta}_N] = \theta_0. \)

- The limit distribution \( H \) does not necessarily have to have zero mean, see, e.g., Example 6.2.3 in [10].

- The variance of \( H \) is called the asymptotic variance of \( \hat{\theta}_N. \)

- By Lemma 3.3.1, if \( H \) has zero mean and variance \( \Sigma_H, \)
  \[ \Sigma_H \leq \liminf_N \text{NMSE}\left[ \hat{\theta}_N \right] \]
3.3. LARGE SAMPLE PROPERTIES OF ESTIMATORS

3.3.3 Asymptotic Efficiency

There is a CRLB for the asymptotic variance as well. Here stated for a scalar $\theta$.

**Theorem 3.3.2** (Theorem 6.2.6 in [10]). Let $Y_1, Y_2, \ldots$ be iid with pdf $p(y; \theta)$, $\theta \in D_\theta \subset \mathbb{R}$ with $\Theta$ being an open interval. Assume that

1. The pdf $p(y; \theta)$ has support $A = \{y : p(y; \theta) > 0\}$ independent of $\theta$.
2. For all $x \in A$, $p$ is twice differentiable with respect to $\theta$, and the second derivative is continuous in $\theta$.
3. The integral $\int p(y; \theta)dy$ can be twice differentiated under the integral sign.
4. The FIM $I_F(\theta)$ satisfies $0 < I_F(\theta) < \infty$.
5. For any $\theta_o \in D_\theta$, there exists positive number $c(\theta_o)$ and a function $M(x; \theta_o)$ such that

$$|\partial^2 \log p(y; \theta)/\partial \theta^2| \leq M(x), \quad \forall x \in A, \theta_o - c < \theta < \theta_o + c$$

and

$$\infty > E_{\theta_o}[M(Y_1; \theta_o)] := \int p(y; \theta_o)M(y; \theta_o)dy$$

Suppose that $\hat{\theta}_N$ is an estimator for which (3.13) holds where $H = N(0, \Sigma(\theta_o))$. Then

$$\Sigma(\theta) \geq I_F^{-1}(\theta_o)$$  (3.14)

except on a set of Lesbesgue measure zero.

The theorem extends to the multiparameter case. Estimators for which (3.13) holds where $H = N(0, \Sigma(\theta_o))$ are said to be asymptotically efficient with respect to $\theta$ if $\Sigma(\theta) = I_F^{-1}(\theta)$.

**Corollary 3.3.1.** Suppose that $g(\theta) : \Theta \to \mathbb{R}^m$ is a differentiable function with $g'(\theta_o) \neq 0$. Let $\hat{\lambda}_N$ be an estimator of $g(\theta)$ such that

$$\sqrt{N} (\hat{\lambda}_N - g(\theta)) \overset{\text{dist}}{\to} N(0, \Lambda(\theta))$$

then

$$\Lambda(\theta) \geq g'(\theta_o)I_F^{-1}(\theta_o) (g'(\theta_o))^T$$  (3.15)

Furthermore, if $\hat{\theta}_N$ is asymptotically efficient for $\theta$, then $g(\hat{\theta}_N)$ is asymptotically efficient with respect to $g(\theta)$.

The last part of the corollary follows from Theorem 3.2.1.

For estimators that are asymptotically normal distributed, Lemma 3.3.1 gives that, under the conditions of Theorem 3.3.2, a necessary condition for efficiency in the limit is that the asymptotic distribution reaches the CRLB (3.14).
3.3.4 Superefficient Estimators

Estimators violating (3.14) (on a set of Lesbegue measure zero) are known as superefficient. However, this type of estimator is known to have poor finite sample properties.

Example 3.3.5 (Examples 6.2.5 and 6.2.7 in [10], due to Hodges). Consider the set-up in Example 3.1.1. Let \( \bar{Y}_N := \frac{1}{N} \sum_{k=1}^{N} Y_k \) and take

\[
\hat{\theta}_N = \begin{cases} 
  \bar{Y}_N & |\bar{Y}_N| \geq N^{1/4} \\
  a\bar{Y}_N & |\bar{Y}_N| < N^{1/4}
\end{cases}
\]

Then (3.13) holds with \( H = N(0, v(\theta)) \) where

\[
v(\theta) = \begin{cases} 
  1 & \theta \neq 0 \\
  a^2 & \theta = 0
\end{cases}
\]

and hence \( \hat{\theta}_N \) violates (3.14) at \( \theta = 0 \) if \( |a| < 1 \).

It can also be shown that

\[
\text{NMSE} \left( \hat{\theta}_N \right) = \begin{cases} 
  1 & \theta \neq 0 \\
  a^2 & \theta = 0
\end{cases}
\]

However, the downside with this estimator is that it can be shown that

\[
\sup_{\theta} \text{NMSE} \left( \hat{\theta}_N \right) \rightarrow \infty
\]

In particular, for each \( N \) setting the true \( \theta \) to \( \theta_N := N^{-1/4} \) will result in that \( \text{NMSE} \left( \hat{\theta}_N \right) \rightarrow \infty \) as \( N \rightarrow \infty \).

3.3.5 The Maximum Likelihood Estimator

As noted the ML estimator is efficient only for the exponential family (2.32) satisfying (2.38) and (2.39). Let us now study its large sample properties.

Below we will assume that \( Y_1, Y_2, \ldots \) are iid each with pdf \( p(y; \theta) \).

Consistency

The ML estimator is also of the form (3.4) with (we use the minus sign to turn maximization of the log-likelihood function into minimization of the negative log-likelihood)

\[
V_N(\theta) := -\frac{1}{N} \sum_{k=1}^{N} \log p(Y_k; \theta)
\]

Since \( \{Y_k\} \) are independent and assuming that

\[
\mathbb{E} [\log p(Y_1; \theta)] < \infty, \quad \mathbb{E} [\log p(Y_1; \theta)] < \infty
\]

we can use the strong law of large numbers (Theorem 5.4.2 in [2]) and obtain

\[
V_N(\theta) \overset{\text{w.p.1}}{\rightarrow} V(\theta) := -\mathbb{E} [\log p(Y_1; \theta)]
\]
which is known as the Shannon entropy for $Y_1$ when $\theta = \theta_o$. Let us introduce

$$D_N(\theta|\theta_o) = V_{N}^{ML}(\theta) - V_{N}^{ML}(\theta_o) = -\frac{1}{N} \sum_{k=1}^{N} \log \left( \frac{p(Y_k; \theta)}{p(Y_k; \theta_o)} \right)$$

(3.19)

$$D_N(\theta|\theta_o) \overset{w.p.1}{\rightarrow} D(\theta|\theta_o) := -E \left[ \log \left( \frac{p(Y_1; \theta)}{p(Y_1; \theta_o)} \right) \right]$$

(3.20)

where $D(\theta|\theta_o)$ is called the Kullback-Leibler divergence. Using Jensen’s inequality

$$-D(\theta|\theta_o) = \mathbb{E} \left[ \log \left( \frac{p(Y_1; \theta)}{p(Y_1; \theta_o)} \right) \right] \leq \log \left( \mathbb{E} \left[ \frac{p(Y_1; \theta)}{p(Y_1; \theta_o)} \right] \right) = \log 1 = 0$$

and hence

$$D(\theta|\theta_o) \geq 0, \quad \text{with equality iff } p(y; \theta) = p(y; \theta_o) \text{ a.e.}$$

This means that only parameters in the set

$$D_c = \{ \theta : p(y; \theta) = p(y; \theta_o) \text{ almost everywhere} \}$$

maximizes the limit function. However, as already discussed, this does not mean that the maximum-likelihood estimate $\hat{\theta}_{ML,N}$ w.p.1 $\rightarrow D_c$. There are a number of counter-examples such as Example 6.4.1 in [10]. However, under the conditions in Theorem 3.3.1 convergence will take place.

Relaxed conditions for consistency can be obtained by considering solutions to the likelihood equations

$$S(Y^N; \theta) = 0$$

Under some technical conditions, it can be shown that there exist a sequence of solutions that is consistent even if $\Theta$ is not compact, but an open possibly unbounded set, see Theorem 6.5.1 in [10]. When there are multiple roots, it is in general not possible to know which ones will lead to a sequence that is consistent.

**Asymptotic Efficiency**

Let us now assume that $\hat{\theta}_{ML,N}$ is consistent and let us denote the Fisher Information Matrix for one of the r.v. $Y_k$ by $I_F(\theta)$. We will follow the approach in (3.8) and the ensuing derivations to analyze the asymptotic distribution of the ML-estimator. Using that $Y_1, Y_2, \ldots$ are iid gives

$$V'_{N}(\theta) = \frac{1}{N} S(Y^N; \theta) = \frac{1}{N} \sum_{k=1}^{N} S(Y_k; \theta)$$

(3.21)

and, again referring to the strong law of large numbers,

$$V''_{N}(\theta) = \frac{1}{N} \frac{\partial}{\partial \theta} S(Y^N; \theta) = \frac{1}{N} \sum_{k=1}^{N} \frac{\partial}{\partial \theta} S(Y_k; \theta) \overset{w.p.1}{\rightarrow} V''(\theta) := \mathbb{E} \left[ \frac{\partial}{\partial \theta} S(Y_k; \theta) \right]$$

(3.22)
but according to (2.25) the right-hand expression is \(-I_F(\theta)\). Thus (3.8) translates into
\[
\hat{\theta}_{ML,N} \approx I_F^{-1}(\theta_o) \frac{1}{N} S(Y^N; \theta_o) + \theta_o = (NI_F(\theta_o))^{-1} S(Y^N; \theta_o) + \theta_o
\]
but recognizing that \(NI_F(\theta)\) is the Fisher matrix for \(Y^N\) we see that the right-hand side is the expression for an efficient estimator. This hints at that when the ML estimator is consistent it is typically also asymptotically efficient. In fact,
\[
\mathbb{E}[V'_N(\theta_o)(V'_N(\theta_o))^T] = I_F(\theta)
\]
so \(\hat{P}(\theta_o)\) in (3.10) is given by
\[
\hat{P}(\theta_o) = I_F(\theta_o)
\]
and (3.22) gives that \(V''(\theta_o) = I_F(\theta_o)\) as well. Hence (3.12) gives
\[
\sqrt{N}(\hat{\theta}_{ML,N} - \theta_o) \xrightarrow{\text{dist.}} N(0, I_F^{-1}(\theta_o))
\]
i.e. the ML estimator is asymptotically efficient (under certain technical assumptions).

It is worth reflecting a little bit on this result. Recall that there were very few distributions for which the ML-estimator could be expressed as the efficient estimator
\[
(NI_F(\theta_o))^{-1} S(Y^N; \theta_o) + \theta_o
\]
which generally depends on the unknown \(\theta_o\). So how is it possible that this can be achieved with large data sets? The answer is that with large data sets we can use the observed data to estimate the properties that we need in (3.25) to build in an efficient estimator. This is done in an indirect way in the ML-estimator but is nevertheless the case. For example we see that the second derivative of the cost-function is the factor \(I_F(\theta_o)\) used in (3.25). This insight also gives a hint on how large data set is required for the ML estimator to be close to efficient: When the above quantities can be accurately estimated.

However, this reflection also suggests another way to use the observations to construct an estimator resembling (3.25). Suppose that a consistent, but not asymptotically efficient, estimator \(\tilde{\theta}_N\) is available. Then \(\hat{\theta}_N \approx \theta_o\) for \(N\) sufficiently large and \(S(Y^N; \tilde{\theta}_N) \approx S(Y^N; \theta_o)\) as well as \(I_F(\hat{\theta}_N) \approx I_F(\theta_o)\) leading to the estimator
\[
\hat{\theta}_N = (NI_F(\tilde{\theta}_N))^{-1} S(Y^N; \tilde{\theta}_N) + \tilde{\theta}_N
\]
Alternatively making use of \(\frac{\partial}{\partial \theta} S(Y^N; \tilde{\theta}_N) \approx -NI_F(\theta_o)\) leads to the estimator
\[
\hat{\theta}_N = -(\frac{\partial}{\partial \theta} S(Y^N; \tilde{\theta}_N))^{-1}(\theta_o)S(Y^N; \tilde{\theta}_N) + \tilde{\theta}_N
\]
However, it is not sufficient that \(\hat{\theta}_N\) is consistent for these two estimators to be asymptotically efficient, the convergence to \(\theta\) has to be fast enough. To specify the exact requirements we need to introduce a few concepts. A sequence \(\{X_N\}\) is bounded in probability if for any \(\varepsilon > 0\) there exist \(C(\varepsilon)\) and \(N(\varepsilon)\) such that
\[
\mathcal{P}(|X_N| \geq C(\varepsilon)) < \varepsilon, \quad N \geq N(\varepsilon) < \infty
\]
3.4. PARAMETER TRANSFORMATIONS

This is a weaker concept than that $X_N$ converges in distribution and hence also implied if $X_N$ converges in mean square sense to 0. Furthermore, $\{X_N\}$ is $O_p(f(N))$ if $\{X_N/f(N)\}$ is bounded in probability.

What we need for (3.26) and (3.27) to be asymptotically efficient is that $\tilde{\theta}_N$ is $\sqrt{N}$-consistent, that is that $\tilde{\theta}_N - \theta_0$ is $O_p(1/\sqrt{N})$. For example any estimator for which

$$\lim_{N \to \infty} N \mathbb{E} \left[ \left| \tilde{\theta}_N - \theta \right|^2 \right] \to 0$$

is $\sqrt{N}$-consistent.

As with many other estimators, the ML-estimate has to be found through numerical optimization. Typically gradient based descent methods are used. The estimator (3.27) can be given an interpretation from this perspective: It is one step in the Newton-Raphson method for minimizing the likelihood function starting from the point $\tilde{\theta}_N$. We will return to how to numerically compute estimates later.

3.4 Parameter Transformations

3.4.1 Consistent Estimators

In Section 2.3.8 we saw that if $\hat{\theta}_{ML}$ is the ML-estimator of $\theta$ and $\alpha = q(\theta)$, then $q(\hat{\theta}_{ML})$ is the ML-estimator of $\alpha$. More generally, assume that $\tilde{\theta}_N$ is a consistent estimator of $\theta$. Then, $q(\tilde{\theta})$ is a consistent estimator of $\alpha$. Furthermore, suppose that

$$\sqrt{N}(\tilde{\theta}_N - \theta) \xrightarrow{\text{dist.}} N(0, P)$$

and that $q'(\theta) \neq 0$. Then Theorem 3.2.1 gives that

$$\sqrt{N}(q(\tilde{\theta}_N) - \alpha) \xrightarrow{\text{dist.}} N(0, q'(\theta)P(q'(\theta))^T)$$

(3.28)

In particular, if $\tilde{\theta}_N$ is asymptotically efficient, then it follows from (3.28) and Corollary 3.3.1 that $q(\tilde{\theta}_N)$ is asymptotically efficient for $\alpha$. From an engineering point of view this is a very appealing result.

Example 3.4.1. Suppose that a feedback controller is to be designed for a system for which a parametric model is available. Let $\theta$ be the model parameter, with $\theta_0$ denoting its true value. The designed controller will be a function of $\theta$ and hence the resulting signals $s$ in the feedback system will depend on both the used model parameter $\theta$ and $\theta_0$ (since the controller is applied to the true system): $s = s(\theta_0, \theta)$. Typically the control design is such that using $\theta = \theta_0$ gives the desired behaviour of the signals, i.e. $s(\theta_0, \theta_0)$ is the ideal closed loop response. The previous result tells us that using the ML-estimate of $\theta$ in the control design will minimize the deviation from this objective in the sense that $\mathbb{E} \left[ \| s(\theta_0, \hat{\theta}_{ML}) - s(\theta_0, \theta_0) \| \right]$ will be minimized.

In summary, under appropriate regularity assumptions:

*Using the ML-estimate in place of the true parameter leads to the best asymptotic performance regardless of the model use.*

Compare this with the situation for efficient estimators where, e.g., for an exponential family there is only one of all possible parametrizations that have parameters that can be efficiently estimated, and where functions of efficient estimators typically are not efficient estimators of their means.
3.4.2 The Extended Invariance Principle (EXIP*)

Now we will return to the situation in Section 2.3.8 under the paragraph Efficient Estimators. That is, as in Section 3.4.1 we will assume that \( q \) maps \( D_\theta \in \mathbb{R}^n \) onto \( D_\alpha \in \mathbb{R}^m \). Suppose that \( \hat{\alpha}_N \) is consistent, and denoting the true parameter by \( \alpha_o \), we can use any consistent estimate of the limit covariance

\[
\sqrt{N}(\hat{\alpha}_N - \alpha_o) \xrightarrow{\text{dist.}} N(0, \Sigma(\alpha_o))
\]  

(3.29)

In Remark 2.3.2 we highlighted that when \( q \) was a linear map, a suitable projection of \( \hat{\alpha}_N \) onto the column span of \( q \) results in an efficient estimator of \( \theta \) if \( \hat{\alpha}_N \) is efficient with respect to \( \alpha \). Furthermore, it follows from Remark 2.3.3 that in this case this estimator of \( \theta \) can be obtained from

\[
\hat{\theta}_N = \arg \min_{\theta \in D_\theta} V_N(\theta), \text{ where }
\]

\[
V_N(\theta) = \frac{1}{N}(\hat{\alpha}_N - q(\theta))^T \Sigma_N^{-1}(\hat{\alpha}_N - q(\theta))
\]  

(3.30)

where \( \Sigma_N \) is the covariance matrix of \( \hat{\alpha}_N \).

We could of course use (3.30) also when \( q \) is a nonlinear function. So what properties will this estimate have? Well, firstly since \( \hat{\alpha}_N \) is consistent and assuming \( N\Sigma_N \rightarrow \Sigma(\alpha_o) \), it follows that \( \hat{\theta}_N \) will be consistent provided \( D_\theta \) is compact. The requirement on \( \Sigma_N \) can be relaxed by noting that we can use any consistent estimate of the limit covariance \( \Sigma(\alpha_o) \), e.g. \( \Sigma(\hat{\alpha}_N) \).

When it comes to the asymptotic distribution, we observe that throughout Section 3.3 a recurring theme has been that when an estimator is consistent, it is sufficient to study the local behaviour around the limit point of the involved quantities. We will follow this path here as well. Denoting the true \( \theta \) by \( \theta_o \) and \( Q := q'(\theta_o) \),

\[
q(\theta) \approx q(\theta_o) + q'(\theta_o)(\theta - \theta_o) = \alpha_o + Q(\theta - \theta_o)
\]

for \( \theta \) in the domain where \( \hat{\theta} \) will be obtained. Thus \( V_N \) in (3.30) can be written

\[
V_N(\theta) \approx \frac{1}{N}(\hat{\alpha}_N - \alpha_o - Q(\theta - \theta_o))^T \Sigma_N^{-1}(\hat{\alpha}_N - \alpha_o - Q(\theta - \theta_o))
\]

This means that (3.30) is fitting the linear in the parameters (seeing \( \theta - \theta_o \) as the parameter instead of \( \theta \) makes no difference) model \( Q(\theta - \theta_o) \) to the “observation” \( \hat{\alpha}_N - \alpha \) which has zero mean and variance \( \Sigma(\alpha_o) \). From (2.52) we obtain

\[
\hat{\theta} - \theta_o \approx (Q^T \Sigma^{-1}(\alpha_o)Q)^{-1} Q^T \Sigma^{-1}(\alpha_o)(\hat{\alpha}_N - \alpha)
\]

with covariance matrix

\[
NCov \left[ \hat{\theta} - \theta_o \right] \approx (Q^T \Sigma^{-1}(\alpha_o)Q)^{-1}
\]

Formally, one can show that

\[
\sqrt{N}(\hat{\theta}_N - \theta_o) \xrightarrow{\text{dist.}} N \left( 0, (Q^T \Sigma^{-1}(\alpha_o)Q)^{-1} \right)
\]  

(3.31)

Using (3.28) we get that the estimate of \( \alpha \) using that \( \alpha = q(\theta) \) has asymptotic distribution

\[
\sqrt{N}(q(\hat{\theta}_N) - \alpha_o) \xrightarrow{\text{dist.}} N \left( 0, Q(Q^T \Sigma^{-1}(\alpha_o)Q)^{-1} Q^T \right)
\]  

(3.32)
3.4. PARAMETER TRANSFORMATIONS

The difference between the covariance matrices in (3.29) and (3.32) is a measure of how much information is gained by knowing that $\alpha = q(\theta)$ asymptotically:

$$\Sigma(\alpha_o) - Q \left( Q^T \Sigma^{-1}(\alpha_o) Q \right)^{-1} Q^T$$  \hspace{1cm} (3.33)

Introducing $X = \Sigma^{1/2}(\alpha_o)$ and $Y = Q^T \Sigma^{-1/2}(\alpha_o)$ and using the “inner product” $\langle X, Y^T \rangle = XY^T$ we can write (3.33) as

$$\langle X, X^T \rangle - \langle X, Y \rangle \langle Y, Y^T \rangle \langle Y, X^T \rangle$$  \hspace{1cm} (3.34)

which comparing with (2.16) means that we can interpret the difference (3.33) as the squared “norm” of what is left when we project the rows of $X = \Sigma^{1/2}(\alpha_o)$ orthogonally on the rows of $Y = Q^T \Sigma^{-1/2}(\alpha_o)$. For example, the diagonal consists of the squared norms of what is left of the rows of $X$ when their projections are removed. This means that in the directions of the rows of $Q^T \Sigma^{-1}(\alpha_o)$, nothing is gained, i.e. the difference (3.33) is zero. This is quite natural as this is the subspace of $\mathbb{R}^m$ which the method based on $\theta$ targets. It is in the complement of this subspace where we have better accuracy using $\theta$ as this method asymptotically does not give any estimate there, and hence the variance of this estimate is zero in this subspace.

Suppose now that $\hat{\alpha}_N$ is obtained by minimizing some criterion $\tilde{V}_N(\alpha)$

$$\hat{\alpha}_N = \arg \min_{\alpha \in D_\alpha} \tilde{V}_N(\alpha)$$  \hspace{1cm} (3.35)

and let us assume that the equivalent to (3.12) holds so that

$$\Sigma(\alpha_o) = (\tilde{V}''(\alpha_o))^{-1} \tilde{P}(\alpha_o)(\tilde{V}''(\alpha_o))^{-1}$$  \hspace{1cm} (3.36)

where

$$\tilde{P}(\alpha) = \lim_{N \to \infty} N \mathbb{E} \left[ \tilde{V}_N'(\alpha)(\tilde{V}_N'(\alpha))^T \right]$$  \hspace{1cm} (3.37)

The asymptotic covariance matrix for $\hat{\theta}_N$ can then be written

$$\tilde{\Sigma}_\theta := \left( Q^T \tilde{V}''(\alpha_o) \tilde{P}^{-1}(\alpha_o) \tilde{V}''(\alpha_o) Q \right)^{-1}$$

In this situation a direct approach of estimating $\theta$ would be to minimize $\tilde{V}_N(\theta) := \tilde{V}_N(q(\theta))$ leading to the asymptotic covariance matrix

$$\tilde{\Sigma}_\theta := (\tilde{V}''(\theta_o))^{-1} \tilde{P}(\theta_o)(\tilde{V}''(\theta_o))^{-1}$$  \hspace{1cm} (3.38)

where

$$\tilde{P}(\theta) = \lim_{N \to \infty} N \mathbb{E} \left[ \tilde{V}_N'(\theta)(\tilde{V}_N'(\theta))^T \right]$$  \hspace{1cm} (3.39)

Now

$$\tilde{V}_N'(\theta) = (q'(\theta))^T \tilde{V}_N'(q(\theta)), \quad \tilde{V}_N''(\theta) = (q''(\theta))^T \tilde{V}_N''(q(\theta)) q'(\theta) + \Delta_N(\theta)$$

where $\Delta_N(\theta)$ is linear in the elements of $\tilde{V}_N'(q(\theta))$ and therefore has zero mean. Thus

$$\tilde{P}(\theta_o) = Q^T \tilde{P}(\theta_o) Q, \quad \tilde{V}''(\theta_o) = Q^T \tilde{V}''(\theta_o) Q$$
and hence

$$\tilde{\Sigma}_\theta := \left( Q^T \tilde{V}''(\theta_o) Q \right)^{-1} Q^T \tilde{P}(\theta_o) Q \left( Q^T \tilde{V}''(\theta_o) Q \right)^{-1} \quad (3.40)$$

To compare $\tilde{\Sigma}_\theta$ and $\Sigma_\theta$, let us consider the difference between their inverses. Dropping arguments gives

$$\tilde{\Sigma}_\theta^{-1} - \Sigma_\theta^{-1} = Q^T \tilde{V}'' \tilde{P}^{-1} \tilde{V}'' Q - Q^T \tilde{V}'' Q \left( Q^T \tilde{P} Q \right)^{-1} Q^T \tilde{V}'' Q$$

$$= Q^T \tilde{V}'' \left( \tilde{P}^{-1} - Q \left( Q^T \tilde{P} Q \right)^{-1} Q^T \right) \tilde{V}'' Q$$

Here we recognize that $Q \left( Q^T \tilde{P} Q \right)^{-1} Q^T$ is the projection of the rows of $\tilde{P}^{-1/2}$ on the rows of $Q^T \tilde{P}^{1/2}$, and hence $\tilde{P}^{-1} - Q \left( Q^T \tilde{P} Q \right)^{-1} Q^T \geq 0$. We can also see this by Schur complement (Section 2.B): The difference is positive semidefinite iff

$$0 \geq \begin{bmatrix} \tilde{P}^{-1} & Q \\ Q^T & Q^T \tilde{P} Q \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & Q^T \end{bmatrix} \begin{bmatrix} \tilde{P}^{-1} & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I & 0 \end{bmatrix}^T$$

but again by Schur complement

$$\begin{bmatrix} \tilde{P}^{-1} & I \\ I & \tilde{P} \end{bmatrix} \geq 0 \iff 0 \geq \tilde{P}^{-1} - I(\tilde{P})^{-1} I = 0$$

Thus we have the, perhaps, surprising result that directly estimating $\theta$, using the natural criterion $J_N(\theta) := \tilde{V}_N(q(\theta))$ gives worse accuracy asymptotically than going through the detour of first estimating $\alpha$ and then $\theta$ through (3.30). Only in the case that $\tilde{V}''(\alpha_o) = \tilde{P}(\alpha_o)$ will the two methods give the same asymptotic accuracy. So how can this be? To answer this, let us first return to the ML-estimator that we studied in Section (3.3.5). We showed there that the ML-estimator is asymptotically efficient, i.e. the asymptotic covariance matrix is equal to the Fisher Information Matrix. In the context here, with $\hat{\alpha}_N$, defined in (3.35), being the ML-estimate of $\alpha$, the reason for the asymptotic efficiency is that both $\tilde{P}(\theta_o)$, defined in (3.37), and $\tilde{V}''(\theta_o)$ are equal to the Fisher Information Matrix. Now when $\tilde{V}_N(\alpha)$ is the likelihood criterion for $\alpha$, clearly $\tilde{V}_N(q(\theta))$ is the likelihood criterion for $\theta$, and hence the direct approach yields the Fisher Information Matrix as asymptotic covariance matrix. Thus the two methods are asymptotically the same when the direct approach is asymptotically efficient. The indirect approach can then be seen as an approximate two-step ML-approach. In the first step, an ML-estimate of $\alpha$ is obtained. Asymptotically, the distribution of this estimate is normal and the second step (3.30) corresponds to an approximate ML-estimation of $\theta$ from $\hat{\alpha}$. However, when $\tilde{V}_N$ is not the ML-criterion (or an asymptotic equivalent), then $\tilde{P}(\theta_o)$ and $\tilde{V}''(\theta_o)$ are not necessarily equal and then the direct approach is not asymptotically efficient. However, the statistic $\hat{\alpha}_N(Y^N)$ contains more information than the direct estimate as the data compression from $Y^N$ to $\hat{\alpha}_N$ is less than from $Y^N$ to the direct estimate of $\theta$ since the search for the direct estimate takes place over $D_\alpha \subset \tilde{D}_\alpha$. Thus using $\hat{\alpha}_N$ to obtain an estimate of $\theta$ in the (asymptotically) statistically optimal way (3.30) yields an estimate with higher accuracy than the direct estimate.

Before concluding we notice that if instead of using (3.30), we use

$$\hat{\theta}_N = \arg \min_{\theta \in D_\theta} V_N(\theta), \text{ where}$$

$$V_N(\theta) = \frac{1}{N} (\hat{\alpha}_N - q(\theta))^T V''(\hat{\alpha}_N)(\hat{\alpha}_N - q(\theta))$$

(3.41)
the two step approach gives the same asymptotic covariance (3.40) as the direct estimate. The approach (3.41) is known as EXIP - the Extended Invariance Principle [15].

### 3.4.3 Beyond the EXIP*

Besides being of theoretical interest in terms of how information propagates through estimates, EXIP and its siblings can sometimes be used in practice to simplify computations of an estimate. Splitting the computations into two steps may lead to considerable simplifications, see, e.g., [15]. However, computationally it may be even more advantageous to let $\tilde{D}_\alpha$ outside $D_\alpha$. We illustrate with a simple example.

**Example 3.4.2.** Let the model be

$$y_t = \frac{\theta}{1-\theta q^{-1}} u_t + e_t, \ u_t = 0, \ t \leq 0$$  

(3.42)

where $\{e_t\}$ is iid Gaussian $N(0,1)$ noise and where $D_\theta \subset (-1, 1)$ so that we know that the model is stable. We have that

$$Y^N = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = m_N(\theta) + E,$$

where

$$m_n(\theta) = \Phi_{N,n} \begin{bmatrix} \theta \\ \theta^2 \\ \vdots \\ \theta^N \end{bmatrix}, \ E = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_N \end{bmatrix}$$

where $\Phi_{N,n}$ is the lower triangular Toeplitz matrix

$$\Phi_{N,n} := \begin{bmatrix} u_1 & 0 & 0 & \ldots & 0 \\ u_2 & u_1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ u_N & u_{N-1} & \ldots & \ldots & u_{N-n+1} \end{bmatrix}$$

so we see that the mean $m(\theta)$ of $Y^N$ is a nonlinear function of $\theta$ and hence the ML-estimator which minimizes

$$V_N(\theta) = \frac{1}{N} \| Y^N - m(\theta) \|^2$$

must be found through numerical nonlinear optimization.

The set of models corresponds to $D_\theta$ in the parameter space but we can also think of the model set as a set in the space of stable impulse responses (which is an infinite dimensional space)

$$D_\alpha = \{ g := \{g_k\}_{k=0}^\infty : g_k = \theta^k, \ k = 0, 1, \ldots, \ \theta \in D_\theta \}$$

Now we may consider the model approximation

$$Y^N = \Phi_{N,n}\alpha + E$$  

(3.43)
where \( \alpha = [\alpha_1 \ldots \alpha_n]^T \) is a truncation of the impulse response without imposing the structure that \( \alpha_k = \theta^k \) for some \( \theta \). In the space of stable impulse responses, this model set can be written as

\[
\tilde{D}_{\alpha,n} = \left\{ g := (g_k)_{k=0}^{\infty} : g_k = \begin{cases} \alpha_k, & k = 0, 1, \ldots, n - 1 \\ 0, & k \geq n \end{cases} \right\}
\]

The attractive feature with this model set is that the ML-estimate of \( \alpha \) in the model (3.43) can be obtained through least-squares

\[
\hat{\alpha}_{N,n} = (\Phi_{N,n}^T \Phi_{N,n})^{-1} \Phi_{N,n}^T Y_N \quad \text{(3.44)}
\]

Following EXIP we would then estimate \( \theta \) by minimizing

\[
\left\| \hat{\alpha}_{N,n} - \begin{bmatrix} \theta \\ \theta^2 \\ \vdots \\ \theta^{n-1} \end{bmatrix} \right\|
\]

with some suitable weighting. However, this is still a non-convex optimization problem so not much is gained here using EXIP.

Instead, notice that with \( g_k(\theta) = \theta^{k+1}, \ k = 0, 1, \ldots \) being the impulse response coefficients of the model (3.42), and \( G(q, \theta) \) being the transfer function \( G(q) = \sum_{k=0}^{\infty} g_k(\theta)q^{-k} \), we have

\[
G(q, \theta) = \frac{\theta}{1 - \theta q^{-1}}
\]

This can be written as

\[
\theta = \sum_{k=0}^{\infty} (1 - \theta q^{-1})g_k(\theta)q^{-k} = g_0 + (g_1 + \theta g_0)q^{-1} + (g_2 + \theta g_1)q^{-1} + \ldots
\]

resulting in one linear equation in \( \theta \) per time lag:

\[
\begin{bmatrix} 1 \\ -g_0 \\ -g_1 \\ \vdots \end{bmatrix} \theta = \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ \vdots \end{bmatrix}
\quad \text{(3.45)}
\]

Thus it is simple to compute \( \theta \) if the model’s impulse response is given. This suggests a two step approach to estimate \( \theta \):

i) Estimate \( \alpha \) using least-squares (3.44). The resulting \( \hat{\alpha}_{N,n} \) is an estimate of the first \( n \) coefficients of the impulse response.

ii) Solve

\[
\begin{bmatrix} 1 \\ -\hat{\alpha}_0 \\ -\hat{\alpha}_1 \\ \vdots \\ -\hat{\alpha}_{n-1} \end{bmatrix} \theta = \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_{n-2} \end{bmatrix}
\quad \text{(3.46)}
\]

\(^1\)We will call such models Finite Impulse Response (FIR) models.
where \( \hat{\alpha}_1, \ldots, \hat{\alpha}_n \) are the elements of \( \hat{\alpha}_{N,n} \), in a least squares sense to obtain an estimate \( \hat{\theta} \) of \( \theta \).

Instead of solving a non-convex optimization problem, we now only have to solve two least-squares problems. The approach can be refined by taking the statistical properties of \( \hat{\alpha}_N \) into account by way of weighting in Step ii).

While the two step procedure in Example 3.4.2 leads to attractive computational properties, we have disregarded an important aspect, namely that \( D_\alpha \) and \( \tilde{D}_\alpha,n \) are disjoint. This is radically different from the situation we studied in the EXIP section when \( D_\alpha \subset \tilde{D}_\alpha \). This means typically that \( \hat{\alpha}_N \) will be a biased of \( \alpha \). In the example, \( \hat{\alpha}_{N,n} \) will be a biased estimate of the first \( n \) impulse response coefficients. Thus we will in the second step try to fit a model in \( D_\alpha \) to “data” that does not belong to this model set.

The solution to this dilemma is to let the model set for which it is simple to compute an estimate, \( \tilde{D}_\alpha \) grow with the sample size. That is we will let \( \tilde{D}_\alpha = \tilde{D}_\alpha(N) \) such that

\[
\inf_{\alpha \in \tilde{D}_\alpha(N)} \| \alpha - \alpha_o \| \to 0, \text{ as } N \to \infty \quad \forall \alpha_o \in D_\alpha
\]

In the example we can achieve this by forcing the model order \( n \) to increase with the sample size \( N \) so that \( n = n(N) \to \infty \) as \( N \to \infty \). Then \( \tilde{D}_\alpha(N) := \tilde{D}_{\alpha,n(N)} \) has the desired property, i.e. all impulse responses in \( D_\alpha \) can be approximated arbitrarily well as the sample size grows.

The rate of \( n(N) \) must be such that the bias vanishes quickly enough so that it is the variance error that dominates in \( \hat{\alpha}_N \). This gives a lower bound on the rate of increase of \( n(N) \). In the example, it may therefore be tempting to take, e.g., \( n(N) = N \). However this leads to a large least-squares problem in Step i). Furthermore, \( \hat{\alpha}_{N,n(n)} \) will then not converge in distribution and it may be difficult to determine how Step ii) should be formulated. This also holds in general. On the contrary, if the rate of \( n(N) \) is upper bounded in a suitable way, \( \hat{\alpha}_N \) will converge in distribution and the analysis and algorithm construction will be simplified significantly. This type of method is called multi-step least-squares [5].

### 3.5 Using Estimation for Parameter Estimation

In Section 2.5 we observed that BLUE fits into the framework of using an estimator of the observations to estimate the parameter vector \( \theta \) underlying the pdf. The BLUE covers the case where the mean of the observations is linear in \( \theta \):

\[
\mathbb{E}[Y] = \Phi_N\theta, \quad Y \in \mathbb{R}^N, \quad \Phi_N \in \mathbb{R}^{N \times n}
\]

In Remark 2.3.3 we found the BLUE could be found in the class of estimators generated by minimizing the cost function

\[
V_N(\theta, W_N) = \frac{1}{N}(Y - \hat{Y}(\theta))^T W_N (Y - \hat{Y}(\theta)), \quad W_N^T = W_N > 0
\]

Choosing \( W_N = \text{Cov}[Y^N] \) yields the BLUE.

We are now ready to pursue this approach in a more general setting. We will still use the simple estimator \( \hat{Y}(\theta) := \mathbb{E}[Y] \) but we will not constrain ourselves to the case where \( \hat{Y}(\theta) \) is linear in \( \theta \). We will not require that the elements of \( Y \) are iid either. However, we will require some regularity conditions for our observations:
We will make the assumption that $V_N(\theta, I_N) \xrightarrow{w.p.1} V(\theta)$ and that $V(\theta)$ is minimized at the true $\theta = \theta_o \in D_\theta$ where $D_\theta$ is a known compact set.

Following the analysis in Section 4.5.10, it follows that with $W_N = I_N$ we will have consistency. It thus makes sense to restrict the sequence of weighting matrices $\{W_N\}_{N=1}^\infty$ to those for which consistency is achieved and in particular we will assume that

$$W_N = \Sigma_N^{-1}(\theta_o) := \text{Cov}^{-1}\{Y\}$$

belongs to this class.

In the following we will be interested in how the choice of $W_N$ affects the asymptotic covariance. Let $\Phi_N = Y'(\theta) \in \mathbb{R}^{N \times n}$. Then

$$V''_N(\theta_o) = -\frac{1}{N} (\Phi_N^T W_N (Y - EY))$$

$$= \frac{1}{N} \Phi_N^T W_N \Phi_N$$

Thus the asymptotic covariance of $\hat{\theta}_N$, given in (3.12), is

$$\Sigma_\theta := (V''(\theta_o))^{-1} P(\theta_o) (V''(\theta_o))^{-1}$$

where

$$P(\theta_o) = \lim_{N \to \infty} N \mathbb{E} \left[ V'_N(\theta_o) (V'_N(\theta_o))^T \right]$$

$$= \lim_{N \to \infty} \frac{1}{N} \Phi_N^T W_N \Sigma_N W_N \Phi_N$$

$$V''(\theta_o) = \lim_{N \to \infty} \mathbb{E} \left[ V''_N(\theta_o) \right] = \lim_{N \to \infty} \frac{1}{N} \Phi_N^T W_N \Phi_N$$

We thus will assume that the pdfs of $Y$ and $\{W_N\}$ are such that these limits exist. We notice that without loss of generality we can take $W_N \geq I_N$ as we can scale $W_N$ without changing the minimum of $V_N$. From Section 2.3.9 we may suspect that $W_N = \Sigma_N^{-1}$ will be optimal. This choice gives

$$\Sigma_\theta = \Sigma_\theta^* := \lim_{N \to \infty} \left( \frac{1}{N} \Phi_N^T \Sigma_N^{-1} \Phi_N \right)^{-1}$$

so let us compare the inverse of the left-hand side of this expression with the corresponding inverse for an arbitrary $W_N$

$$\Phi_N^T \Sigma_N^{-1} \Phi_N - \Phi_N^T W_N \Phi_N \Phi_N^T W_N \Sigma_N W_N \Phi_N \Phi_N^T W_N \Phi_N$$

This difference is positive definite iff

$$0 \geq \begin{bmatrix} \Phi_N^T \Sigma_N^{-1} \Phi_N & \Phi_N^T W_N \Phi_N \\ \Phi_N^T W_N \Phi_N & \Phi_N^T W_N \Sigma_N W_N \Phi_N \end{bmatrix} = \begin{bmatrix} \Phi^T & 0 \\ 0 & \Phi^T \end{bmatrix} \begin{bmatrix} \Sigma_N^{-1} & W_N \\ W_N & W_N \Sigma_N^{-1} \end{bmatrix} \begin{bmatrix} \Phi \\ 0 \end{bmatrix}$$

$$\Leftrightarrow$$

$$0 \leq \Sigma_N^{-1} - W_N (W_N \Sigma_N W_N)^{-1} W_N = 0$$

Thus if $\Sigma_\theta^*$ exists, $\Sigma_\theta - \Sigma_\theta^* \geq 0$ for any sequence of $\{W_N\}$ of weighting matrices for which $\Sigma_\theta$ exists. Thus it is asymptotically optimal to use the inverse covariance of the observations as weighting. We dig deeper into this later on in the course.
3.6. REFERENCES

3.5.1 Exercises

3.5.1. Let \( Y_1, Y_2, \ldots \) be iid \( N(0, \lambda) \), where \( \lambda = 1 \).

(a) What is the CRLB for \( \lambda \) given \( N \) measurements?

(b) What is an efficient estimator for \( \lambda \)?

(c) Construct a consistent estimator of \( \lambda \) using only fourth moments \( Y_k^4 \) of the observations.

(d) Derive the asymptotic distribution for your estimator in (c). Does it reach the CRLB?

(e) Propose an estimator based on (3.26) and your estimator in (c).

(f) Perform a Monte Carlo study for the three estimators considered above, with sample sizes \( N = 10, 100, 1000, 10000 \) and \( m = 500 \) simulations for each \( N \). For each \( N \), plot the histograms of \( \sqrt{N}(\hat{\lambda} - \lambda) \). For which \( N \) does the sample distribution of the estimator in (c) correspond to the asymptotic distribution you derived in (d). Also compute the normalized (by \( N \)) sample variances of the estimates. Does it seem like the estimator in (e) is asymptotically efficient? How does it compare with the efficient estimator?

3.5.2. * Let \( Y_1, \ldots, Y_N \) be iid \( N(\theta, a^2) \) where \( a \) is known.

(a) Compute the CRLB. Compare this with the CRLB when \( Y_1, \ldots, Y_N \) are \( N(\theta, \lambda) \) with \( \lambda \) known and equal to \( a^2 \) (a). How much does it seem to be possible to gain by knowing \( a \)?

(b) Derive explicit expressions for the ML-estimators for the two cases in (a).

(c) Perform a Monte Carlo simulation study for the two ML estimators where you study the sample properties for increasing sample size \( N \). Over a large number \( M \) of realizations, compute the sample means, the sample variance and the sample MSE of the estimates and compare the results with the respective CRLBs.

3.6 References

Section 3.2

Section 1.8 in [10], Sections 4.1–4.2, 4.5, 5.1*–5.4* in [2].

Section 3.3

Chapter 6 in [10].
CHAPTER 3. ASYMPTOTIC THEORY

3.A  Asymptotic Dominance of the Variance Term

3.A.1  MSE Dominating the CRLB

Here we will show that the variance error dominates the MSE for large sample sizes if the MSE is minimized. For simplicity we will consider the scalar case $\theta \in \mathbb{R}$ and we will assume that the elements of $Y^N \in \mathbb{R}^N$ are iid with pdf $p(y; \theta)$. Then the Fisher information for $Y^N \in \mathbb{R}^N$ is $NI_F(\theta)$ where $I_F(\theta)$ is the Fisher information for one of the elements in $Y^N$.

Let $b(\theta)$ be the bias of the estimator $\hat{\theta}$ i.e. $E[\hat{\theta}] = \theta + b(\theta)$. Then (2.30) gives

$$\text{MSE}[\hat{\theta}] = b^2(\theta) + \text{Cov} [\hat{\theta}] \geq b^2(\theta) + \frac{(1 + b'(\theta))^2}{NI_F(\theta)} \quad (3.48)$$

We will study the conditions on the bias when this lower bound is smaller than the CRLB $1/(NI_F(\theta))$

$$b^2(\theta) + \frac{(1 + b'(\theta))^2}{NI_F(\theta)} \leq \frac{1}{NI_F(\theta)} \quad (3.49)$$

for an interval $\theta \in [a, c]$ and when the sample size $N$ grows. We can write (3.49) as

$$NI_F(\theta)b^2(\theta) + 2b'(\theta) + \left(b'(\theta)\right)^2 \leq 0$$

or

$$b'(\theta) \leq -\frac{NI_F(\theta)}{2}b^2(\theta) - \frac{1}{2}\left(b'(\theta)\right)^2 \quad (3.50)$$

We start by observing that this inequality cannot always be satisfied. Solving for equality gives

$$b'(\theta) = -1 \pm \sqrt{1 - NI_F(\theta)b^2(\theta)}$$

and thus

$$b^2(\theta) \leq \frac{1}{NI_F(\theta)} \quad (3.51)$$

is a necessary and sufficient condition. From this inequality we see that the bias term in the MSE is upper bounded by the variance term. Below we will show that (3.50) imposes even stronger conditions on the bias term. However, before this let us note that (3.51) implies that the derivative $b'(\theta)$ cannot be small over a large interval. To see this, define first

$$\gamma = \inf_{\theta \in [a, c]} I_F(\theta)$$

Then

$$b^2(\theta) \leq \frac{1}{N\gamma} \quad (3.52)$$

is a sufficient condition for (3.51). Now, suppose that if $b'(\theta) \leq -1/2$ on a subset of $[a, c]$ of measure $l$, then

$$b(a) - b(c) \geq l/2$$
3.A. ASYMPTOTIC DOMINANCE OF THE VARIANCE TERM

since \( b'(\theta) \leq 0 \) on the entire interval. In view of (3.52) this implies that

\[
\frac{l}{2} \leq \frac{1}{\sqrt{N\gamma}} - \frac{1}{\sqrt{N\gamma}}
\]

is an upper bound on \( l \) for it to be possible for (3.51) to hold. Thus, the variance term \( \text{Cov} \left[ \hat{\theta} \right] \) is of the order \( O(1/N) \) except on a set that tends to measure zero as \( N \to \infty \).

Now we will study the bias term \( b^2(\theta) \). Consider the relaxation

\[
b'(\theta) \leq -\frac{N\gamma}{2} b^2(\theta)
\]

of (3.50). Clearly if \( b(\theta) \) satisfies (3.50) it satisfies (3.53). Now let \( \tilde{b}(\theta) \) be the solution to

\[
\tilde{b}'(\theta) = -\frac{N\gamma}{2} \tilde{b}^2(\theta)
\]

Then any \( b(\theta) \) that satisfies (3.53) is subject to \( b(\theta) \leq \tilde{b}(\theta) \). Hence any \( b(\theta) \) that satisfies (3.50) satisfies \( b(\theta) \leq \tilde{b}(\theta) \) as well.

Clearly \( \tilde{b} \) is a decreasing function and (3.54) has a continuous solution over \([a, c]\) given by

\[
\tilde{b}(\theta) = \frac{b(a)}{1 + \frac{N\gamma}{2} b(a)(\theta - a)} \quad (3.55)
\]

when \( b(a) > 0 \), and

\[
\tilde{b}(\theta) = \frac{b(c)}{1 + \frac{N\gamma}{2} b(c)(\theta - a)} \quad (3.56)
\]

when \( b(c) < 0 \).

Clearly \( \tilde{b}(\theta) \) is decreasing and suppose now that for a given value \( b(a) > 0 \), \( b^2(a) \leq 1/(N\gamma) \) at the left end-point of the interval, we want to examine when \( \tilde{b}^2(\theta) \) has decreased below \( 1/(N\gamma)^{1+\delta} \) for some \( \delta > 0 \). From (3.55) we get the condition

\[
\frac{b^2(a)}{\left(1 + \frac{N\gamma}{2} b(a)(\theta - a)\right)^2} = \frac{1}{(N\gamma)^{1+\delta}}
\]

giving

\[
(N\gamma)^{\frac{1}{2} + \frac{\delta}{2}} b(a) = 1 + \frac{N\gamma}{2} b(a)(\theta - a)
\]

or

\[
\theta - a = \frac{(N\gamma)^{\frac{1}{2} + \frac{\delta}{2}} b(a) - 1}{N\gamma b(a)} \leq 2(N\gamma)^{-\frac{1}{2} + \frac{\delta}{2}}
\]

Thus if \( 0 < \delta < 1/2 \) and \( b(a) > 0 \),

\[
\tilde{b}^2(\theta) \leq \frac{1}{(N\gamma)^{1+\delta}} \quad \forall \theta \geq a + aN, \; 0 \leq aN \leq 2(N\gamma)^{-\frac{1}{2} + \frac{\delta}{2}} \to 0, \; \text{as} \; N \to \infty
\]
Similarly we can show that if \(0 < \delta < 1/2\) and \(b(b) < 0\),
\[
\bar{b}^2(\theta) \leq \frac{1}{(N\gamma)^{1+\delta}} \quad \forall \theta \leq c - c_N, \quad 0 \leq c_N \leq 2(N\gamma)^{-\frac{1}{2} + \frac{\delta}{2}} \to 0, \quad \text{as } N \to \infty
\]

Consider now that \(b(\theta)\) satisfies (3.50) on an interval \([a, c]\). Since \(b'(\theta) \leq 0\) must hold, it follows that there are three cases to consider:

i) \(b(a) \geq 0\) and \(b(c) \geq 0\). Since \(b(\theta) \leq \tilde{b}(\theta)\), the above gives that
\[
b^2(\theta) \leq \frac{1}{(N\gamma)^{1+\delta}}
\]

except on an interval with \(a\) as left end-point and where the length of the interval tends to zero as \(N \to \infty\).

ii) \(b(a) < 0\) and \(b(c) < 0\). We know that \(b(\theta) \leq \tilde{b}(\theta)\) but that \(\tilde{b}^2(\theta)\) will increase from an given value larger than \(1/(N\gamma)^{1+\delta}\) to \(1/(N\gamma)\) over an interval that tends to zero as \(N \to \infty\). Thus for \(b(\theta)\) to be feasible, i.e. to satisfy \(b^2(\theta) \leq 1/(NI_F(\theta))\), \(b^2(\theta)\) must be less than \(1/(N\gamma)^{1+\delta}\), except, possibly, on an interval with \(c\) as right end-point where the length of this interval tends to zero as \(N \to \infty\).

iii) \(b(a) > 0\) and \(b(c) < 0\). We can split the interval into two subintervals \([a, d]\) and \([d, c]\), defined by \(b(d) = 0\). Then we use i) for \([a, d]\) and ii) for \([d, c]\) arriving at similar conclusions as in these two cases.

The above shows that an estimator for which the lower bound on the MSE is no larger than the unbiased CRLB, for large sample sizes \(N\) must have a bias term in the MSE that is \(O(N^{1+\delta})\) for some \(0 < \delta < 1/2\) and thus this term is dominated by the variance term \((1 + b'(\theta))^2/(NI_F(\theta))\), which we above showed is of size \(O(1/N)\).

### 3.A.2 Optimal MSE

Here we consider the problem of minimizing the integral of the lower bound (2.30) of an interval \([a, c]\). As in the previous section we consider the scalar case \(\theta \in \mathbb{R}\)

\[
L(b(\theta), b'(\theta), \theta) = \int_a^c b^2(\theta) + \frac{(1 + b'(\theta))^2}{NI_F(\theta)} d\theta
\]

We will restrict ourselves to the case \(I_F(\theta) = 1\). The Euler-Lagrange equation

\[
\frac{\partial}{\partial b} L = \frac{d}{d\theta} \frac{\partial}{\partial b'} L
\]

becomes

\[
2b(\theta) = \frac{d}{d\theta} \frac{2}{N} (1 + b'(\theta)) = \frac{2}{N} b''(\theta)
\]

which has solution

\[
b(\theta) = C_1 e^{-\sqrt{N}\theta} + C_2 e^{\sqrt{N}\theta}
\]
3.A. ASYMPTOTIC DOMINANCE OF THE VARIANCE TERM

For given end-point conditions \( b(a) \) and \( b(c) \) we can determine \( C_1 \) and \( C_2 \):

\[
\begin{bmatrix}
e^{-\sqrt{N}a} & e^{\sqrt{N}a} \\
e^{-\sqrt{N}b} & e^{\sqrt{N}b}
\end{bmatrix}
\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} =
\begin{bmatrix} b(a) \\ b(b) \end{bmatrix}
\Rightarrow
\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} =
\frac{1}{e^{\sqrt{N}(b-a)} - e^{-\sqrt{N}(b-a)}}
\begin{bmatrix} e^{\sqrt{N}b} & -e^{\sqrt{N}a} \\ -e^{-\sqrt{N}b} & e^{-\sqrt{N}a} \end{bmatrix}
\begin{bmatrix} b(a) \\ b(b) \end{bmatrix}
\]

so that

\[
b(\theta) = \frac{1}{e^{\sqrt{N}(b-a)} - e^{-\sqrt{N}(b-a)}} \begin{bmatrix} e^{-\sqrt{N}\theta} & e^{\sqrt{N}\theta} \\ -e^{\sqrt{N}\theta} & e^{-\sqrt{N}\theta} \end{bmatrix}
\begin{bmatrix} e^{\sqrt{N}(b-\theta)} & -e^{\sqrt{N}(b-\theta)} \\ -e^{-\sqrt{N}(b-\theta)} & e^{-\sqrt{N}(b-\theta)} \end{bmatrix}
\begin{bmatrix} b(a) \\ b(b) \end{bmatrix}
\]

\[
= \frac{1}{1 - e^{-2\sqrt{N}(b-a)}} \begin{bmatrix} e^{-\sqrt{N}(\theta-a)} & e^{-\sqrt{N}((b-\theta)+(b-a))} \\ e^{-\sqrt{N}(b-\theta)} & e^{-\sqrt{N}((\theta-a)+(b-a))} \end{bmatrix}
\begin{bmatrix} b(a) \\ b(b) \end{bmatrix}
\]

from which we see that \( b(\theta) \) decays as \( e^{-\sqrt{N}l} \) where \( l \) is the minimum distance to the end-points.
Chapter 4

Biased Estimators

4.1 The Bias-Variance Error Trade-Off

It is easy to construct examples where the performance of an efficient or UMVU estimator is poor. In fact, any physical system is built from sub-atomic particles and one would expect that an unbiased estimator would need to model all these, leading to an immensely complex model with an overwhelming number of parameters, resulting in a catastrophically large MSE. Despite this practitioners routinely estimate models of physical systems from data. So how can this be? Well, since these estimators are uniformly optimal in the class of unbiased estimators, to improve on these we either have to drop the ambition to have the estimator be uniformly (over all parameters in the set of possible parameters) best or the requirement of unbiasedness, or both. As in (2.1) we can split the MSE for an estimator \( \hat{\theta} \) into a variance term and a bias term

\[
\text{MSE} \left[ \theta, \hat{\theta} \right] = \mathbb{E} \left[ \| \hat{\theta} - \theta \|^2 \right] = \mathbb{E} \left[ \| \hat{\theta} - \mathbb{E} \left[ \hat{\theta} \right] \|^2 \right] + \| \mathbb{E} \left[ \hat{\theta} \right] - \theta \|^2
\]

(4.1)

Since both terms are positive it may seem as if nothing is gained by using a biased estimator: A biased estimator can always be improved by removing the bias. The catch with this is that the bias is typically \( \theta \) dependent and therefore cannot be removed (as \( \theta \) is unknown). That using biased estimators opens new possibilities is easily seen from an example.

**Example 4.1.1.** Let our scalar observation be

\[ Y = \theta + E \]

where \( E \) is \( N(0, 1) \). Let the estimator be

\[ \hat{\theta} = aY + b \]

Since \( \hat{\theta} = a\theta + aE + b \), this estimator has

\[
\text{MSE} \left[ \theta, \hat{\theta} \right] = a^2 + (a\theta + b - \theta)^2 = a^2 + (1 - a)^2\theta^2 - 2(1 - a)\theta b + b^2
\]

\[ = a^2 + \left( \theta - \frac{b}{1-a} \right) (1-a)^2 \left( \theta - \frac{b}{1-a} \right) \]

(4.2)

The Cramér-Rao lower bound is 1 and is achieved by the ML-estimator \( \hat{\theta}_{ML} = Y \) which is achieved by taking \( a = 1 \) and \( b = 0 \). Since the second term is positive, we see that it is necessary that \( |a| < 1 \)
for the biased estimator to beat the CRLB. Under this condition, the biased estimator beats the CRLB iff

\[ a^2 + \left( \theta - \frac{b}{1-a} \right)^2 \left( 1 - a \right)^2 \left( \theta - \frac{b}{1-a} \right) < 1 \]

\[ \iff \left( \theta - \frac{b}{1-a} \right) \frac{1-a}{1+a} \left( \theta - \frac{b}{1-a} \right) < 1 \]

i.e. \( \theta \) has to belong to the interior of a disc centered at \( b/(1-a) \) with radius \( \sqrt{(1-a)/(1+a)} \).

If we consider the case where \( b = 0 \), we have

\[
\text{MSE}\left[\theta, \hat{\theta}\right] = a^2 + (1-a)^2 \theta^2 = a^2 + \theta^2 a^2 - 2\theta^2 a + \theta^2 \\
= \left( a - \frac{\theta^2}{1+\theta^2} \right) (1+\theta^2) \left( a - \frac{\theta^2}{1+\theta^2} \right) - \frac{\theta^4}{1+\theta^2} + \theta^2 \\
= \left( a - \frac{\theta^2}{1+\theta^2} \right) (1+\theta^2) \left( a - \frac{\theta^2}{1+\theta^2} \right) + \frac{\theta^2}{1+\theta^2}
\]

We see that starting at \( a = 1 \) and then successively decreasing \( a \) results in that the MSE decreases from the CRLB (MSE=1) to the minimum \( \theta^2/(1+\theta^2) \) at \( a = \theta^2/(1+\theta^2) \) and then starts to increase until at \( a = (\theta^2 - 1)/(\theta^2 + 1) \) it reaches the CRLB 1 again. What happens is that when we decrease \( a \), the variance term \( a^2 \) decreases at the expense of increasing the bias term. Initially the gain outweighs the loss but eventually the bias grows too large. For this reason the decomposition (4.1) is known as the bias-variance trade-off.

The example also illustrates that there is no uniformly optimal biased estimator. Taking \( a = 0 \) gives an estimator that has MSE=0 when \( \theta = b \). Therefore our ambitions have to be set lower than for unbiased estimators. However, as in that case we will still consider what we can achieve over a set \( D_\theta \) of parameters.

### 4.2 The Cramér-Rao Lower Bound

For the family of estimators \( \hat{\theta} \) with a certain bias \( b(\theta) := \mathbb{E} \left[ \hat{\theta} \right] - \theta \), the optimal estimator is the one with the lowest variance error. The Cramér-Rao Lower Bound (2.30) gives a lower bound for this part of the error

\[
\text{Cov} \left[ \hat{\theta}(Y) \right] \geq (I + b'(\theta)) \ I_F^{-1}(\theta) \ (I + b'(\theta))^T
\]  

(4.4)

Thus for this family we have the following lower bound on the MSE

\[
\text{MSE} \geq \|b(\theta)\|^2 + \text{Tr} \left[ (I + b'(\theta)) \ I_F^{-1}(\theta) \ (I + b'(\theta))^T \right]
\]  

(4.5)

We conclude that even for biased estimators it is important to have try to make the variance as small as possible.
4.3 Principles for finding biased estimators

Let, as usual, $Y \in \mathbb{R}^N$ be the random variable representing our observations. Since, as we saw in Section 4.1, there is no estimator that uniformly (over $\theta$) minimizes the MSE, we have to relax the requirement of uniformity and use some other criteria. Starting with a general measure of risk

$$R(\theta, \hat{\theta}(Y)) = \mathbb{E} \left[ L(\theta, \hat{\theta}(Y)) \right],$$

where $L$ is a convex function (of which perhaps the MSE is the most common choice), we will discuss three approaches

i) Minimax estimation

ii) Average risk minimization

iii) Pointwise risk estimation

In i) we define the worst-case risk

$$\sup_{\theta \in D_0} R(\theta, \hat{\theta}(Y))$$

and we then try to find the estimator that minimizes this criterion

$$\inf_{\hat{\theta}} \sup_{\theta \in D_0} R(\theta, \hat{\theta}(Y))$$

A number of variations of (4.7) are possible. For example we may compare with the unbiased CRLB

$$\sup_{\theta \in D_0} R(\theta, \hat{\theta}(Y)) - \text{Tr} \left\{ I_F^{-1}(\theta) \right\}$$

In the second approach we define an average risk measure

$$r(\hat{\theta}, f_\Theta) := \int_{D_0} R(\theta, \hat{\theta}(Y)) f_\Theta(\theta) d\theta$$

where $f_\Theta$ is a positive weighting function indicating how important it is to have a low risk for different values of $\theta$. Without loss of generality, $f_\Theta$ can be normalized to have

$$\int_{D_0} f_\Theta(\theta) d\theta = 1$$

The problem is now to find the estimator that minimizes (4.9).

Finally, in iii) we form an estimate $R(\hat{\theta}(Y))$ for the risk $R(\hat{\theta}(Y), \theta_o)$ at the parameter $\theta_o$ underlying the true distribution of the data.

4.4 Admissible Estimators

**Definition 4.4.1.** An estimator is dominated by another estimator if there is no $\theta \in D_0$ for which the former has lower risk than the latter.

An estimator is called inadmissible if it is dominated by another estimator. An estimator which is not inadmissible is called admissible.
Clearly, it is necessary to look for optimal estimators only in the family of admissible estimators.

**Example 4.4.1** (Example 4.1.1 continued). For $a > 1$, $\hat{\theta}$ has larger MSE than $\hat{\theta}_{ML}$ as we saw the example. Thus $a > 1$ gives an inadmissible estimator.

For $a < 0$, $1 - a > 0$ and hence the MSE (4.2) satisfies

$$\text{MSE} \left[ \theta, \hat{\theta} \right] = a^2 + \left( \theta - \frac{b}{1 - a} \right)^2 \left( \theta - \frac{b}{1 - a} \right) \geq \left( \theta - \frac{b}{1 - a} \right)^2$$

where we recognize that the right-hand side is the MSE for the constant estimator $b/(1 - a)$. Thus $a < 0$ also leads to an inadmissible estimator.

Finally, for $a = 1$ gives the estimator $Y + b$ which has bias $1 + b^2 \geq 1$, and hence only $b = 0$ is a candidate for an admissible estimator.

In summary, admissible estimators have $a \in [0, 1]$ and for $a = 1$, $b = 0$ is necessary.

In the example we never made use of the fact that $E$ is normal distributed and thus our findings hold for a general distribution. In the example we proved inadmissibility for some estimators. It is typically much more difficult to prove that an estimator is admissible.

### 4.5 Average Risk Minimization

#### 4.5.1 The Bayes Estimator

Inserting (4.6) in (4.9) gives

$$r(\hat{\theta}, f_{\Theta}) = \int_{D_{\theta}} \mathbb{E} \left[ L(\theta, \hat{\Theta}(Y)) \right] f_{\Theta}(\theta) d\theta = \int_{D_{\theta}} \int L(\theta, \hat{\Theta}(y)) p_{Y}(y; \theta) f_{\Theta}(\theta) dy f_{\Theta}(\theta) d\theta \quad (4.10)$$

An estimator minimizing (4.10) is called a **Bayes estimator**.

If we introduce

$$f_{Y;\Theta}(y, \theta) := p_{Y}(y; \theta) f_{\Theta}(\theta)$$

we can write (4.10) as

$$r(\hat{\theta}, f_{\Theta}) = \int_{D_{\theta}} \int L(\theta, \hat{\Theta}(y)) p_{Y}(y; \theta) f_{\Theta}(\theta) dy d\theta = \int_{D_{\theta}} \int L(\theta, \hat{\Theta}(y)) f_{Y;\Theta}(y, \theta) dy d\theta. \quad (4.11)$$

Now since

$$\int_{D_{\theta}} \int f_{Y;\Theta}(y, \theta) dy d\theta = 1$$

we recognize (4.11) as

$$r(\hat{\theta}, f_{\Theta}) = \mathbb{E} \left[ L(\Theta, \hat{\Theta}(Y)) \right] \quad (4.12)$$

where $\mathbb{E}$ is the expectation over the random variables $\{Y, \Theta\}$ with joint pdf $f_{Y;\Theta}$. Restricting attention to the case where the loss $L$ is the squared error $L(\theta, \hat{\theta}) = ||\hat{\theta} - \theta||^2$, this expression is the MSE for the estimator $\hat{\Theta}(Y)$ that uses the random variable $Y$ to estimate the *random variable* $\Theta$ when the joint distribution of $Y$ and $\Theta$ is given by $f_{Y;\Theta}(y, \theta)$. From Section 2.2 we have the optimal solution to this
4.5. AVERAGE RISK MINIMIZATION

Problem: We should take as \( \hat{\theta} \) the conditional mean of \( \Theta \) given \( Y = y \) i.e. the mean of \( \Theta \) with respect to the pdf

\[
f_{\Theta|Y}(\theta|y) := \frac{f_{Y,\Theta}(y, \theta)}{f_Y(y)} = \frac{p_Y(y; \theta)f_\Theta(\theta)}{f_Y(y)}, \quad \text{where } f_Y(y) := \int_{\Theta} f_{Y,\Theta}(y, \theta)d\theta = \int_{\Theta} p_Y(y; \theta)f_\Theta(\theta)d\theta
\]

When \( L \) is quadratic, the Bayes estimator is thus given by

\[
\hat{\theta}_{fa}(y) = \int_{\Theta} \theta f_{\Theta|Y}(\theta|y)d\theta
\]

(4.13)

Recall from Section 2.2 that this estimator minimizes

\[
\mathbb{E}\left[L(\theta, \hat{\theta}(y))|Y = y\right] = \int_{\Theta} L(\theta, \hat{\theta}(y))f_{\Theta|Y}(\theta|y)d\theta \propto \int_{\Theta} L(\theta, \hat{\theta}(y))p_Y(y; \theta)f_\Theta(\theta)d\theta \quad (4.14)
\]

when the cost function \( L \) is quadratic.

The minimum average risk, i.e. the average risk for a Bayes estimator is called the Bayes risk.

4.5.2 A Bayesian Interpretation

We arrived at (4.13) by interpreting \( f_{Y,\Theta}(y, \theta) \) as a probability density function over both \( y \) and \( \theta \). This does not mean that we consider \( \theta \) to be the outcome of a random variable; it is just that the solution to our problem coincides with the solution to the problem of estimating the random variable \( \Theta \), with pdf \( p_\Theta(\theta) := f_\Theta(\theta) \) from the random variable \( Y \), with pdf \( p_Y(y) := f_Y(y) \) when their joint pdf is \( f_{Y,\Theta}(y, \theta) = p_Y(y; \theta)f_\Theta(\theta) = p_Y(y; \theta)p_\Theta(\theta) \), i.e. when the conditional distribution of \( Y \) given \( \Theta \) is \( p_Y(y; \theta) \). The conditional distribution of \( \Theta \) given \( Y \) is given by the pdf \( p_{\Theta|Y} := f_{\Theta|Y} \).

In the latter problem, \( p_\Theta \) is known as the prior distribution and \( p_{\Theta|Y} \) as the posterior distribution of \( \Theta \), respectively.

The Bayes estimator can thus also be motivated directly by assuming that \( \theta \) actually is an outcome of a random variable \( \Theta \). This can be motivated when we are trying to estimate a number of systems with some spread in between them, e.g. manufacturing.

The prior \( p_\Theta \) can also be thought as a state of mind of the experimenter reflecting a subjective feeling about the likelihood of different values of \( \theta \), e.g. obtained from prior experience of similar problems.

Example 4.5.1. Suppose that

\[
Y = \Phi \theta + E, \quad Y \in \mathbb{R}^N, \quad \theta \in \mathbb{R}^n, \quad n \leq N, \quad E \sim N(0, \lambda_e I)
\]

i.e. \( Y \sim N(\Phi \theta, \lambda_e I) \).

Now let the weighting function \( f_\Theta \) be the pdf of the \( N(\hat{\theta}, \lambda_\theta I) \) distribution. When deriving the Bayes estimator we can (temporarily) think of \( \theta \) being a realization of \( \Theta \sim N(\hat{\theta}, \lambda_\theta I) \) so that the conditional distribution of \( Y \) given \( \Theta \) is \( N(\Phi \theta, \lambda_e I) \) and the joint distribution is

\[
\begin{bmatrix} \Theta \\ Y \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ \Phi \hat{\theta} \end{bmatrix}, \begin{bmatrix} \lambda_\theta I & \lambda_\theta \Phi^T \\ \lambda_\theta \Phi & \lambda_\theta \Phi \Phi^T + \lambda_e I \end{bmatrix} \right)
\]

Here we recognize that the lower right corner corresponds to \( f_Y \), i.e. \( f_Y \sim N(\Phi \hat{\theta}, \lambda_\theta \Phi \Phi^T + \lambda_e I) \).
Since we have a joint normal distribution, the posterior distribution of $\Theta$ given $Y$ is also normal with mean

$$\text{Cov} [\Theta, Y] \text{Cov} [Y, Y]^{-1} (Y - E[Y]) + E[\Theta]$$

and covariance

$$\text{Cov} [\Theta, \Theta] - \text{Cov} [\Theta, Y] \text{Cov} [Y, Y]^{-1} \text{Cov} [Y, \Theta]$$

(see Exercise 2.2.1) which gives

$$\Theta | Y \sim N \left( \lambda_\theta \Phi^T \left( \lambda_\theta \Phi \Phi^T + \lambda_e I \right)^{-1} (Y - \Phi \bar{\theta}) + \lambda_\theta I - \lambda_\theta^2 \Phi^T \left( \lambda_\theta \Phi \Phi^T + \lambda_e I \right)^{-1} \Phi \right)$$

(4.15)

Using Exercise 4.9.1

$$\lambda_\theta \Phi^T \left( \lambda_\theta \Phi \Phi^T + \lambda_e I \right)^{-1} = \left( \frac{1}{\lambda_e} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \Phi^T \frac{1}{\lambda_e}$$

(4.16)

the Bayes estimator (the mean of (4.15)) can be expressed as

$$\hat{\theta}_{fe} = \left( \frac{1}{\lambda_e} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \left( \Phi^T \frac{1}{\lambda_e} Y + \frac{1}{\lambda_\theta} \bar{\theta} \right)$$

(4.17)

That the solution to the average risk estimation problem and the posterior mean, when $\theta$ is assumed to be stochastic, coincide is a source of significant confusion in the literature and great care has to be taken when interpreting the results. In particular the different measures of uncertainty is a source misinterpretation and misuse. So far in this exposition we have used the MSE as a quality measure of an estimator, or, even more general, the distribution of the estimator. For example, in Section 3.3.1 we discussed estimators that become normal distributed for large sample sizes. The distribution reflects the spread of the estimator around the true parameter when the same experiment is repeated over and over again with the same system generating the data (i.e. keeping the parameter underlying the pdf governig the distribution of the observations fix). This is known as a frequentistic perspective. In a Bayesian setting, the posterior $p_\Theta | Y$ reflects the spread of the estimator when, instead, in each experiment a new value $\theta$ is drawn from the prior distribution of $\Theta$. These two spreads can be very different.

Overall, whether to use a Bayesian or a frequentistic approach is historically probably the largest controversial issue in statistics and many interesting discussion on this topic can be found in the literature. We will remain neutral on this issue, but stress the importance of knowing the differences between the two approaches and to use the right concept for the right problem. Our perspective will, however, continue to be the frequentist’s so next we analyse the bias and variance of the Bayes estimator.

### 4.5.3 Bias and Variance of the Bayes Estimator

We continue with Example 4.5.1.
**Example 4.5.2** (Example 4.5.1 continued). The Bayes estimator (4.17) has mean (remember that now we are back to that $Y \sim N(\Phi \theta, \lambda_\theta I)$)

$$
\mathbb{E} \left[ \hat{\theta}_{fe} \right] = \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \Phi^T \frac{1}{\lambda_\theta} \mathbb{E} [Y] = \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \left( \frac{1}{\lambda_\theta} \Phi^T \Phi \theta + \frac{1}{\lambda_\theta} \bar{\theta} \right)
$$

Introducing

$$
M := \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \frac{1}{\lambda_\theta} \Phi^T \Phi
$$

(4.18)
gives

$$
\mathbb{E} \left[ \hat{\theta}_{fe} \right] = M \theta + (I - M) \bar{\theta}
$$

and the bias is thus

$$
b(\theta) := \mathbb{E} \left[ \hat{\theta}_{fe} \right] - \theta = (M - I)(\theta - \bar{\theta}) \quad (4.19)
$$

which is non-zero for all $\theta$ except for $\theta = \bar{\theta}$.

The covariance is given by

$$
\text{Cov} \left[ \hat{\theta}_{fe} \right] = \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \Phi^T \frac{1}{\lambda_\theta} \text{Cov} \left[ Y \right] \frac{1}{\lambda_\theta} \Phi \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1}
$$

$$
= \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \frac{1}{\lambda_\theta} \Phi^T \Phi \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1}
$$

$$
= M \left( \frac{1}{\lambda_\theta} \Phi^T \Phi \right)^{-1} M^T
$$

(4.20)

We can compare (4.20) with the posterior covariance given in (4.15). Using (4.16) the posterior covariance is given by

$$
\text{Cov} \left[ \Theta | Y \right] = \lambda_\theta I - \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \frac{\lambda_\theta}{\lambda_\theta} \Phi^T \Phi = \lambda_\theta (I - M)
$$

which we also can write as

$$
\text{Cov} \left[ \Theta | Y \right] = \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1}
$$

(4.21)

Comparing with (4.20), we see that the posterior covariance differs from the covariance of the Bayes estimator, with the posterior covariance being the larger of the two.

In the general case the Bayes estimator has mean

$$
\mathbb{E} \left[ \hat{\theta}_{fe}(y) \right] = \int \int_D \hat{\theta}_{fe}(y|\theta)d\theta p_Y(y; \theta)dy = \int \int_D \hat{\theta} \frac{p_Y(y; \bar{\theta}) f_{\Theta}(\bar{\theta})}{f_Y(y)} d\theta p_Y(y; \theta)dy
$$

(4.22)

for quadratic loss $L$. This is a relatively complicated expression. As in Example 4.5.2, one can show (see Exercise 4.9.3) that the Bayes estimator is biased except for the degenerate case where the average risk is zero.

As for the bias (4.22), the covariance of the Bayes estimator is given by an integral expression.
4.5.4 The Prior Viewed as Fictitious Observations

Example 4.5.3 (Example 4.5.2 continued). Suppose that
\[
\begin{bmatrix} Y \\ \hat{\theta} \end{bmatrix} \sim N \left( \begin{bmatrix} \Phi \theta \\ \lambda_e I \end{bmatrix}, \begin{bmatrix} \lambda_e I & 0 \\ 0 & \lambda_\theta I \end{bmatrix} \right)
\]
Then the UMVU estimator of \( \theta \) given observations of both \( Y \) and \( \hat{\theta} \) is given by
\[
\hat{\theta}_{\text{UMVU}} = \left( \frac{1}{\lambda_e} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \left( \frac{\Phi^T}{\lambda_e} Y + \frac{\lambda_\theta}{\lambda_e} \hat{\theta} \right)
\]
In view of (4.17), the UMVU estimator for the above problem is identical to the Bayes estimator in our running example. This prompts the interpretation that a prior corresponds to adding fictitious measurements. However, the covariance of the UMVU estimator above is given by
\[
\text{Cov} \left[ \hat{\theta}_{\text{UMVU}} \right] = \left( \frac{1}{\lambda_e} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1}
\]
and coincides with the posterior covariance (4.21) for the Bayes estimator and not its covariance (4.20).

Notice that the UMVU estimator can be written as
\[
\hat{\theta}_{\text{UMVU}} = \left( \frac{1}{\lambda_e} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \Phi^T \Phi \hat{\theta}_{\text{ML}} + \left( \frac{1}{\lambda_e} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} I \frac{\lambda_\theta}{\lambda_e} \hat{\theta}
\]
where \( \hat{\theta}_{\text{ML}} \) is the ML estimate of \( \theta \) based on \( Y \). The expression weighs \( \hat{\theta}_{\text{ML}} \) and \( \hat{\theta} \) in proportion to their covariances, compare with BLUE.

4.5.5 Sufficiency and the Bayes Estimator

Suppose that \( T(Y) \) is a sufficient statistic for \( Y \), i.e. \( p_Y(y; \theta) = g(T(y); \theta) h(y) \). Now, the Bayes estimator (4.13) depends only on \( f_{\theta|Y} \) which can be expressed as
\[
f_{\theta|Y}(y|\theta) = \frac{f_Y(y, \theta)}{f_Y(y)} = \frac{p(y; \theta) f_{\theta}(\theta)}{\int p(y; \theta) f_{\theta}(\theta) d\theta} = \frac{g(T(y); \theta) h(y) f_{\theta}(\theta)}{\int g(T(y); \theta) h(y) f_{\theta}(\theta) d\theta} = \frac{g(T(y); \theta) f_{\theta}(\theta)}{\int g(T(y); \theta) f_{\theta}(\theta) d\theta}
\]
Thus we can use \( T(Y) \) to compute the Bayes estimator. We simply replace \( p_Y(y; \theta) \) with \( g(T(Y); \theta) \).

Example 4.5.4 (Example 4.5.2 continued). Since
\[
p_Y(y; \theta) = ce^{-\frac{1}{2\lambda_e} \|Y - \Phi \theta\|^2} = ce^{-\frac{1}{2\lambda_e} (\|Y\|^2 - 2\theta^T \Phi^T Y + \|\Phi \theta\|^2)}
\]
a sufficient statistic is \( \Phi^T Y \). However, if \( \Phi \) has full column rank, the ML estimate \( \hat{\theta}_{\text{ML}} := (\Phi^T \Phi)^{-1} \Phi^T Y \) is also a sufficient statistic and its distribution is given by
\[
\hat{\theta}_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T Y = \theta + (\Phi^T \Phi)^{-1} \Phi^T E \sim N \left( \theta, \lambda_e (\Phi^T \Phi)^{-1} \right)
\]
(4.23)
Thus the joint distribution of (our imaginary) \( \Theta \) and \( \hat{\theta}_{\text{ML}} \) is
\[
\begin{bmatrix} \Theta \\ \hat{\theta}_{\text{ML}} \end{bmatrix} \sim N \left( \begin{bmatrix} \theta \\ \hat{\theta} \end{bmatrix}, \begin{bmatrix} \lambda_e I & \lambda_\theta I \\ \lambda_\theta I & \lambda_\theta I + \lambda_e (\Phi^T \Phi)^{-1} \end{bmatrix} \right)
\]
(4.24)
from which we get that the Bayes estimator can be expressed as

\[ \hat{\theta}_{f_\theta} = \lambda_\theta \left( \lambda_\theta I + \lambda_\theta (\Phi^T \Phi)^{-1} \right)^{-1} \left( \hat{\theta}_{\text{ML}} - \hat{\theta} \right) + \hat{\theta} \]

\[ = \left( \frac{1}{\lambda_\theta} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \left( \frac{1}{\lambda_\theta} \Phi^T \Phi \hat{\theta}_{\text{ML}} + \frac{1}{\lambda_\theta} \hat{\theta} \right) \]

\[ = M \hat{\theta}_{\text{ML}} + (I - M) \hat{\theta} \quad (4.25) \]

which is exactly (4.17).

Now, besides being a sufficient statistic, \( \hat{\theta}_{\text{ML}} \) is an efficient estimator of \( \theta \). Thus we see that the Bayes estimator modifies the estimator \( \hat{\theta}_{\text{ML}} \) by a linear transformation corresponding to the symmetric matrix \( M \). We notice that this is a contraction since all the eigenvalues of \( M \) are strictly less than 1. Notice in particular that \( M \) tends to 0 as \( \lambda_\theta \to 0 \). This means that the Bayes estimator shrinks \( \hat{\theta}_{\text{ML}} \) more and more towards 0 the smaller \( \lambda_\theta \) becomes. At the same time the second term approaches \( \bar{\theta} \). So why is this? What happens when \( \lambda_\theta \) becomes small is that the weighting function \( f_\theta(\theta) \) becomes more and more concentrated around \( \bar{\theta} \). It is thus important (from the average risk function’s perspective) to have an accurate estimate when \( \theta \) is in a neighbourhood of \( \bar{\theta} \). This can be achieved by introducing a large variance reduction in this neighbourhood at the cost of a small bias. From (4.20) we see that indeed the covariance of the Bayes estimator tends to zero as \( \lambda_\theta \to 0 \) and from (4.19) follows that the bias tends to \( -(\theta - \bar{\theta}) \), which is also small in a neighbourhood of \( \bar{\theta} \). The downside of this strategy is that, while the variance becomes uniformly (in \( \theta \)) small, the bias (\( \approx -(\theta - \bar{\theta}) \)) becomes large except in a neighbourhood of \( \bar{\theta} \). However, the Bayes estimator balances this trade-off optimally for a given weight \( f_\theta \).

The shrinkage effect we saw in Example 4.5.4 is typical for Bayes estimators.

### 4.5.6 The Bayes Estimator for Exponential Families*

We saw in Example 4.5.1 that it is easy to compute the Bayes estimator for the case of the normal distributions. More generally, when \( p_Y(y; \theta) \) is an exponential family the following result can be used.

**Theorem 4.5.1** (Theorem 4.3.2 in [10]). Suppose that \( p_Y(y; \theta) \) is given by an exponential family on the canonical form (2.33):

\[ p_Y(y; \theta) = e^{\theta^T T(y) - A(\theta) h(y)} \quad (4.26) \]

Then the Bayes estimator is given by

\[ \hat{\theta}_{f_\theta}(y) = (T'(y) T'(y))^{-1} \left( \frac{d}{dy} \log f_Y(y) - \frac{d}{dy} \log h(y) \right) \]

### 4.5.7 Admissibility of the Bayes Estimator*

We say that the Bayes estimator is unique if any two Bayes estimators only differ on a set \( A \subset \Omega \) which has zero probability measure regardless of \( \theta \).

**Theorem 4.5.2** (Theorem 5.2.4 in [10]). Any unique Bayes estimator is admissible.

**Proof.** Suppose that the Bayes estimator \( \hat{\theta}_{f_\theta} \) is dominated by \( \hat{\theta} \). Then \( R(\theta, \hat{\theta}) \leq R(\theta, \hat{\theta}_{f_\theta}) \) implying \( r(\hat{\theta}) \leq r(\hat{\theta}_{f_\theta}) \) but this contradicts the uniqueness. \( \square \)
If the risk function is continuous, the uniqueness assumption can be dropped (Theorem 5.7.9 in [10]).

The family of Bayes estimators constitute a very large class of biased estimators since there is a large degree of freedom in choosing the weighting function $f_\Theta$. The next result establishes the Bayes estimators as essentially the only family of estimators that need to be considered.

**Theorem 4.5.3.** Let $Y$ have pdf $p_y(y; \theta)$ such that $p_y(y; \theta) > 0 \forall y \in \mathcal{Y}$, $\theta \in \mathcal{D}_\theta$. Let the loss function $L(\theta, \hat{\theta})$ be continuous, strictly convex in $\hat{\theta}$ for every $\theta \in \mathcal{D}_\theta$. Let

$$\lim_{|\theta| \to \infty} L(\theta, \hat{\theta}) = \infty \quad \forall \theta \in \mathcal{D}_\theta$$

Then, to every admissible estimator $\hat{\theta}(Y)$ there corresponds a sequence $f_{\Theta,n}$ of weighting functions with support on a finite set (and hence with finite Bayes risk) for which the corresponding sequence of Bayes estimators $\{\hat{\theta}_{f_{\Theta,n}}(Y)\}_{n=1}^\infty$ satisfies

$$\hat{\theta}_{f_{\Theta,n}}(Y) \to \hat{\theta}(Y) \quad a.e.$$

If $f_\Theta$ is not a proper pdf then the Bayes estimator may still be well defined. If it still minimizes the posterior expected loss $\mathbb{E} \left[ L(\theta, \hat{\theta}(Y)) | Y = y \right]$ for all $y$, then it is said to be a generalized Bayes estimator. For exponential families, all admissible estimators belong to the family of generalized Bayes estimators under the assumptions of Theorem 4.5.3.

### 4.5.8 Hierarchical Bayes

The weighting function $f_\Theta$ can be built up in several stages, e.g.

$$f_\Theta(\theta) = \int f_{\Theta|\Gamma}(\theta | \gamma) f_{\Gamma}(\gamma) d\gamma$$

Taking $f_{\Gamma}(\gamma)$ and $f_{\Theta|\Gamma}(\theta | \gamma)$ as pdfs, the stochastic interpretation is that $\theta$ is generated by first drawing a value $\gamma$ from the distribution $f_{\Gamma}(\gamma)$ and then $\theta$ from the conditional distribution $f_{\Theta|\Gamma}(\theta | \gamma)$. This is sometimes a systematic way of constructing a weighting function that can be given some interpretation. Several steps can be included as well as parallel branches leading to a *Bayesian network* structure of $f_\Theta$.

### 4.5.9 Empirical Bayes

The idea of using a weighting function $f_\Theta$ is to give higher weights in regions where it is thought that the true model parameter $\theta$ can be. A natural idea is to use the data itself to decide this. This can be achieved by parametrizing $f_\Theta(\theta; \eta)$ and then to tune $\eta$ using data. This approach is called Empirical Bayes.

A popular way to tune $\eta$ is by maximum likelihood minimization. Adapting the notion that $\theta$ is drawn from the distribution $f_\Theta(\theta; \eta)$. Stressing that we now think of $f_\Theta(\theta; \eta)$ as a prior distribution, we write $p_\Theta(\theta; \eta) := f_\Theta(\theta; \eta)$. The likelihood function is given by

$$p_Y(Y; \eta) = \int p_{Y|\Theta}(Y | \theta) p_{\Theta}(\theta; \eta) d\theta$$
4.5. AVERAGE RISK MINIMIZATION

\[ \hat{\eta}_{\text{ML}} = \arg \max_{\eta \in \mathcal{D}} p_Y(Y; \eta) \]

It should be stressed that even though \( \eta \) can be estimated using ML as above, this parameter does not govern the distribution of \( Y \); that distribution depends on \( \theta \): \( p_Y(y; \theta) \). To distinguish the roles of \( \theta \) and \( \eta \), the latter is called a hyperparameter.

**Example 4.5.5** (Example 4.5.4 continued). Suppose that \( \lambda_\theta \) in Example 4.5.4 is left as a free hyperparameter. From (4.24), assuming for simplicity \( \Phi^T \Phi = I \) (e.g. \( \Phi = I \)), we obtain that

\[
-2 \log p_Y(y; \lambda_\theta) \propto \log \det (\lambda_\theta I + \lambda_e I) + (\hat{\theta}_{\text{ML}} - \bar{\theta})^T (\lambda_\theta I + \lambda_e I)^{-1} (\hat{\theta}_{\text{ML}} - \bar{\theta})^T
\]

and hence

\[
-2 \frac{\partial}{\partial \lambda_\theta} \log p_Y(y; \lambda_\theta) = n \frac{1}{\lambda_\theta + \lambda_e} - \frac{\|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2}{(\lambda_\theta + \lambda_e)^2}
\]

giving that

\[
\hat{\lambda}_{\theta,\text{ML}} = \frac{1}{n} \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2 - \lambda_e
\]  

(4.27)

Thus \( f_{\theta} \sim N \left( \bar{\theta}, \frac{1}{n} \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2 - \lambda_e \right) \), i.e. the spread of the independent elements of \( \hat{\theta}_{\text{ML}} \) is used as an estimate of the variance \( \lambda_\theta \). A quite reasonable approach. The term \(-\lambda_e\) takes into account that the spread of \( \hat{\theta}_{\text{ML}} \) also depends on \( \lambda_e \).

In our case (4.18) becomes

\[
M = \frac{1}{\lambda_e + \frac{1}{\lambda_\theta}} I = \frac{\lambda_\theta}{\lambda_\theta + \lambda_e} I
\]

(4.29)

and replacing \( \lambda_\theta \) by \( \hat{\lambda}_{\theta,\text{ML}} \) gives

\[
M = \frac{\frac{1}{n} \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2 - \lambda_e}{\frac{1}{n} \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2} I = \left( 1 - \frac{\lambda_e}{\frac{1}{n} \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2} \right) I
\]

and then (4.25) gives the Empirical Bayes estimator as a convex combination of \( \hat{\theta}_{\text{ML}} \) and \( \bar{\theta} \)

\[
\hat{\theta}_{\text{EB}} = M \hat{\theta}_{\text{ML}} + (I - M) \bar{\theta} = \left( 1 - \frac{\lambda_e}{\frac{1}{n} \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2} \right) \hat{\theta}_{\text{ML}} + \frac{\lambda_e}{\frac{1}{n} \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2} \bar{\theta}
\]

(4.30)

This expression can be given an interesting interpretation. Recall from (4.23) that \( \hat{\theta}_{\text{ML}} \sim N(\theta, \lambda_e I) \). Now, under the hypothesis that \( \theta = \bar{\theta}, T := \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2/\lambda_e \) is \( \chi^2(n) \)-distributed and the probability that it exceeds \( n \) is approximately 50%. Thus we see that the more the hypothesis \( \hat{\theta}_{\text{ML}} \sim N(\bar{\theta}, \lambda_e I) \) is violated, the more the EB-estimate will rely on data (i.e. \( \hat{\theta}_{\text{ML}} \)), while the more the hypothesis is confirmed (i.e. the smaller \( T \) is), the more it will rely on \( \bar{\theta} \).
CHAPTER 4. BIASED ESTIMATORS

We observe that \( \hat{\lambda}_{\theta,ML} \) can be negative, which is inconsistent with our model of \( \lambda_\theta \) as a variance. Introducing \( (x)^+ = \max(0, x) \), the EB-estimator using \( (\hat{\lambda}_{\theta,ML})^+ \) can be written

\[
\hat{\theta}_{\text{EB}} = \left(1 - \frac{\lambda_e}{\frac{1}{n}\|\theta_{\text{ML}} - \bar{\theta}\|^2}\right)^+ \hat{\theta}_{\text{ML}} + \min\left(\frac{\lambda_e}{\frac{1}{n}\|\theta_{\text{ML}} - \bar{\theta}\|^2}, 1\right) \tilde{\theta}, \tag{4.31}
\]

which again can be given the interpretation as being based on hypothesis test. Now, if data strongly contradicts the hypothesis, only data will be used.

The estimator (4.30) was studied by James and Stein in [6] and was first derived without any reference to Empirical Bayes. It can be shown that (4.30) dominates the least-squares estimate \( \hat{\theta}_{\text{ML}} \), which has constant risk \( n\lambda_e \), but in turn (4.30) is dominated by (4.31).

Notice that under our assumption \( \Phi^T\Phi = I \), the elements of \( \hat{\theta}_{\text{ML}} \) are independent. Now, assuming for simplicity that \( \bar{\theta} = 0 \), element \( i \) of (4.30) is given by

\[
\left(1 - \frac{\lambda_e}{\frac{1}{n}\sum_{k=1}^{n} \theta_{\text{ML},k}^2}\right) \hat{\theta}_{\text{ML},i}
\]

This means that in comparison with \( \hat{\theta}_{\text{ML},i} \), the estimator (4.30) improves on the estimate of \( \theta_i \) using observations \( \hat{\theta}_{\text{ML},1}, \ldots, \hat{\theta}_{\text{ML},i-1}, \hat{\theta}_{\text{ML},i+1}, \hat{\theta}_{\text{ML},n} \) that do not contain any information about \( \theta_i \). A rather disconcerting fact known as the Stein effect.

The number of estimated parameters in relation to the number of observations will greatly influence the quality of the estimate.

**Example 4.5.6** (Example 4.5.5 continued). Suppose that instead \( \bar{\theta} \) is the hyperparameter. Then (4.27) gives that \( \hat{\theta}_{\text{ML}} = \hat{\theta}_{\text{ML}} \), and inserting this (4.25) gives that the EB-estimator is \( \hat{\theta}_{\text{ML}} \) itself.

### 4.5.10 Constrained Bayes Estimation

While the Bayes estimator (4.13) minimizes (4.14), there may be reasons for not choosing this estimator. For example, the application may require a particular structure of the estimator, i.e. \( \hat{\theta} \in \bar{D}_\theta \). The estimator would then be taken as

\[
\hat{\theta}(y) = \arg\min_{\theta \in \bar{D}_\theta} \mathbb{E}\left[ L(\Theta, \hat{\theta}(y))|Y = y \right]
\]

With quadratic loss function

\[
\mathbb{E}\left[ L(\Theta, \hat{\theta}(y))|Y = y \right] = \mathbb{E}\left[ \|\Theta - \hat{\theta}(y)\|=y\|^2 \right] = \mathbb{E}\left[ \|\Theta - \mathbb{E}[\Theta|Y = y]\|^2 \right] + \|\mathbb{E}[\Theta|Y = y] - \hat{\theta}(y)\|^2
\]

implying that

\[
\hat{\theta}(y) = \arg\min_{\theta \in \bar{D}_\theta} \|\mathbb{E}[\Theta|Y = y] - \hat{\theta}(y)\|^2
\]

However, computationally it may be easier to minimize (4.14) since computation of the normalization \( f(y) \) can be avoided.

\(^1\)Typically \( \bar{\theta} = 0 \) is used.
4.6 Minimax Estimation

In this section we will study estimators that minimize the worst-case risk (4.7)

$$\sup_{\theta \in D_\theta} R(\theta, \hat{\theta}(Y))$$  (4.32)

Such estimators are called *minimax* estimators.

There is a strong connection between such estimators and Bayes estimators. Let \(\hat{\theta}\) be an arbitrary estimator with finite risk. Then for any pdf \(f_\theta\)

$$r(\hat{\theta}, f_\theta) = \int R(\theta, \hat{\theta}) f_\theta(\theta) d\theta \leq \sup_{\hat{\theta}} R(\theta, \hat{\theta}) \int f_\theta(\theta) d\theta = \sup_{\hat{\theta}} R(\theta, \hat{\theta})$$

Thus, the worst-case risk for a given estimator \(\hat{\theta}\) is lower bounded by the average risk for any weighting function \(f_\theta\). Now, the average risk for the Bayes estimator is no greater than that of \(\hat{\theta}\), implying

$$r(\hat{\theta}, f_\theta) \leq \int f_\theta(\theta) d\theta \leq \sup_{\theta} R(\theta, \hat{\theta})$$

Suppose now that the risk of the Bayes estimator is constant for \(f^*_\theta\). Then

$$\sup_{\theta} R(\theta, \hat{\theta} f^*_\theta) = \int R(\theta, \hat{\theta} f^*_\theta) f^*_\theta(\theta) d\theta \leq \int R(\theta, \hat{\theta}) f^*_\theta(\theta) d\theta = \sup_{\theta} \int R(\theta, \hat{\theta} f^*_\theta) f^*_\theta(\theta) d\theta = r(\hat{\theta}, f^*_\theta)$$

Since this holds for any estimator \(\hat{\theta}\), \(\hat{\theta} f^*_\theta\) must be minimax.

We remark that \(f^*_\theta\) is least favorable in the sense that it has the largest average risk for the Bayes estimator. Let \(r(f_\theta)\) denote the minimum of the average risk (4.10), i.e. \(r(f_\theta) = r(\hat{\theta} f_\theta, f_\theta)\). Then

$$r(f_\theta) = \int R(\theta, \hat{\theta} f_\theta) f_\theta(\theta) d\theta \leq \int R(\theta, \hat{\theta} f^*_\theta) f_\theta(\theta) d\theta \leq \sup_{\theta} R(\theta, \hat{\theta}) = \int R(\theta, \hat{\theta} f^*_\theta) f^*_\theta(\theta) d\theta = r(f^*_\theta)$$

Our findings generalize to the following result.

**Theorem 4.6.1** (Theorem 1.4 in [10]). Let \(P_\Theta\) be a probability measure on \(D_\theta\) and let \(\hat{\theta}_{P_\Theta}\) denote the corresponding Bayes estimator. Suppose that

$$r(\hat{\theta}_{P_\Theta}, P_\Theta) = \int R(\theta, \hat{\theta}_{P_\Theta}) P_\Theta(d\theta) = \sup_{\theta} R(\theta, \hat{\theta}_{P_\Theta})$$

Then

i) \(\hat{\theta}_{P_\Theta}\) is minimax.

ii) If \(\hat{\theta}_{P_\Theta}\) is the unique Bayes estimator with respect to \(P_\Theta\), it is the unique minimax estimator.

iii) \(P_\Theta\) is the least favorable distribution.

Often a least favorable distribution does not exist. Then the following limit can be used instead.

**Definition 4.6.1.** A sequence of probability measures \(\{P_{\Theta,n}\}\) is least favorable if for every probability measure \(P\) we have

$$r(P_\Theta) \leq r = \lim_{n \to \infty} r(P_{\Theta,n})$$  (4.33)
Theorem 4.6.2. Suppose that \( \{P_{\Theta, n}\} \) have Bayes risks satisfying (4.33) and that \( \hat{\theta} \) is an estimator satisfying
\[
\sup_{\theta} R(\theta, \hat{\theta}) = r
\]
Then
i) \( \hat{\theta} \) is minimax.

ii) The sequence \( P_{\Theta, n} \) is the least favorable.

Example 4.6.1. Consider \( Y \in \mathbb{R}^N \) with elements being iid \( N(\theta, 1) \). Then the sample mean
\[
\bar{Y} := \frac{1}{N} \sum_{k=1}^{N} Y_k
\]
is minimax. To see this consider the Bayes estimator corresponding to \( f_{\Theta} \sim N(\bar{\theta}, \lambda) \). This is a special case of Example 4.5.1 with \( \Phi = \text{1} \) (a vector with 1 in all elements) and, hence, (4.17) gives
\[
\hat{\theta}_{f_{\Theta}} = \left( \frac{N}{\lambda_e} + \frac{1}{\lambda} \right)^{-1} \left( \frac{N}{\lambda_e} \frac{\bar{Y}}{\lambda} + \frac{\bar{\theta}}{\lambda} \right)
\]
(4.34)
The posterior covariance of the Bayes estimator is given by (4.21)
\[
\text{Cov} [\Theta | Y] = \left( \frac{1}{\lambda_e} \Phi^T \Phi + \frac{1}{\lambda} I \right)^{-1} = \left( \frac{N}{\lambda_e} + \frac{1}{\lambda} \right)^{-1}
\]
which is independent of \( Y \) and hence the Bayes risk (which is given by the expectation of \( \text{Cov} [\Theta | Y] \) with respect to \( Y \)) is
\[
r(\hat{\theta}_{f_{\Theta}}) = \left( \frac{N}{\lambda_e} + \frac{1}{\lambda} \right)^{-1} \frac{\lambda_e}{N} =: r \quad \text{as} \quad \lambda \to \infty
\]
Now \( \bar{Y} \) has risk
\[
R(\theta, \bar{Y}) = \mathbb{E} [ |\bar{Y} - \theta|^2 ] = \frac{\lambda_e}{N} = r
\]
and thus by Theorem 4.6.2 it is minimax.

4.7 Pointwise Risk Estimation

The last approach to systematically derive biased estimators - pointwise risk estimation - has a strong connection to Empirical Bayes in that the hyperparameters in an Empirical Bayes estimator can be determined using pointwise risk estimation. Typically also here the family of estimators is parametrized by a hyperparameter \( \eta: \hat{\theta}_\eta(Y), \eta \in \mathcal{D}_\eta \). The key step is to estimate the risk corresponding to the true distribution, i.e. \( R(\theta_o, \hat{\theta}_\eta(Y)) \), and then to pick the estimator that minimizes the estimated risk:

i) Compute an estimate \( \hat{R}(\eta) \) of \( R(\theta_o, \hat{\theta}_\eta(Y)) \) for \( \eta \in \mathcal{D}_\eta \).

ii) \( \hat{\eta} = \arg \min_{\eta \in \mathcal{D}_\eta} \hat{R}(\eta) \).

iii) Compute \( \hat{\theta}_\eta(Y) \).

This approach covers a wide range of methods.
### 4.7. POINTWISE RISK ESTIMATION

#### 4.7.1 Estimating Risk

As the risk depends on $\theta_0$, this means that some type of estimate of $\hat{\theta}$ must be used and the properties of the resulting estimator depends critically on this estimate. However, as we will see it is not specifically an estimate of $\theta_0$ itself that may be required.

### ML Risk Estimation

**Theorem 4.7.1.** The ML estimator of $R(\theta_0, \hat{\theta})$ is given by $R(\theta_0, \hat{\theta})|_{\theta_0=\hat{\theta}_{ML}}$ for any estimator $\hat{\theta}$.

Furthermore, let the optimal hyperparameter and the optimal estimator be defined by

$$
\eta(\theta_0) = \arg\min_{\eta \in D} R(\theta_0, \hat{\theta}_{\eta}(Y)),
$$

$$
\hat{\theta}^*(Y) = \hat{\theta}_{\eta(\theta_0)}(Y)
$$

respectively. Then the ML estimators of these two quantities are given by

$$
\hat{\eta}_{ML}(Y) = \arg\min_{\eta \in D} R(\theta_0, \hat{\theta}_{\eta}(Y))|_{\theta_0=\hat{\theta}_{ML}} = \eta(\hat{\theta}_{ML})
$$

(4.35)

$$
\hat{\theta}^*_{ML} = \hat{\theta}_{\hat{\eta}_{ML}}(Y)
$$

(4.36)

**Proof.** The theorem follows directly from the invariance principle discussed in Section 2.3.8. □

From our knowledge of ML-estimation, we would expect $\hat{\theta}^*_{ML}$ to have nice asymptotic properties. However, we also know from Section 3.1 that asymptotically it is the variance error that dominates.

**Example 4.7.1** (Example 4.5.6 continued). We consider the Bayes estimator (4.25) as parametrized by $\lambda_\theta$ but with $\bar{\theta}$ fix:

$$
\hat{\theta}_{\lambda_\theta} = M \hat{\theta}_{ML} + (I - M) \bar{\theta}
$$

(4.37)

With $\Phi^T\Phi = I$, $M$ is given by (4.29)

$$
M = \frac{\lambda_\theta}{\lambda_\theta + \lambda_e} I
$$

(4.38)

and hence the risk of $\hat{\theta}_{\lambda_\theta}$ is given by

$$
R(\theta, \hat{\theta}_{\lambda_\theta}) = \text{Tr} \left[ M^2 \text{Cov} \left[ \hat{\theta}_{ML} \right] \right] + (\theta - \bar{\theta})^T (I - M)^2 (\theta - \bar{\theta})
$$

$$
= \lambda_e \text{Tr} \left[ M^2 \right] + (\theta - \bar{\theta})^T (I - M)^2 (\theta - \bar{\theta})
$$

$$
= n \frac{\lambda_e \lambda_\theta^2}{(\lambda_\theta + \lambda_e)^2} + \frac{\lambda_e^2}{(\lambda_\theta + \lambda_e)^2} \|\theta - \bar{\theta}\|^2
$$

(4.39)

It is straightforward to minimize the risk $R(\theta_0, \hat{\theta}_{\lambda_\theta})$ with respect to $\lambda_\theta$ and the minimizer is given by

$$
\lambda_\theta^*(\theta_0) = \frac{1}{n} \|\theta_0 - \bar{\theta}\|^2
$$

which gives the optimal estimator

$$
\hat{\theta}^* = \frac{1}{n} \frac{\|\theta_0 - \bar{\theta}\|^2}{\|\theta_0 - \theta\|^2 + \lambda_e} \hat{\theta}_{ML} + \frac{\lambda_e}{\frac{1}{n} \|\theta_0 - \theta\|^2 + \lambda_e} \bar{\theta}
$$

$$
= \left( 1 - \frac{\lambda_e}{\frac{1}{n} \|\theta_0 - \theta\|^2 + \lambda_e} \right) \hat{\theta}_{ML} + \frac{\lambda_e}{\frac{1}{n} \|\theta_0 - \theta\|^2 + \lambda_e} \bar{\theta}
$$
From Theorem 4.7.1 it follows that the ML estimate of the risk \( R(\theta_o, \hat{\theta}_{\lambda_e}) \) is given by

\[
\hat{R}_{\text{ML}}(\lambda_e) = n \frac{\lambda_e^2}{(\lambda_e + \lambda_e)^2} + \frac{\lambda_e^2}{(\lambda_e + \lambda_e)^2} \| \hat{\theta}_{\text{ML}} - \bar{\theta} \|^2,
\]

the ML-estimate of \( \lambda^*_e(\theta) \) is given by

\[
\hat{\lambda}^*_{\theta,\text{ML}} := \frac{1}{n} \| \hat{\theta}_{\text{ML}} - \bar{\theta} \|^2 \tag{4.40}
\]

and the ML-estimate of the optimal estimator is given by

\[
\hat{\theta}^*_{\text{ML}} = \left( 1 - \frac{1}{n} \| \hat{\theta}_{\text{ML}} - \bar{\theta} \|^2 + \lambda_e \right) \hat{\theta}_{\text{ML}} + \frac{\lambda_e}{\| \hat{\theta}_{\text{ML}} - \bar{\theta} \|^2 + \lambda_e} \bar{\theta}.
\]

We recognize that this estimator is similar to, yet different from, (4.30). We stress that while we in both cases estimated \( \lambda_e \) referring to ML, the approaches were completely different. In the Empirical Bayes setting we marginalized out \( \theta \) and considered the likelihood of \( \lambda_e \), while in the pointwise risk estimation approach we used the ML estimate of \( \theta \) to obtain the ML estimate of \( \lambda_e \).

**Unbiased Risk Estimation**

Other estimators of the risk than the ML-estimator can be used. Typically, as in the example above, using the ML-estimator leads to a biased estimate of the risk. Stein proposed the use of an unbiased estimator [14] when estimating the mean of a multivariate normal distribution, leading to what is known as SURE (Stein’s Unbiased Risk Estimate). Consider the normal distribution \( N(\hat{\theta}, \Sigma_{XY}) \)

\[
p_Y(y; \theta) = ce^{-\frac{1}{2}(y - \theta)^T \Sigma_{YY}^{-1}(y - \theta)} = e^{y^T \Sigma_{XX}^{-1} \theta - \frac{1}{2} \theta^T \Sigma_{XX}^{-1} \theta} h(y), \text{ where}
\]

\[
h(y) = ce^{-\frac{1}{2}y^T \Sigma_{YY}^{-1} y}
\]

The risk of an estimator \( \hat{\theta} = \hat{\theta}(Y) \) when the quadratic is loss is given by (expectation is with respect to \( p_Y(y; \theta_o) \))

\[
R(\theta_o, \hat{\theta}) = \mathbb{E} \left[ || \hat{\theta} - \theta_o ||^2 \right] = \text{Tr} \left[ \mathbb{E} \left[ \hat{\theta}\hat{\theta}^T \right] - \mathbb{E} \left[ \hat{\theta} \right] \theta_o^T - \theta_o \mathbb{E} \left[ \hat{\theta}^T \right] + \theta_o \theta_o^T \right]
\]

\[
= \text{Tr} \left[ \mathbb{E} \left[ (\hat{\theta} - Y) (\hat{\theta} - Y)^T \right] + \mathbb{E} \left[ \hat{\theta} Y^T \right] + \mathbb{E} \left[ Y \hat{\theta}^T \right] - \mathbb{E} \left[ \theta Y^T \right] - \mathbb{E} \left[ \theta \right] \theta_o^T - \theta_o \mathbb{E} \left[ \theta^T \right] + \theta_o \theta_o^T \right]
\]

\[
= \text{Tr} \left[ \mathbb{E} \left[ (\hat{\theta} - Y) (\hat{\theta} - Y)^T \right] + \mathbb{E} \left[ \hat{\theta} (Y - \theta_o)^T \right] + \mathbb{E} \left[ (Y - \theta_o) \hat{\theta}^T \right] - \Sigma_{YY} \right]
\]

Now it holds that \( \frac{d}{d \theta} \log h(Y) = -\Sigma_{YY}^{-1} y \) so Stein’s identity (Lemma 4.41) gives

\[
\mathbb{E} \left[ g'(Y) \right] = -\mathbb{E} \left[ (-\Sigma_{YY}^{-1} Y + \Sigma_{YY}^{-1} \theta_o) g(Y) \right], \text{ i.e. } \Sigma \mathbb{E} \left[ g'(Y) \right] = \mathbb{E} \left[ (Y - \theta_o) g(Y) \right] \tag{4.41}
\]

Taking \( g(Y) = \hat{\theta}(Y) \) gives

\[
R(\theta_o, \hat{\theta}) = \text{Tr} \left[ \mathbb{E} \left[ (\hat{\theta}(Y) - Y) (\hat{\theta}(Y) - Y)^T \right] + \mathbb{E} \left[ \hat{\theta}'(Y) \right]^T \Sigma_{YY} + \Sigma_{YY} \mathbb{E} \left[ \hat{\theta}'(Y) \right] \right] - \Sigma_{YY}
\]
4.7. POINTWISE RISK ESTIMATION

The reason for this exercise is that if we now remove the expectation from the right-hand side, we obtain an unbiased estimate of the risk:

\[
\hat{R}_{SURE}(\hat{\theta}(Y)) := \text{Tr} \left[ (\hat{\theta}(Y) - Y)^T (\hat{\theta}(Y) - Y) + \theta'(Y)^T \Sigma_{YY} + \Sigma_{YY} \hat{\theta}'(Y) - \Sigma_{YY} \right]
\]

\[
= \|\hat{\theta}(Y) - Y\|^2 + 2\text{Tr} \left[ \Sigma_{YY} \hat{\theta}'(Y) \right] - \text{Tr} [\Sigma_{YY}]
\]

(4.42)

**Example 4.7.2** (Example ?? continued). With \(\Phi^T \Phi = I, M\) is given by (4.29) and the Bayes estimator given by (4.37) can be written

\[
\hat{\theta}_{\lambda_\theta} = \frac{\lambda_\theta}{\lambda_\theta^\gamma + \lambda_e^\gamma} \hat{\theta}_{\text{ML}} + \frac{\lambda_e}{\lambda_\theta^\gamma + \lambda_e^\gamma} \bar{\theta}
\]

so that

\[
\frac{d}{d\hat{\theta}_{\text{ML}}} \hat{\theta}_{\lambda_\theta} = \frac{\lambda_\theta}{\lambda_\theta^\gamma + \lambda_e^\gamma} - I
\]

which gives that the SURE estimate of the risk is

\[
\hat{R}_{SURE}(\lambda_\theta) = \|\hat{\theta}_{\lambda_\theta} - \hat{\theta}_{\text{ML}}\|^2 + 2\text{Tr} \left[ \frac{\lambda_\theta}{\lambda_\theta^\gamma + \lambda_e^\gamma} I \right] - \text{Tr} [\lambda_e I] = \|\hat{\theta}_{\lambda_\theta} - \hat{\theta}_{\text{ML}}\|^2 + 2n \frac{\lambda_\theta \lambda_e}{\lambda_\theta^\gamma + \lambda_e^\gamma} - n \lambda_e
\]

\[
= \|\hat{\theta}_{\lambda_\theta} - \hat{\theta}_{\text{ML}}\|^2 + n \lambda_e \frac{\lambda_\theta - \lambda_e}{\lambda_\theta^\gamma + \lambda_e^\gamma} + \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2 + n \lambda_e \frac{\lambda_\theta - \lambda_e}{\lambda_\theta^\gamma + \lambda_e^\gamma}
\]

(4.43)

The minimizer is given by

\[
\hat{\lambda}_{\theta, \text{SURE}} = \frac{1}{n} \|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2 - \lambda_e
\]

which is nothing but the Empirical Bayes estimate (4.28)!

We do not need to use Stein’s identity to derive the SURE estimate when the estimator is affine in the data

\[
\hat{\theta} = MY + (I - M) \bar{\theta}
\]

As in (4.39),

\[
R(\theta_o, \hat{\theta}) = \text{Tr} \left[ M \Sigma_{YY} M^T \right] + \text{Tr} \left[ (I - M)(\theta_o - \bar{\theta})(\theta_o - \bar{\theta})^T (I - M)^T \right]
\]

For the second term, replacing the unknown \(\theta_o\) with the ML-estimate \(\hat{\theta}_{\text{ML}} = Y\) (recall that \(Y \sim N(\theta, \Sigma_{YY})\) so the data itself is the ML-estimate) and taking expectation gives

\[
\text{Tr} \left[ (I - M) \mathbb{E} \left[ (Y - \hat{\theta})(Y - \hat{\theta})^T \right] (I - M)^T \right]
\]

\[
= \text{Tr} \left[ (I - M)(\theta_o - \bar{\theta})(\theta_o - \bar{\theta})^T (I - M)^T \right] + \text{Tr} \left[ (I - M) \Sigma_{YY} (I - M)^T \right]
\]

The second term is a bias so if we subtract this term we obtain the unbiased estimate of \(R(\theta_o, \hat{\theta})\)

\[
\text{Tr} \left[ M \Sigma_{YY} M^T \right] + \text{Tr} \left[ (I - M)(Y - \hat{\theta})(Y - \hat{\theta})^T (I - M)^T \right] - \text{Tr} \left[ (I - M) \Sigma_{YY} (I - M)^T \right]
\]

\[
= \|\hat{\theta} - \bar{\theta}\|^2 + 2\text{Tr} [\Sigma_{YY} M] - \text{Tr} [\Sigma_{YY}]
\]

which is exactly the SURE estimate (4.42).
4.8 Rapprochement Between Empirical Bayes and Pointwise Risk Estimation

We have seen above that there are strong similarities between Empirical Bayes and pointwise risk estimation. It is possible to make the link even closer. In certain cases, one can interpret Empirical Bayes as a special case of pointwise risk estimation.

Example 4.8.1 (Example 4.7.2 continued). Recall the estimator

\[ \hat{\theta}_{\lambda_0} = M\hat{\theta}_{\text{ML}} + (I - M)\bar{\theta} \]

with risk (4.39)

\[ R(\theta, \hat{\theta}_{\lambda_0}) = \lambda_e \text{Tr} \left[ M^2 \right] + (\theta - \bar{\theta})^T (I - M)^2 (\theta - \bar{\theta}) \]

Seen as a function of \( M \) and \( G = (\theta - \bar{\theta})(\theta - \bar{\theta})^T \) we can write the risk as

\[ W(M, G) := \lambda_e \text{Tr} \left[ M^2 \right] + \text{Tr} \left[ (I - M)^2 G \right] \]

By completing the square we see that

\[ W(M, G) \geq W(M^*(G, G), G) \]

where

\[ M^*(\hat{G}, G)) = \hat{G} (\lambda_e + G)^{-1} \]

i.e. with \( G \) given, the optimal choice of \( M \) is \( M = G \). In pointwise risk estimation we estimate \( G \), with \( G \) restricted to some family \( \mathcal{G} \), and then optimize \( M \), restricted to some family \( \mathcal{M} \). In our examples we have used

\[ M = M(\lambda_0) := \frac{\lambda_0}{\lambda_0 + \lambda_e} I \quad (4.45) \]

Now, from the reasoning above, if \( \mathcal{M} = \mathcal{G} \) then the second step in the pointwise risk minimization procedure in Section 4.7 gives the optimal \( M \) as \( M = M^*(\hat{G}, \hat{G}) \), where \( \hat{G} \) is the estimate of \( G \). In other words the optimization step has become trivial. The resulting estimator is

\[ \hat{\theta}_{\lambda_0} = \hat{G}\hat{\theta}_{\text{ML}} + (I - \hat{G})\bar{\theta} \]

In ML risk estimation we used \( \hat{\theta}_{\text{ML}} \) as estimate of \( \theta_o \) which corresponds to using

\[ \hat{G} = \hat{G}_{\text{ML}} := (\hat{\theta}_{\lambda_0} - \bar{\theta})(\hat{\theta}_{\lambda_0} - \bar{\theta})^T \]

Using this as \( M \) leads to very bad performance. The reason is that we have now estimated as many hyperparameters as there are data samples (recall that the “data” is \( \hat{\theta}_{\text{ML}} \))^2

However, there are other possibilities for how to choose \( \hat{G} \). For example, we can see \( \hat{G}_{\text{ML}} \) as “data” and fit an auxiliary model to this data. Now

\[ \mathbb{E} \left[ \hat{G}_{\text{ML}} \right] = (\theta_o - \bar{\theta})(\theta_o - \bar{\theta})^T + \lambda_e I \]

\(^2\)Compare this with (4.45) where in case we use \( \hat{G}_{\text{ML}} \) as model of \( G \), these parameters are condensed into one in the optimization step ii), leading to improved statistical properties.
and hence we can write

\[ \hat{G}_{\text{ML}} = (\theta_o - \bar{\theta})(\theta_o - \bar{\theta})^T + \lambda_e I + E = G_o + \lambda_e I + E \]

where \( E \) has zero mean. From above we see that \( \hat{G}_{\text{ML}} \) is biased and replacing it by the unbiased estimate

\[ \hat{G}_{\text{ML}} - \lambda_e I \]

gives the SURE estimate (4.44) when (4.45) is used as \( M \).

Alternatively, we can use a parametric model \( G(\eta) \) for \( G_o \) and fit it to \( \hat{G}_{\text{ML}} \). There are many possibilities for \( G(\eta) \) but suitable for our purpose is to use

\[ G(\eta) = \eta I \]  
(4.46)

which gives

\[ M^*(G(\eta), G(\eta)) = \frac{\eta}{\eta + \lambda_e} I \]

which has the same structure as (4.45). Thus, if we restrict \( M \) to have the structure given in (4.45), and if \( \hat{\eta} \) is an estimate of \( \eta \), pointwise risk estimation will result in

\[ M = M^*(G(\hat{\eta}), G(\hat{\eta})) = \frac{\hat{\eta}}{\hat{\eta} + \lambda_e} I \]  
(4.47)

The model (4.46) may seem bizarre. For example, \( G_o \) is rank one while \( G(\eta) \) is full rank. However, recall that biased models may be good from a risk perspective. In this step we are interested in a model that has low risk when it comes to estimating \( G_o \). We would thus like to estimate \( \eta \) in the model

\[ \hat{G}_{\text{ML}} = \eta I + \lambda_e I + E = (\eta + \lambda_e)I + E \]

where \( \mathbb{E}[E] = 0 \). Data is here in matrix form and there exists many different estimators for \( \eta \). Disregarding the noise term, \( \hat{G}_{\text{ML}}((\eta + \lambda_e)I)^{-1} = I \). To turn this idea into an optimization problem we need the following elementary result.

**Lemma 4.8.1.** It holds that

\[ n = \min_{0 < M = M^T \in \mathbb{R}^{n \times n}} \text{Tr}\{M\} + \log \det M^{-1} \]  
(4.48)

The unique minimizer is \( M = I \).

**Proof.** See Appendix 4.B. \( \square \)

Lemma 4.8.1 suggests that we can estimate \( \theta \) by minimizing

\[ = \text{Tr} \left[ \hat{G}_{\text{ML}} ((\eta + \lambda_e)I)^{-1} \right] + \log \det \left( ((\eta + \lambda_e)I) \hat{G}_{\text{ML}}^{-1} \right) \]

\[ = (\hat{\theta}_{\text{ML}} - \bar{\theta})^T ((\eta + \lambda_e)I)^{-1} (\hat{\theta}_{\text{ML}} - \bar{\theta}) + \log \det ((\eta + \lambda_e)I) - \log \det \left( \hat{G}_{\text{ML}} \right) \]

The last term is \( \eta \) independent and can be dropped, leading to the criterion

\[ J(\eta) = \frac{\|\hat{\theta}_{\text{ML}} - \bar{\theta}\|^2}{\eta + \lambda_e} + n \log(\eta + \lambda_e) \]  
(4.49)
CHAPTER 4. BIASED ESTIMATORS

but with $\eta = \lambda_\theta$ we recognize this as the marginalized likelihood criterion (4.27) that was used to obtain the Empirical Bayes estimator $\hat{\lambda}_{BM,ML}$. But then we see that the estimator resulting from pointwise risk estimation (which uses (4.47)) is identical to the Empirical Bayes estimator.

Thus, using the variance of the weighting function $f_\Theta$ as model of $G_\Theta$ and then using the trace-logdet criterion (4.49) to estimate this model results in that the pointwise risk estimation procedure becomes identical to the Empirical Bayes procedure when the variance is estimated using marginalized ML. The reason is that the optimization step (ii) in pointwise risk estimation becomes trivial in the setting we considered.

Thus, for the estimation problem considered in this example, Empirical Bayes can be derived from a frequentistic perspective as pointwise risk estimation. □

4.9 Asymptotics of Biased Estimators

4.9.1 Asymptotic Efficiency of the Bayes Estimator

Example 4.9.1 (Example 4.5.4 continued). Suppose that

$$R_N := \frac{1}{N} \Phi^T \Phi \to R > 0$$

in Example 4.5.4. Then we see that

$$M = \left( \frac{1}{\lambda_\epsilon} \Phi^T \Phi + \frac{1}{\lambda_\theta} I \right)^{-1} \frac{1}{\lambda_\epsilon} \Phi^T \Phi = \left( \frac{1}{\lambda_\epsilon} NR_N + \frac{1}{\lambda_\theta} I \right)^{-1} \frac{1}{\lambda_\epsilon} NR_N = \left( I + \frac{\lambda_\epsilon}{N\lambda_\theta} R_N^{-1} \right)^{-1}$$

and hence $\hat{\theta}_{f_\Theta} = \hat{\theta}_{ML,N} + \delta$ where $\delta$ is of order $1/N$. This indicates that the influence of the “prior” $f_\Theta$ vanishes as $N \to \infty$. Furthermore, since $\hat{\theta}_{ML,N}$ is consistent and asymptotically efficient $\hat{\theta}_{f_\Theta}$ will be consistent and asymptotically efficient. □

For iid observations, the result in Example 4.9.1 generalizes under the assumptions that render the ML-estimator asymptotically efficient strengthened as follows.

Let $V_N$ denote the normalized log likelihood function

$$V_N(\theta) = \frac{1}{N} \sum_{k=1}^{N} \log p(y; \theta)$$

and consider the expansion

$$V_N(\theta) = V_N(\theta_o) + V_N'(\theta_o)(\theta - \theta_o) - \frac{1}{2}(\theta - \theta_o) \left( I_{F(\theta_o)} + \tilde{R}_N(\theta) \right)$$

Then for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathcal{P} \left( \left\{ \sup_{\|\theta - \theta_o\| \leq \delta} \| \tilde{R}_N(\theta) \| \geq \varepsilon \right\} \right) \to 0$$
Furthermore, for any $\delta > 0$ there exists $\varepsilon > 0$ such that
\[
P \left( \sup_{\|\theta - \theta_0\| \geq \delta} V_N(\theta) - V_N(\theta_0) \leq -\varepsilon \right) \rightarrow 0
\]
Regarding the weighting (“prior”) $f_{\Theta}$, it has to be continuous and positive on $D_{\theta}$, the domain of definition of $\theta$, and
\[
\int_{D_{\theta}} \|\theta\| f_{\Theta}(\theta) d\theta < \infty
\]

### 4.9.2 Excess MSE of Biased Estimators

We continue with our running example.

**Example 4.9.2 (Example 4.7.1 continued).** For simplicity we let $\bar{\theta} = 0$ and assume $\Phi^T \Phi = NI$, or equivalently that $\lambda_e = \lambda_e / N$ for some $\lambda_e > 0$.

The MSE of the ML estimate is then $n \lambda_e / N$. Let us compare the risk of $\hat{\theta}_{\lambda \theta}$ and the risk (MSE) of the ML estimate by taking the difference. Using (4.39) gives
\[
\frac{\lambda_e^2}{(\lambda_\theta + \lambda_e)^2} \|	heta\|^2 + n \frac{\lambda_e \lambda_\theta^2}{(\lambda_\theta + \lambda_e)^2} - n \lambda_e = \frac{\lambda_e^2}{(\lambda_\theta + \lambda_e)^2} \|	heta\|^2 + \frac{n \lambda_e}{(\lambda_\theta + \lambda_e)^2} \left( \lambda_\theta^2 - (\lambda_\theta + \lambda_e)^2 \right)
\]
\[
= \frac{\lambda_e^2}{(\lambda_\theta + \lambda_e)^2} \|	heta\|^2 - \frac{n \lambda_e^2}{(\lambda_\theta + \lambda_e)^2} \frac{2 \lambda_\theta + \lambda_e}{(\lambda_\theta + \lambda_e)^2}
\]

We see that the size of this difference is of the order $1/N^2$. Thus if we define the excess MSE of $\hat{\theta}_{\lambda \theta}$, let us call it $\text{XMSE} [\theta, \hat{\theta}_{\lambda \theta}]$, as the limit of this difference normalized by $N^2$ we get
\[
\text{XMSE} [\theta, \hat{\theta}_{\lambda \theta}] = \frac{\lambda_e^2}{\lambda_\theta^2} \|	heta\|^2 - \frac{2n \lambda_e^2}{\lambda_\theta}
\]

From the example we see that asymptotically the improvement over the ML-estimator is of the order $1/N$. Let us now establish the above type of result in a more general setting.

### The Excess Cramér-Rao Lower Bound

Suppose that the bias of an estimator based on sample size $N$ satisfies
\[
Nb_N(\theta) \rightarrow b(\theta), \quad \text{as } N \rightarrow \infty
\]
so that
\[
Nb_N(\theta) \rightarrow b'(\theta), \quad Nb'_N(\theta) \rightarrow b''(\theta) \quad \text{as } N \rightarrow \infty
\]
and that the FIM is $NI_F(\theta)$ where $I_F$ is the per sample information matrix (e.g. that we have iid observations). Then the biased CRLB (2.30) gives the following lower bound for the MSE
\[
\text{MSE} \left[ \theta, \hat{\theta}_N \right] \geq b_N(\theta)b_N(\theta)^T + \frac{1}{N} \left( I + b'_N(\theta) \right) I_F^{-1}(\theta) \left( I + b'_N(\theta) \right)^T
\]
and hence
\[
N^2 \left( \text{MSE} \left[ \theta, \hat{\theta}_N \right] - (NI_F(\theta))^{-1} \right) \\
\geq N^2 b_N(\theta)b_N(\theta)^T + b_N'(\theta)I_F^{-1}(\theta) + N b_N'(\theta)b_N'(\theta)^T \\
\rightarrow b(\theta)b(\theta)^T + b'(\theta)I_F^{-1}(\theta) + I_F^{-1}(\theta)b'(\theta)^T
\]
i.e. we have the following lower bound on the excess MSE
\[
\text{XMSE} \left[ \theta, \hat{\theta}_N \right] \geq b(\theta)b(\theta)^T + b'(\theta)I_F^{-1}(\theta) + I_F^{-1}(\theta)b'(\theta)^T
\] (4.52)

A General Expression for the Excess MSE

Now we turn to the asymptotic analysis of biased estimators. We will consider estimators that are functions of \( \hat{\theta}_{ML} \). We will also make the dependence on the sample size \( N \) explicit and for this it is convenient to introduce \( \gamma = 1/N \). To motivate the approach we take we return once more to our running example.

Example 4.9.3 (Example 4.5.4 continued). Suppose that
\[
\Phi^T \Phi = N \frac{1}{N} \Phi^T \Phi = N R_N
\]
where \( R_N \rightarrow R > 0 \). Taking \( \bar{\theta} = 0 \), then the estimator (4.25) can be expressed as
\[
\hat{\theta}_{f_0} = f_N(\hat{\theta}_{ML,N}, 1/N)
\]
where
\[
f_N(\hat{\theta}_{ML,N}, \gamma) := \left( I + \gamma \frac{\lambda_e}{\lambda_\theta} R_N^{-1} \right)^{-1} \hat{\theta}_{ML,N}
\]

Following this we will consider estimators that are given by \( \hat{\theta}_N = f_N(\hat{\theta}_{ML,N}, 1/N) \) where
\[
f_N(\hat{\theta}_{ML,N}, \gamma) \rightarrow f(\theta_o, \gamma), \quad \text{as } N \rightarrow \infty
\] (4.53)
and further that
\[
f_N(\theta, 0) = 0
\] (4.54)

The latter is just a constraint that when the information from data tends to infinity \( (N \rightarrow \infty \text{ or, equivalently } \gamma \rightarrow 0) \), the estimator is consistent. Condition (4.54) implies
\[
\frac{\partial f_N(\theta, 0)}{\partial \theta} = I
\] (4.55)
\[
\frac{\partial^2 f_N(\theta, 0)}{\partial \theta_j \partial \theta_i} = 0, \quad i, j = 1, \ldots, n
\] (4.56)
Using this when expanding \( f_N \) in \( \theta \) and \( \gamma \) around \( \theta = \theta_o \) (the true parameter) and \( \gamma = 0 \) gives

\[
f_N(\theta, \gamma) = f_N(\theta_o, 0) + \frac{\partial f_N(\theta_o, 0)}{\partial \theta}(\theta - \theta_o) + \frac{\partial f_N(\theta_o, 0)}{\partial \gamma} \gamma
\]

\[+ \frac{1}{2} \sum_{i,j} \frac{\partial^2 f_N(\theta_o, 0)}{\partial \theta_i \partial \theta_j} (\theta_i - \theta_o,i)(\theta_j - \theta_o,j)\]

\[+ \gamma \frac{\partial^2 f_N(\theta_o, 0)}{\partial \gamma \partial \theta^T}(\theta - \theta_o)\]

\[+ \frac{\gamma^2}{2} \frac{\partial^2 f_N(\theta_o, 0)}{\partial \gamma^2} + \text{Higher Order Terms}\]

\[= \theta + \gamma \frac{\partial f_N(\theta_o, 0)}{\partial \gamma} + \gamma \frac{\partial^2 f_N(\theta_o, 0)}{\partial \gamma \partial \theta^T}(\theta - \theta_o) + \text{Higher Order Terms}\]

giving

\[
f_N(\hat{\theta}_{ML,N}, \gamma) - \theta_o = \gamma \frac{\partial f_N(\theta_o, 0)}{\partial \gamma} + \left( I + \gamma \frac{\partial^2 f_N(\theta_o, 0)}{\partial \gamma \partial \theta^T} \right)(\hat{\theta}_{ML,N} - \theta_o) + \text{H.O.T} \quad (4.57)
\]

from which the excess MSE can be obtained

\[
\text{XMSE} \left[ \theta, f_N(\hat{\theta}_{ML,N}, \gamma) \right] = \frac{\partial f(\theta_o, 0)}{\partial \gamma} \left( \frac{\partial f(\theta_o, 0)}{\partial \gamma} \right)^T + \frac{\partial^2 f(\theta_o, 0)}{\partial \gamma \partial \theta^T} I_F^{-1}(\theta) + I_F^{-1}(\theta) \frac{\partial^2 f(\theta_o, 0)}{\partial \gamma \partial \theta^T} \quad (4.58)
\]

Notice now that the bias is given by

\[
b_N(\theta) = \mathbb{E} \left[ f_N(\hat{\theta}_{ML,N}, \gamma) \right] - \theta_o = \gamma \frac{\partial f_N(\theta_o, 0)}{\partial \gamma} + \mathbb{E} [\text{H.O.T}] \]

so that

\[
\frac{1}{\gamma} b_N(\theta) \to \frac{\partial f(\theta_o, 0)}{\partial \gamma}
\]

and

\[
\frac{1}{\gamma} b'_N(\theta) \to \frac{\partial^2 f(\theta_o, 0)}{\partial \gamma \partial \theta^T}
\]

Comparing with (4.50) and (4.51), we can identify \( \frac{\partial f(\theta_o, 0)}{\partial \gamma} \) with \( b(\theta) \) and \( \frac{\partial^2 f(\theta_o, 0)}{\partial \gamma \partial \theta^T} \) with \( b'(\theta) \), and hence (4.58) shows that \( f_N(\hat{\theta}_{ML,N}, 1/N) \) reaches the CRLB (4.52).

**Reaching the Excess CRLB**

Suppose now that a given estimator has bias \( b_N(\theta) \) satisfying (4.50). Then we can construct a new estimator based on the ML-estimator which reaches the CRLB (4.52) for the excess MSE. Consider

\[
\hat{\theta}_N^* := \hat{\theta}_{ML,N} + b_N(\hat{\theta}_{ML,N})
\]
which we expand as
\[
\hat{\theta}_N = \theta + b_N(\theta) + (I + b'_N(\theta))(\hat{\theta}_{ML,N} - \theta) + \sum_{i,j} \frac{\partial^2 b_N(\theta)}{\partial \theta_i \partial \theta_j} \bigg|_{\theta = \xi_N} (\hat{\theta}_{ML,N,i} - \theta_i)(\hat{\theta}_{ML,N,j} - \theta_j)
\]
for some \(\xi_N\) between \(\theta\) and \(\hat{\theta}_{ML,N}\). We thus get
\[
\mathbb{E} \left[ (\hat{\theta}_N^* - \theta)(\hat{\theta}_N - \theta)^T \right] = b_N(\theta)b_N^T + (I + b'_N(\theta))\text{Cov} \left[ \hat{\theta}_{ML,N} \right] (I + b'_N(\theta))^T + \Delta_N(\theta)
\]
where the remainder term \(\Delta_N(\theta)\) is given by
\[
\Delta_N(\theta) = \Delta_{N,1}(\theta) + \Delta_{N,2}(\theta) + \Delta_{N,3}(\theta)
\]
where
\[
\Delta_{N,1}(\theta) = \sum_{i,j,k,l} \mathbb{E} \left[ \frac{\partial^2 b_N(\theta)}{\partial \theta_i \partial \theta_j} \bigg|_{\theta = \xi_N} \left( \frac{\partial^2 b_N(\theta)}{\partial \theta_k \partial \theta_l} \bigg|_{\theta = \xi_N} (\hat{\theta}_{ML,N,i} - \theta_i)(\hat{\theta}_{ML,N,j} - \theta_j)(\hat{\theta}_{ML,N,k} - \theta_k)(\hat{\theta}_{ML,N,l} - \theta_l) \right)^T \right]
\]
\[
\Delta_{N,2}(\theta) = \sum_{i,j} \mathbb{E} \left[ \frac{\partial^2 b_N(\theta)}{\partial \theta_i \partial \theta_j} \bigg|_{\theta = \xi_N} (\hat{\theta}_{ML,N,i} - \theta_i)(\hat{\theta}_{ML,N,j} - \theta_j) b_N^T(\theta) \right]
\]
Assuming that (4.50) is uniform in \(\theta\) it follows that \(\Delta_N = O \left( \frac{1}{\sqrt{N}} \right) \) and hence
\[
N^2 \left( \mathbb{E} \left[ (\hat{\theta}_N^* - \theta)(\hat{\theta}_N - \theta)^T \right] - (NI_F(\theta))^{-1} \right) = N^2b_N(\theta)b_N^T + b'_N(\theta)I_F^{-1}(\theta)(b'_N(\theta))^T + N^2b''_N(\theta)I_F^{-1}(\theta)(b'_N(\theta))^T + N^2\Delta_N
\]
i.e. \(\hat{\theta}_N^*\) has the smallest possible excess MSE among estimators with bias \(\{b_N(\theta)\}\). For this reason we can limit the asymptotic analysis of biased estimators satisfying (4.50) to those that are functions of the ML-estimator.

**Excess MSE of the Bayes Estimator**

We will now use the results in the previous sections to analyze the asymptotic behaviour of the Bayes estimator in more detail. Recall that the Bayes estimator is given by
\[
\hat{\theta}_{f\varphi,N} = \frac{1}{f_Y(Y^N)} \int \theta p_Y(Y^N; \theta) f\varphi(\theta) d\theta
\]
where
\[
f_Y(Y; \eta) := \int p_Y(Y^N; \theta) f\varphi(\theta; \eta) d\theta
\]
Assuming, as usual, that the observations are independent and expanding \(\log p_Y(Y^N; \theta)\) around \(\hat{\theta}_{ML,N}\) gives
\[
\log p_Y(Y^N; \theta) = \sum_{k=1}^{N} \log p_Y(Y_k; \theta) \quad \text{(4.59)}
\]
\[
\approx \sum_{k=1}^{N} \log p_Y(Y_k; \hat{\theta}_{ML,N}) - \frac{1}{2} \left( \theta - \hat{\theta}_{ML,N} \right)^T N \hat{I}_N(\hat{\theta}_{ML,N}) \left( \theta - \hat{\theta}_{ML,N} \right) \quad \text{(4.60)}
\]
where

\[ \hat{I}_N(\hat{\theta}_{ML,N}) = -\frac{\partial^2}{\partial \theta \partial \theta^T} \frac{1}{N} \sum_{k=1}^{N} \log p_Y(Y_k; \theta) \]

gives

\[
f_Y(Y) \approx p_Y(Y^N; \hat{\theta}_{ML,N}) \int f_\Theta(\theta) e^{-\frac{1}{2}(\theta - \hat{\theta}_{ML,N})^T \hat{I}_N(\hat{\theta}_{ML,N})(\theta - \hat{\theta}_{ML,N})} d\theta \tag{4.61}
\]

\[
= p_Y(Y^N; \hat{\theta}_{ML,N}) (2\pi)^{n/2} \det^{1/2} \left( N \hat{I}_N(\hat{\theta}_{ML,N}) \right) \mathbb{E}[f_\Theta(\Theta)] \tag{4.62}
\]

where \( \Theta \) has distribution \( N(\hat{\theta}_{ML,N}, \frac{1}{N} \hat{I}_N^{-1}(\hat{\theta}_{ML,N})) \). Similarly

\[
\int \theta p_Y(Y^N; \hat{\theta}_{ML,N}) f_\Theta(\theta) d\theta \approx p_Y(Y^N; \hat{\theta}_{ML,N}) (2\pi)^{n/2} \det^{1/2} \left( N \hat{I}_N(\hat{\theta}_{ML,N}) \right) \mathbb{E}[\Theta f_\Theta(\Theta)]
\]

Thus

\[ \hat{\theta}_{f\Theta,N} \approx \frac{\mathbb{E}[\Theta f_\Theta(\Theta)]}{\mathbb{E}[f_\Theta(\Theta)]} \tag{4.63} \]

Now as the sample size \( N \) grows, the distribution \( N(\hat{\theta}_{ML,N}, \frac{1}{N} \hat{I}_N^{-1}(\hat{\theta}_{ML,N})) \) becomes more and more peaked around \( \hat{\theta}_{ML,N} \), which in turn moves to \( \theta_0 \). To compute the means above we can thus use Taylor expansions around \( \hat{\theta}_{ML,N} \):

\[
f_\Theta(\theta) \approx f_\Theta(\hat{\theta}_{ML,N}) + f_\Theta'(\hat{\theta}_{ML,N})(\theta - \hat{\theta}_{ML,N}) + \frac{1}{2}(\theta - \hat{\theta}_{ML,N})^T f_\Theta''(\hat{\theta}_{ML,N})(\theta - \hat{\theta}_{ML,N}) \tag{4.64}
\]

\[
\theta f_\Theta(\theta) \approx \theta f_\Theta(\hat{\theta}_{ML,N}) + (\theta f_\Theta'(\hat{\theta}_{ML,N}) + \hat{\theta}_{ML,N}(f_\Theta'(\hat{\theta}_{ML,N})))^T(\theta - \hat{\theta}_{ML,N})
\]

\[ + \frac{1}{2}(\theta - \hat{\theta}_{ML,N})^T \left( 2f_\Theta'(\hat{\theta}_{ML,N}) + f_\Theta''(\hat{\theta}_{ML,N})\hat{\theta}_{ML,N} \right)(\theta - \hat{\theta}_{ML,N}) \tag{4.65}
\]

(the last term is not correctly expressed) which gives

\[
\mathbb{E}[f_\Theta(\Theta)] \approx f_\Theta(\hat{\theta}_{ML,N}) + \frac{1}{2N} \text{Tr} \left[ \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta''(\hat{\theta}_{ML,N}) \right]
\]

\[
\mathbb{E}[\Theta f_\Theta(\Theta)] \approx f_\Theta(\hat{\theta}_{ML,N})\hat{\theta}_{ML,N} + \frac{1}{2N} \left( 2\hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta'(\hat{\theta}_{ML,N}) + \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta''(\hat{\theta}_{ML,N})\hat{\theta}_{ML,N} \right)
\]

and hence

\[
\hat{\theta}_{f\Theta,N} \approx \frac{f_\Theta(\hat{\theta}_{ML,N})\hat{\theta}_{ML,N} + \frac{1}{2N} \left( 2\hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta'(\hat{\theta}_{ML,N}) + \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta''(\hat{\theta}_{ML,N})\hat{\theta}_{ML,N} \right)}{f_\Theta(\hat{\theta}_{ML,N}) + \frac{1}{2N} \text{Tr} \left[ \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta''(\hat{\theta}_{ML,N}) \right]}
\]

\[
= \frac{\hat{\theta}_{ML,N} + \frac{1}{2N f_\Theta(\hat{\theta}_{ML,N})} \left( 2\hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta'(\hat{\theta}_{ML,N}) + \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta''(\hat{\theta}_{ML,N})\hat{\theta}_{ML,N} \right)}{1 + \frac{1}{2N f_\Theta(\hat{\theta}_{ML,N})} \text{Tr} \left[ \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta''(\hat{\theta}_{ML,N}) \right]}
\]

\[
\approx \left( \hat{\theta}_{ML,N} + \frac{1}{2N f_\Theta(\hat{\theta}_{ML,N})} \left( 2\hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta'(\hat{\theta}_{ML,N}) + \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta''(\hat{\theta}_{ML,N})\hat{\theta}_{ML,N} \right) \right)
\]

\[
\left( 1 - \frac{1}{2N f_\Theta(\hat{\theta}_{ML,N})} \text{Tr} \left[ \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta''(\hat{\theta}_{ML,N}) \right] \right)
\]

\[
\approx \hat{\theta}_{ML,N} + \frac{1}{N f_\Theta(\hat{\theta}_{ML,N})} \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) f_\Theta'(\hat{\theta}_{ML,N}) = \hat{\theta}_{ML,N} + \frac{1}{N} \hat{I}_N^{-1}(\hat{\theta}_{ML,N}) \frac{d}{d\theta} \log f_\Theta(\hat{\theta}_{ML,N})
\]
where the Bayes estimator can be written on the form \( f_N(\hat{\theta}_{ML,N}, \gamma) \) which has limit
\[
\lim_{\gamma \to \infty} f_N(\hat{\theta}_{ML,N}, \gamma) = \theta_0 + \gamma I_F^{-1}(\theta_0) \frac{\partial}{\partial \theta} \log f_\Theta(\theta) \bigg|_{\theta = \theta_0}, \quad \text{as } N \to \infty
\]  
(4.66)

For the Bayes estimator, the limit function in (4.53) is thus
\[
f_{f_\Theta}(\theta, \gamma) := \theta + \gamma I_F^{-1}(\theta) \frac{\partial}{\partial \theta} \log f_\Theta(\theta)
\]
(4.67)

for which we have
\[
\frac{\partial}{\partial \gamma} f_{f_\Theta}(\theta, \gamma) = I_F^{-1}(\theta) \frac{\partial}{\partial \theta} \log f_\Theta(\theta)
\]
(4.68)

\[
\frac{\partial^2}{\partial \gamma \theta^T} f_{f_\Theta}(\theta, \gamma) = I_F^{-1}(\theta) \frac{\partial^2}{\partial \theta \theta^T} \log f_\Theta(\theta) - I_F^{-1}(\theta) \left( \frac{d}{d \theta} I_F(\theta) \right) I_F^{-1}(\theta) \frac{\partial}{\partial \theta} \log f_\Theta(\theta)
\]
(4.69)

where the second term in \( \frac{\partial^2}{\partial \gamma \theta^T} f_{f_\Theta}(\theta, \gamma) \) has to be interpreted correctly (due to dimension issues). Assuming for simplicity that \( I_F \) is constant and using (4.58)
\[
\text{XMSE} \left[ \theta, \hat{\theta}_{f_\Theta,N} \right] = \frac{\partial f(\theta_0, 0)}{\partial \gamma} \left( \frac{\partial f(\theta_0, 0)}{\partial \gamma} \right)^T + \frac{\partial^2 f(\theta_0, 0)}{\partial \gamma \partial \theta^T} I_F^{-1}(\theta) + I_F^{-1}(\theta) \frac{\partial^2}{\partial \theta \gamma^T} I_F^{-1}(\theta)
\]

with \( f = f_{f_\Theta} \), the excess MSE can be expressed as
\[
\text{XMSE} \left[ \theta, \hat{\theta}_{f_\Theta,N} \right] = I_F^{-1} \left( \frac{\partial}{\partial \theta} \log f_\Theta(\theta_0) \frac{\partial}{\partial \theta} \log f_\Theta(\theta_0) + 2 \frac{\partial^2}{\partial \theta \theta^T} \log f_\Theta(\theta_0) \right) I_F^{-1}
\]
(4.70)

Since
\[
\frac{d^2}{d \theta d \theta^T} \log f_\Theta(\theta_0) = \frac{f''_\Theta(\theta_0)}{f_\Theta(\theta_0)} - \frac{d}{d \theta} \log f_\Theta(\theta_0) \frac{d}{d \theta^T} \log f_\Theta(\theta_0)
\]

we can also in this case write
\[
\text{XMSE} \left[ \theta, \hat{\theta}_{f_\Theta,N} \right] = I_F^{-1} \frac{f''_\Theta(\theta_0)}{f_\Theta(\theta_0)} I_F^{-1}
\]
(4.71)

Since \( f_\Theta(\theta) > 0 \), the Bayes estimator asymptotically improves upon the ML-estimator in the region where the second derivative of \( f_\Theta \) is negative.

Since the Bayes estimator minimizes the average risk, it also minimizes the average excess MSE
\[
\text{exr}(\hat{\theta}, f_\Theta) := \int \text{XMSE} \left[ \theta, \hat{\theta} \right] f_\Theta(\theta) d\theta
\]

For the case where the per sample information matrix \( I_F \) is constant, the minimum average excess MSE can be obtained from (4.85) as
\[
\text{exr}(\hat{\theta}_{f_\Theta}, f_\Theta) = \text{Tr} \left[ I_F^{-2} \int \left( \frac{d}{d \theta} \log f_\Theta(\theta_0) \frac{d}{d \theta^T} \log f_\Theta(\theta_0) + 2 \frac{d^2}{d \theta d \theta^T} \log f_\Theta(\theta_0) \right) f_\Theta(\theta) d\theta \right]
\]
\[
= - \text{Tr} \left[ I_F^{-2} I_F(\theta_0) \right] = - \text{Tr} \left[ I_F^{-1} I_P(\theta_0) I_F^{-1} \right]
\]
(4.72)

where \( I_F \) is the prior “information matrix”
\[
I_P(\theta) = \mathbb{E} \left[ \frac{d}{d \theta} \log f_\Theta(\theta) \frac{d}{d \theta^T} \log f_\Theta(\theta) \right] = - \mathbb{E} \left[ \frac{d^2}{d \theta d \theta^T} \log f_\Theta(\theta) \right]
\]

As expected, we see that the average excess MSE for the Bayes estimator is negative, i.e. the average risk of the Bayes estimator is lower than the average risk for the ML-estimator.
4.9.3 Asymptotics for Hyperparameter Tuning

Asymptotically Optimal Hyperparameters

Suppose now that our estimator is parametrized by some hyperparameters $\eta$: $\hat{\theta}_{N,\eta} = f_N(\hat{\theta}_{ML,N}, 1/N, \eta)$. Then the excess MSE will depend on $\eta$ and we can define the asymptotically optimal hyperparameter

$$\eta^*(\theta) := \arg\min_{\eta} \text{Tr} \left[ \text{XMSE} \left[ \theta, f_N(\hat{\theta}_{ML,N}, 1/N, \eta) \right] \right]$$

(4.73)

which under the assumptions above can be written

$$\eta^*(\theta) := \arg\min_{\eta} \| b(\theta, \eta) \|^2 + 2 \text{Tr} \left[ \frac{\partial b(\theta)}{\partial \theta} I_F^{-1}(\theta) \right]$$

Example 4.9.4 (Example 4.9.2 continued). In Example 4.9.2 the excess MSE was given by

$$\text{XMSE} \left[ \theta, \hat{\theta}_{\lambda_0} \right] = \frac{\lambda^2_0}{\lambda_0^2} \| \theta \|^2 - \frac{2n\lambda^2_0}{\lambda_0}$$

Regarding $\lambda_0$ as a hyperparameter we see that it is minimized by $\lambda_0 = \lambda^*_\theta := \frac{1}{n} \| \theta \|^2$ giving the minimum excess MSE

$$\text{XMSE}^*[\theta, \hat{\theta}_{\lambda_0}] = -\frac{n^2\lambda^2_0}{\| \theta \|^2}$$

In previous sections we have discussed different ways to determine the hyperparameters. An interesting question is which methods will recover the asymptotically optimal hyperparameter when the sample size grows.

The ML and SURE Estimators

The ML-estimator is given by (4.35) which, indicating the sample size dependency, is given by

$$\hat{\eta}_{ML,N} = \hat{\eta}_{ML,N}(Y^N) = \eta_N(\hat{\theta}_{ML})$$

(4.74)

where

$$\eta_N(\theta) = \arg\min_{\eta \in D_\eta} R(\theta, \hat{\theta}_\eta(Y^N))$$

(4.75)

For the MSE, with the estimator expressed as $f_N$, $\hat{\eta}_{ML,N}$ can be written

$$\hat{\eta}_{ML,N} = \arg\min_{\eta \in D_\eta} \text{MSE} \left[ \theta_o, f_N(\hat{\theta}_{ML,N}, 1/N, \eta) \right] \big|_{\theta_o = \hat{\theta}_{ML}}$$

Analysing the behaviour of $\hat{\eta}_{ML,N}$ is the same type of analysis as we did in Section 4.5.10. Asymptotically we have seen that

$$\text{MSE} \left[ \theta_o, \hat{\theta}_{N,\eta} \right] \big|_{\theta_o = \hat{\theta}_{ML}} = \frac{1}{N} I_F(\hat{\theta}_{ML,N}) + \frac{1}{N^2} \text{XMSE} \left[ \theta_o, f_N(\hat{\theta}_{ML,N}, 1/N, \eta) \right] \big|_{\theta_o = \hat{\theta}_{ML}}$$
and thus, asymptotically, minimizing $\text{MSE} \left[ \theta_o, \hat{\theta}_{N, \eta} \right]$ with respect to $\eta$ is equivalent to minimizing $\text{XMSE} \left[ \theta_o, f_N(\hat{\theta}_{\text{ML}, N}, 1/N, \eta) \right] |_{\theta_o = \hat{\theta}_{\text{ML}}}$. Now,

$$\hat{\theta}_{\text{ML}, N} \overset{w.p.1}{\to} \theta_o$$

and assuming $f$ and its derivatives are smooth functions and that $D_\eta$ is compact one can establish uniform convergence of the cost function

$$\lim_{N \to \infty} \sup_{\eta \in D_\eta} \left| \text{XMSE} \left[ \theta_o, f_N(\hat{\theta}_{\text{ML}, N}, 1/N, \eta) \right] \right|_{\theta_o = \hat{\theta}_{\text{ML}}} = 0$$

and hence

$$\hat{\eta}_{\text{ML}, N} \to \eta^*(\theta_o)$$

The same type of analysis can be performed for the SURE estimate.

The Empirical Bayes Estimator

Recall that the Empirical Bayes estimator is based on minimizing the marginalized likelihood for $Y$

$$f_Y(Y; \eta) := \int p_Y(Y^N; \theta) f_\Theta(\theta; \eta) d\theta$$

From (4.61)

$$f_Y(Y; \eta) \approx p_Y(Y^N; \hat{\theta}_{\text{ML}, N}) (2\pi)^{N/2} \det^{1/2} \left( N \hat{I}_F(\hat{\theta}_{\text{ML}, N}) \right) \mathbb{E} \left[ f_\Theta(\Theta; \eta) \right]$$

where $\Theta$ has distribution $N(\hat{\Theta}_{\text{ML}, N}, \frac{1}{N} \hat{I}_F^{-1}(\hat{\Theta}_{\text{ML}, N}))$. Now as the sample size $N$ grows this distribution becomes more and more peaked around $\hat{\Theta}_{\text{ML}, N}$, which in turn moves to $\theta_o$, so that in the limit the density acts as a dirac delta-function so that

$$\log f_Y(Y; \eta) \overset{w.p.1}{\to} \log f_\Theta(\theta_o; \eta)$$

modulo $\eta$ independent terms. Thus, in case this convergence is uniform,

$$\hat{\eta}_{EB, N} \to \eta^*_{EB}(\theta_o) := \arg \max_{\eta \in D_\eta} \log f_\Theta(\theta_o; \eta)$$

(4.76)

Notice that in general $\eta^*_{EB}(\theta_o) \neq \eta^*(\theta_o)$ since they correspond to different criteria.

4.9.4 Accounting for Hyperparameter Variability in the Bayes Estimator

Above we have seen that the sequence $\{ \hat{\eta}_N \}$ of estimated hyperparameters will converge to some limit $\eta^*(\theta_o)$ for standard estimation procedures. From this one may be led to believe the excess MSE then becomes $\text{XMSE} \left[ \theta_o, f_N(\hat{\theta}_{\text{ML}, N}, 1/N, \eta^*(\theta_o)) \right]$. However, the variability of $\hat{\eta}_N$ will incur an increase in the MSE, and thus in the excess MSE.
4.9. ASYMPTOTICS OF BIASED ESTIMATORS

We will consider the case where \( f_\Theta \) in the Bayes estimator is parametrized: \( f_\Theta(\theta) = f_\Theta(\theta; \eta) \), \( \eta \in D_\eta \). The asymptotic approximation (4.63) is then given by

\[
\hat{\eta}_f(\cdot, \eta), N \approx \frac{\mathbb{E}[\Theta f_\Theta(\Theta; \eta)]}{\mathbb{E}[f_\Theta(\Theta; \eta)]}
\]

where \( \Theta \) has distribution \( N(\hat{\Theta}_{ML,N}, \frac{1}{N} I_N^{-1}(\hat{\Theta}_{ML,N})) \). Following the arguments before (4.66), for fix \( \eta \) the estimator asymptotically can be expressed as a function \( f_N(\hat{\Theta}_{ML,N}, \gamma; \eta) \), now depending on \( \eta \).

Assuming now instead that \( \gamma \) is a function of the ML-estimate, i.e. \( \gamma = \gamma_N(\hat{\Theta}_{ML,N}) \), where \( \gamma_N \) converges uniformly to \( \gamma^* \) on \( D_\theta \), it follows that \( f_N(\hat{\Theta}_{ML,N}, \gamma) = f_N(\hat{\Theta}_{ML,N}, \gamma(\hat{\Theta}_{ML,N})) \) and the limit (4.66) becomes

\[
f_N(\hat{\Theta}_{ML,N}, \gamma) \to \theta_o + \gamma I_F^{-1}(\theta_o) \frac{\partial}{\partial \theta} \log f_\Theta(\theta_o; \eta^*(\theta_o)) \quad \text{as} \quad N \to \infty \quad (4.77)
\]

When hyperparameters are estimated, the limit function defined in (4.53) thus changes from the limit (4.67)

\[
f_{f_\Theta}(\theta, \gamma) := \theta + \gamma I_F^{-1}(\theta) \frac{\partial}{\partial \theta} \log f_\Theta(\theta)
\]

of the Bayes estimator to

\[
f(\theta, \gamma) := \theta + \gamma I_F^{-1}(\theta) \frac{\partial}{\partial \theta} \log f_\Theta(\theta^*(\theta)) \quad (4.78)
\]

The excess MSE when the estimator is the Bayes estimator, but with \( \eta \) estimated from data according to \( \eta = \eta^*_N(\hat{\Theta}_{ML,N}) \), is thus given by (4.58)

\[
\text{XMSE} \left[ \theta, f_N(\hat{\Theta}_{ML,N}, \gamma) \right] = \frac{\partial f(\theta_o, 0)}{\partial \gamma} \left( \frac{\partial f(\theta_o, 0)}{\partial \gamma} \right)^T + \frac{\partial^2 f(\theta_o, 0)}{\partial \gamma \partial \theta^T} I_F^{-1}(\theta) + I_F^{-1}(\theta) \frac{\partial^2 f(\theta_o, 0)}{\partial \theta \partial \gamma} \quad (4.79)
\]

with \( f \) given by (4.78).

The first term in the expression (4.58) for the excess MSE corresponds to the bias contribution. Since

\[
\frac{\partial f}{\partial \gamma}(\theta, \gamma) = I_F^{-1}(\theta) \frac{\partial}{\partial \theta} \log f_\Theta(\theta^*(\theta)) = \frac{\partial f_{\Theta}(\cdot, \eta^*(\theta_o))}{\partial \gamma} \quad (4.79)
\]

the bias term contributing to the excess MSE will be the same as if \( \eta = \eta^*(\theta_o) \) had been used. However, as

\[
\frac{\partial^2 f}{\partial \gamma \partial \theta^T}(\theta, \gamma) = \frac{\partial}{\partial \theta^T} \left( I_F^{-1}(\theta) \frac{\partial}{\partial \theta} \log f_\Theta(\theta^*(\theta)) \right) = \frac{\partial^2 f_{\Theta}(\cdot, \eta^*(\cdot))}{\partial \gamma \partial \theta^T} \quad (4.80)
\]

the variance contribution to the excess MSE will differ in comparison with the Bayes estimator using \( f_{\Theta}(\theta, \eta^*(\theta_o)) \) as weighting function. The variance contribution will thus depend on the sensitivity of
the function \( \eta^*(\theta) \). We study the most interesting cases below. For simplicity we will assume that \( I_F \) is constant so that (4.69) reduces to
\[
\frac{\partial^2}{\partial \eta \partial \theta^T} f_{\Theta}(\theta, \gamma) = I_F^{-1} \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta)
\]
giving
\[
\frac{\partial^2 f}{\partial \gamma \partial \theta^T} (\theta, \gamma) = I_F^{-1} \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) + I_F^{-1} \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) \frac{d\eta^*(\theta)}{d\theta}
\]
\[= I_F^{-1} \left( \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) + \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) \frac{d\eta^*(\theta)}{d\theta} \right)
\]
(4.81)

**The Empirical Bayes Estimator**

For the Empirical Bayes estimator, \( \eta^* \) is given by \( \eta_{EB}^*(\theta_o) \) in (4.76)
\[\eta_{EB}^*(\theta_o) := \arg \max_{\eta \in D_\eta} \log f_{\Theta}(\theta_o; \eta) \]
which, hence, satisfies
\[\frac{\partial}{\partial \eta} \log f_{\Theta}(\theta; \eta^*(\theta)) = 0, \quad \forall \theta \in D_\theta \]
Taking the full derivative of this expression with respect to \( \theta \) gives
\[\frac{\partial^2}{\partial \eta \partial \theta^T} \log f_{\Theta}(\theta_o; \eta^*(\theta)) + \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) \frac{d\eta^*(\theta)}{d\theta} = 0 \]
giving
\[\frac{d\eta^*(\theta)}{d\theta} = - \left( \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) \right)^{-1} \frac{\partial^2}{\partial \eta \partial \theta^T} \log f_{\Theta}(\theta_o; \eta^*(\theta)) \]
which inserted in (4.81) gives
\[\frac{\partial^2 f}{\partial \gamma \partial \theta^T} (\theta, \gamma) = I_F^{-1} \left( \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta) + \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) \frac{d\eta^*(\theta)}{d\theta} \right)
\]
\[= I_F^{-1} \left( \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta) - \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) \left( \frac{\partial^2}{\partial \eta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) \right)^{-1} \frac{\partial^2}{\partial \eta \partial \theta^T} \log f_{\Theta}(\theta_o; \eta^*(\theta)) \right)
\]
(4.82)

Recall now that \( \eta^*(\theta) \) is a maximum of \( \log f_{\Theta}(\theta; \eta) \) and hence
\[\frac{\partial^2}{\partial \eta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) < 0 \]
from which, referring to (4.79), we conclude that Empirical Bayes adds
\[-2I_F^{-1} \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) \left( \frac{\partial^2}{\partial \eta \partial \eta^T} \log f_{\Theta}(\theta; \eta^*(\theta)) \right)^{-1} \frac{\partial^2}{\partial \eta \partial \theta^T} \log f_{\Theta}(\theta_o; \eta^*(\theta)) I_F^{-1} > 0 \]
to the excess MSE of the Bayes estimator that uses \( f_{\Theta}(\theta; \eta_{EB}^*(\theta_o)) \) as weighting when the information matrix is constant.
The ML-estimator

As already observed, the bias term in the excess MSE is identical to the bias term for the Bayes estimator with \( \eta \) fixed to \( \eta^*(\theta_o) \). What we are after is a more transparent expression for the second term in (4.81)

\[
\frac{\partial^2 f}{\partial \gamma \partial \theta^T}(\theta, \gamma) = I_F^{-1} \left( \frac{\partial^2 f}{\partial \theta^T} \log f_\theta(\theta; \eta^*(\theta)) + \frac{\partial^2 f}{\partial \theta \partial \eta^T} \log f_\theta(\theta; \eta^*(\theta)) \frac{d\eta^*(\theta)}{d\theta} \right)
\]

In the case of the ML-estimator of \( \eta_{N, \eta} \), \( \eta^*(\theta) \) is given as the minimizer of the excess MSE, see (4.73). Since we are restricting attention to the Bayes estimator, \( f_N(\hat{\theta}_{ML,N}, 1/N, \eta) \) is the Bayes estimator \( \hat{\theta}_{f_\theta(\cdot, \cdot), N} \). Now, similarly to the EB-case,

\[
\frac{d\eta^*(\theta)}{d\theta} = - \left( \frac{\partial^2}{\partial \eta \partial \eta^T} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\theta(\cdot, \cdot), N} \right] \right)^{-1} \frac{\partial^2}{\partial \eta \partial \theta^T} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\theta(\cdot, \cdot), N} \right]
\]

where the excess MSE is given by (4.85)

\[
\text{XMSE} \left[ \theta, \hat{\theta}_{f_\theta(\cdot, \cdot), N} \right] = I_F^{-1}(\theta_o) \left( \frac{\partial}{\partial \eta} \log f_\theta(\theta_o; \eta) \frac{\partial}{\partial \theta^T} \log f_\theta(\theta_o; \eta) + \frac{\partial^2}{\partial \theta \partial \eta^T} \log f_\theta(\theta_o; \eta) \right) I_F^{-1}(\theta_o)
\]

To proceed without too much technicalities, we will consider our running example.

**Example 4.9.5** (Example 4.5.1 Continued). Recall the model

\[ Y = \Phi \theta + E, \quad Y \in \mathbb{R}^N, \quad \theta \in \mathbb{R}^n, \quad n \leq N, \quad E \sim N(0, \lambda_e I) \]

The Bayes estimator based on the weighting function \( f_\theta \sim N(0, P(\eta)) \) is given by

\[
\hat{\theta}_{N, \eta} = (I + \frac{1}{N} I_{F, N}^{-1}(\eta))^{-1} \hat{\theta}_{ML,N}
\]

where \( I_{N, F} \) is the per sample information matrix \( I_{N, F} := \frac{1}{N} \Phi^T \Phi \) which we assume converges to some \( I_F > 0 \). For simplicity we assume that \( \eta \) is scalar.

We thus have

\[
\log f_\theta(\theta, \eta) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \log P(\eta) - \frac{1}{2} \theta^T P^{-1}(\eta) \theta
\]

\[
\frac{\partial}{\partial \theta} \log f_\theta(\theta, \eta) = -P^{-1}(\eta) \theta
\]

\[
\frac{\partial^2}{\partial \theta \partial \theta^T} \log f_\theta(\theta, \eta) = -P^{-1}(\eta)
\]

\[
\frac{\partial^2}{\partial \theta \partial \eta^T} \log f_\theta(\theta, \eta) = - \frac{\partial}{\partial \eta} P^{-1}(\eta) \theta
\]

giving

\[
\text{XMSE} \left[ \theta, \hat{\theta}_{f_\theta(\cdot, \cdot), N} \right] = I_F^{-1} \left( P^{-1}(\eta) \theta \theta^T P^{-1}(\eta) - 2P^{-1}(\eta) \right) I_F^{-1}
\]

so that its scalar version is

\[
\text{XMSE} \left[ \theta, \hat{\theta}_{f_\theta(\cdot, \cdot), N} \right] = \theta^T P^{-1}(\eta) I_F^{-2} P^{-1}(\eta) \theta - 2 \text{Tr} \left[ I_F^{-1} P^{-1}(\eta) I_F^{-1} \right]
\]
with derivatives
\[
\frac{\partial}{\partial \theta} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] = 2P^{-1}(\eta)I_F^{-2}P^{-1}(\eta)\theta
\]
\[
\frac{\partial^2}{\partial \theta \partial \omega^T} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] = 2P^{-1}(\eta)I_F^{-2}P^{-1}(\eta)
\]
\[
\frac{\partial}{\partial \eta} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] = -2\theta^T P^{-1}(\eta)I_F^{-2} \frac{\partial P^{-1}(\eta)}{\partial \eta} \theta + \theta \text{ independent terms}
\]
\[
\frac{\partial^2}{\partial \eta \partial \theta} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] = 4P^{-1}(\eta)I_F^{-2} \frac{\partial}{\partial \eta} P^{-1}(\eta)\theta
\]
\[
= 4 \frac{\partial^2}{\partial \theta \partial \omega^T} \log f_\theta(\theta, \eta)I_F^{-2} \frac{\partial}{\partial \theta \partial \omega} \log f_\theta(\theta, \eta)
\]
The last expression can be written
\[
\frac{\partial^2}{\partial \eta \partial \theta^T} \log f_\theta(\theta, \eta) = \frac{1}{4} I_F^{-1} \left( \frac{\partial^2}{\partial \theta \partial \omega^T} \log f_\theta(\theta, \eta) \right)^{-1} \frac{\partial^2}{\partial \theta \partial \omega^T} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right]
\]
Inserting this and (4.84) in (4.83) gives
\[
\frac{\partial^2 f}{\partial \gamma \partial \theta^T}(\theta, \gamma) = I_F^{-1} \left( \frac{\partial^2}{\partial \theta \partial \omega^T} \log f_\theta(\theta, \eta^*()) + \frac{1}{4} I_F^2 \left( \frac{\partial^2}{\partial \theta \partial \omega^T} \log f_\theta(\theta, \eta^*()) \right) \right)^{-1}
\]
\[
\frac{\partial^2}{\partial \theta \partial \eta} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] \left( \frac{\partial^2}{\partial \eta \partial \theta} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] \right)^{-1} \frac{\partial^2}{\partial \eta \partial \theta^T} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right]
\]
\[
= \left( I + \frac{1}{4} I_F \left( \frac{\partial^2}{\partial \theta \partial \omega^T} \log f_\theta(\theta, \eta^*()) \right)^{-1} Z \left( \frac{\partial^2}{\partial \theta \partial \omega^T} \log f_\theta(\theta, \eta^*()) \right)^{-1} I_F \right) I_F^{-1} \frac{\partial^2}{\partial \theta \partial \omega^T} \log f_\theta(\theta, \eta^*())
\]
where
\[
Z := \frac{\partial^2}{\partial \theta \partial \omega^T} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] \left( \frac{\partial^2}{\partial \eta \partial \theta} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] \right)^{-1} \frac{\partial^2}{\partial \eta \partial \omega^T} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right]
\]
Since \( \eta^*() \) minimizes \( \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] \)
\[
\frac{\partial^2}{\partial \eta \partial \theta} \text{XMSE} \left[ \theta, \hat{\theta}_{f_\omega(\cdot, N)} \right] > 0
\]
and hence \( Z > 0 \). Now since \( I_F^{-1} \frac{\partial^2}{\partial \omega \partial \theta^T} \log f_\theta(\theta, \eta^*()) \) is \( \frac{\partial^2 f}{\partial \eta \partial \omega^T}(\theta, \gamma) \) for the Bayes estimator using the fix \( \eta = \eta^*() \), ML-estimation of \( \eta^*() \) increases the variance contribution to the excess MSE by the factor within the big parenthesis in (4.86) which is greater than the identity matrix.

### 4.9.5 Exercises

4.9.1. Let \( A \in \mathbb{R}^{m \times m} \) and \( B \in \mathbb{R}^{m \times n} \). Show that
\[
B(AB + I_{n \times n})^{-1} = (BA + I_{m \times m})^{-1} B
\]
when the inverses exist.
4.9.2. Suppose that

\[ Y = \Phi \theta + E \]

where \( E \sim N(0, \lambda_e I) \).

(a) Compute the Bayes estimator when \( f_\Theta \) corresponds to a \( N(0, \bar{P}) \) distribution where \( \bar{P} > 0 \).

(b) Compute the bias and variance of the Bayes estimator.

Assuming that \( \Phi \) has full column rank, show that the covariance is strictly smaller than the covariance of an efficient estimator (which one is that?).

(c) Compute the posterior covariance of \( \Theta \) given \( Y \) when \( f_\Theta \) in a) is the prior distribution of \( \Theta \) and

\[ Y = \Phi \Theta + E \]

(d) * Show that when

\[ Y = \Phi \theta_o + E \]

and \( \bar{P} = G_o := \theta_o \theta_o^T \) the posterior covariance is identical to the matrix valued MSE

\[
\text{Cov} \left[ \hat{\theta}_{f_o} \right] + \left( \mathbb{E} \left[ \hat{\theta}_{f_o} \right] - \theta_o \right) \left( \mathbb{E} \left[ \hat{\theta}_{f_o} \right] - \theta_o \right)^T
\]

You may assume (incorrectly) that \( G_o \) is full rank.
Can you interpret this result?

4.9.3. Let \( X \) and \( Y \) be two random variables. We know that \( \mathbb{E} [X|Y] \) has the same mean as \( X \). However, show that \( \mathbb{E} [X|Y] \) is a biased estimator of \( X \) itself. More precisely, show that if

\[ \mathbb{E} [\mathbb{E} [X|Y] | X = x] = x \quad \forall x \]

then

\[ \mathbb{E} [(X - \mathbb{E} [X|Y])^2] = 0 \]

4.9.4. Let the risk be the MSE and suppose that \( \hat{\theta} \) is admissible for \( \theta \). Show that \( a\hat{\theta} + b \) is admissible for \( a\theta + b \).

4.9.5. Let the risk be the MSE and suppose that \( \hat{\theta} \) is minimax for \( \theta \). Show that \( a\hat{\theta} + b \) is minimax for \( a\theta + b \).

4.10 References

Sections 4.1–4.6, Chapter 4-5 in [10].
See also [4].
Section 4.5.10: [16].
### 4.A Stein’s Identity

**Lemma 4.A.1.** Let $Y$ be distributed according to the canonical exponential family (2.33)

$$ p_Y(y; \theta) = e^{\theta^T y - A(\theta) h(y)} \tag{4.87} $$

and let $g$ be any differentiable function such that $\mathbb{E} \|g'(Y)\| < \infty$. Then

$$ \mathbb{E} [g'(Y)] = - \mathbb{E} \left[ \left( \frac{d}{d} \log h(Y) + T'(Y)^T \theta \right) g(Y) \right] \tag{4.88} $$

### 4.B Proof of Lemma 4.8.1

To see (4.48) notice that, with $\lambda_k$ denoting the eigenvalues of $M$,

$$ \text{Tr}\{M\} + \log \det M^{-1} = \sum_k \lambda_k + \log \left( \prod_k \frac{1}{\lambda_k} \right) $$

$$ = \sum_k \lambda_k - \sum \log \lambda_k $$

Differentiating this expression with respect to $\lambda_l$ gives

$$ 1 - 1/\lambda_l \tag{4.89} $$

so that clearly $\lambda_1 = \ldots = \lambda_n = 1$ at a stationary point. But this corresponds to a unique $M$, namely $M = I$. Differentiating (4.89) shows that the Hessian at the stationary point is $I$, i.e. positive definite. Thus we have shown that the global minimum of $\text{Tr}\{M\} + \log \det M^{-1}$ is the identity matrix.
Chapter 5

Dynamical Models

5.1 Noise Models

5.1.1 Introduction

When models from time-series measurements a key issue is how to handle temporally correlated measurements. We illustrate this with an example.

**Example 5.1.1.** Suppose that our data generating mechanism is given by

\[ y(t) = \theta_0 u(t) + e(t), \quad \theta_0 = 0 \]

One possibility to determine \( \theta_0 \) from measurements of \( u(t) \) and \( y(t) \) is to minimize the least-squares criterion

\[ V_{id}(\theta) = \frac{1}{N} \sum_{t=1}^{N} (y(t) - \theta u(t))^2 \tag{5.1} \]

With the data given in Figure 5.1 we get the estimate in Figure 5.2 and we obviously have a very good estimate.

Suppose now instead that the data is generated according to

\[ y(t) = \theta_0 u(t) + H(q)e(t), \quad \theta_0 = 0.5, \quad H(q) = \frac{1}{1 - 0.9q^{-1}} \tag{5.2} \]

The resulting estimate is shown in Figure 5.3 which is significantly worse than in the previous case.

An explanation to this is obtained if we take a look at Figure 5.4 which shows the data we have used in this case.

Clearly this is a more difficult case and the reason is that the correlation in the noise terms misleads the least-squares algorithm. So can we fix this? Well, an idea could be to try to make the noise white. We could pre-filter both the input and the output through the filter \( 1/H(q) = 1 - 0.9q^{-1} \). This would give

\[ y_f(t) = \theta_0 u_f(t) + e(t), \quad \theta_0 = 0.5 \]

where \( y_f(t) = H^{-1}(q)y(t) \) and \( u_f(t) = H^{-1}(q)u(t) \). We are then back to the white measurement noise case so let us try on the data in Figure 5.4. The pre-filtered data is given in Figure 5.5. The
CHAPTER 5. DYNAMICAL MODELS

Figure 5.1: The input, noise free and noise contaminated output as function of time in Example 5.1.1.

Figure 5.2: The estimated output as function of time as well as the true noise free output in Example 5.1.1.
5.1. NOISE MODELS

Figure 5.3: The estimated output as function of time as well as the true noise free output in Example 5.1.1 when data are generated by (5.2).

Figure 5.4: The input, noise free and noise contaminated output as function of time in Example 5.1.1.
estimate minimizing

\[ V_{id}(\theta) = \frac{1}{N} \sum_{t=1}^{N} (y_f(t) - \theta u_f(t))^2 \]

is given in Figure 5.6.

Clearly a significant improvement. However, it is rare to know the noise model \( H \) in advance. However, we could, in addition to \( \theta_1 \), estimate the parameter in \( H \) as well.

We can write the cost function

\[ V_{id}(\theta) = \frac{1}{N} \sum_{t=1}^{N} (H^{-1}(q)y(t) - \theta_1 H^{-1}(q)u(t))^2 = \frac{1}{N} \sum_{t=1}^{N} H^{-1}(q)(y(t) - \theta_1 u(t))^2 \]

and then we extend the parameter vector \( \theta \) with \( \theta_2 \) in \( H(q, \theta_2) = 1/(1 + \theta_2 q^{-1}) \) and use

\[ V_{id}(\theta) = \frac{1}{N} \sum_{t=1}^{N} H^{-1}(q, \theta_2)(y(t) - \theta_1 u(t))^2 \]

This is now a non-linear least-squares problem. The resulting estimate is shown in Figure 5.7. Clearly a significant improvement over the case when the least-squares cost function (5.1) was used, see Figure 5.3.

### 5.1.2 Stochastic Models

Example 5.1.1 shows the importance of modeling temporal noise and disturbance dependencies. So what type of models can be used for this? One possibility is to, as in the example, see a noise sequence
Figure 5.6: The estimated output as function of time as well as the true noise free output in Example 5.1.1 when data are generated by (5.2) and pre-filtering is used.

Figure 5.7: The estimated output as function of time as well as the true noise free output in Example 5.1.1 when data are generated by (5.2) and the noise model is estimated.
\{v(t)\}$ as a sequence of random variables, i.e. a stochastic process. A stochastic process $\{v(t)\}$ is characterized by its finite dimensional distributions $p_{t_1,...,t_n}(v_1,\ldots,v_n) := p_{v(t_1),\ldots,v(t_n)}(v_1,\ldots,v_n)$ for all $n$ and $t_1,\ldots,t_n$. Often the characteristics of the noise does not vary with time. Stationary models can then be used. A (strictly) stationary stochastic process $\{v(t)\}$ has a finite dimensional distributions $p_{t_1,...,t_n}$ that are invariant to time shifts:

$$p_{t_1+\tau,...,t_n+\tau} = p_{t_1,...,t_n}, \quad \forall n \in \mathbb{Z}_+ \text{ and } \forall t_1,\ldots,t_n, \tau \in \mathbb{N}$$

A weak (or wide-sense) stationary process $\{v(t)\}$ has constant mean and autocorrelations which are invariant to time-shifts:

$$\mathbb{E}[v(t)] = m_v, \quad \mathbb{E}[v(t)v^T(t-\tau)] = R_v(\tau), \quad \forall t, \tau \in \mathbb{N}$$

An alternative is to use quasi-stationarity which allows for deterministic components of the process, we refer to Chapter 2 in [11]. Many of the results that apply to weakly stationary processes apply to quasi-stationary processes as well.

### 5.1.3 Models of Stationary Processes

So how do we model a stationary stochastic process $\{v(t)\}_{t=-\infty}^{\infty}$? The starting point is what is known as the Wold decomposition which amounts to a linear projection of $v(t)$ on the span of its past $v(t-1), v(t-2), \ldots$. Denoting this projection $\hat{v}(t|t-1)$ and the prediction error $v(t) - \hat{v}(t|t-1)$ by $e(t)$, we can write

$$v(t) = e(t) + \hat{v}(t|t-1) = e(t) + \sum_{k=1}^{\infty} \alpha_k v(t-k)$$

for some $\{\alpha_k\}$. However, we can now replace $v(t-1)$ by the same type of expansion giving

$$v(t) = e(t) - \alpha_1 e(t-1) + \sum_{k=2}^{\infty} \beta_k v(t-k)$$

Continuing like this gives

$$v(t) = \sum_{k=0}^{\infty} h_k e(t-k)$$

for some sequence $\{h_k\}$. Since $e(t)$ is orthogonal to the past and $e(s)$, $s < t$ is a function of the past, $\{e(t)\}$ form an uncorrelated sequence. This suggests that filtered white (uncorrelated) noise could be a useful model of correlated noise.

In order to be precise in how such models should be constructed we will need a bit of formal theory for the decomposition above. For this purpose, we introduce

$$S_t = \text{Span}\{v(s) : -\infty < s \leq t\}, \text{ and } S_{-\infty} = \bigcap_{t=-\infty}^{\infty} S_t$$

where $S_{-\infty}$ can be viewed as the remote past.

**Theorem 5.1.1** (Wold Decomposition, Theorem 6.11 in [17]). A weakly stationary stochastic process $\{v(t)\}_{t=-\infty}^{\infty}$ can be decomposed as

$$v(t) = v_r(t) + v_d(t), \quad v_r(t) = \sum_{k=0}^{\infty} h_k e(t-k) \quad (5.3)$$

where
5.1. NOISE MODELS

i) $\mathbb{E}[e(t)] = 0$ and $\mathbb{E}[e(t)e^T(t-\tau)] = \delta(\tau)\Sigma$ where $\Sigma \geq 0$ is the prediction error matrix

$$\Sigma := \mathbb{E}\left[(v(t) - v(t)_{|S_{t-1}})(v(t) - v(t)_{|S_{t-1}})^T\right]$$

ii) $\mathbb{E}\left[v(0)e^T(t-k)\right] = h_k\Sigma$, where $h_0\Sigma = \Sigma = \Sigma h_k^T$, and

$$\sum_{k=0}^{\infty} h_k\Sigma h_k^T < \infty \quad (5.4)$$

iii) $e(t) \in S_t$

iv) $\mathbb{E}[e(t)v^T_d(t)] = 0$ for all $s, t \in \mathbb{N}$

v) $v_d(t) \in S_{-\infty}$ for all $t \in \mathbb{N}$.

vi) $v_d$ is deterministic, meaning that

$$\mathbb{E}\left[(v_d(t) - v_d(t)_{|S_{t-1}})(v_d(t) - v_d(t)_{|S_{t-1}})^T\right] = 0$$

The rank of a stationary process $\{v(t)\}$ is the rank of its prediction error matrix.

**Corollary 5.1.1.** If the process is full rank the decomposition in Theorem 5.1.1 is unique and $h_0 = I$.

If $S_{-\infty} = \{0\}$ we say that the process is regular and then $v_d \equiv 0$.

For a stationary process $\{v(t)\}$ we define its spectral density as the Fourier transform of its autocorrelation function

$$\Phi_v(e^{i\omega}) = \sum_{\tau=-\infty}^{\infty} R_v(\tau)e^{-i\omega\tau}$$

if this quantity exists.

**Theorem 5.1.2** (7.9 Main Lemma I and 7.10 Main Theorem I in [17], Theorem 1 in [1]). *For a regular process the spectrum is given by

$$\Phi_v(e^{i\omega}) = H(e^{i\omega})\Sigma H^*(e^{i\omega})$$

where $H(z)$ is the $z$-transform of $\{h_k\}$ in Theorem 5.1.1

$$H(z) = \sum_{k=0}^{\infty} h_k z^{-k}$$

and $H^*(e^{i\omega})$ denotes the complex conjugate of $H(e^{i\omega})$.

A regular process is full rank if and only if

$$\int_{-\pi}^{\pi} \log \det \Phi_v(e^{i\omega})d\omega < \infty \quad (5.5)$$

in which case the determinant of the prediction error matrix is given by

$$\det \Sigma = e^{\frac{1}{4\pi}\int_{-\pi}^{\pi} \log \det \Phi_v(e^{i\omega})d\omega}$$

For a regular full rank process, $H^{-1}(z)$ is analytic outside of the unit disc, i.e. in $D = \{z : |z| > 1\}$. 
CHAPTER 5. DYNAMICAL MODELS

Condition (5.5) implies that a regular full rank process cannot have a spectrum with zeros on the unit circle while the last result in the theorem implies that it cannot have zeros outside of the unit disc either. Furthermore, as \( \Sigma > 0 \) for a full rank process (5.4) requires that \( H(z) \) has no poles on or outside the unit circle. This means that when first and second order moments are concerned, any stationary stochastic process \( \{v(t)\} \) can be modelled as

\[
v(t) = H(q)e(t) + m \tag{5.6}
\]

where \( \{e(t)\} \) is a sequence of zero mean uncorrelated random variables with \( \mathbb{E}[e(t)e^T(t)] = \Sigma \) for some \( \Sigma > 0 \), and where \( H(z) = \sum_{k=0}^{\infty} h_k z^{-k} \) with \( h_0 = I \), has no zeros or poles on or outside the unit disc. The constant \( m \) is the mean of the process.

It is common to restrict the model class to some parametrized family, such as

\[
H(q) = \frac{C(q)}{D(q)} = \frac{1 + c_1 q^{-1} + \ldots + c_{n_c} q^{-n_c}}{1 + d_1 q^{-1} + \ldots + d_{n_d} q^{-n_d}} \tag{5.7}
\]

where \( z^{n_c} C(z) \) and \( z^{n_d} D(z) \) have no zeros on or outside of the unit disc and where \( c_i, i = 1, \ldots, n_c \) and \( d_i, i = 1, \ldots, n_d \) are parameters (possibly constrained to some set). We stress, however, that when disturbances and noise undergo non-linear transformations in the model, the above model class is not sufficient as only second order characteristics are modeled. Then the finite dimensional distributions of the process need further modeling.

An important exception is when \( \{v(t)\} \) is a Gaussian process. Then \( \{e(t)\} \) will be an iid Gaussian process since uncorrelated Gaussian random variables are independent. This is an important result so we state it separately.

**Theorem 5.1.3.** Any regular full rank stationary Gaussian stochastic process \( \{v(t)\} \) can be decomposed as

\[
v(t) = H(q)e(t) + m \tag{5.8}
\]

where \( \{e(t)\} \) is a sequence of zero mean iid Gaussian random variables with \( \mathbb{E}[e(t)e^T(t)] = \Sigma \) for some \( \Sigma > 0 \), and where \( H(z) = \sum_{k=0}^{\infty} h_k z^{-k} \) with \( h_0 = I \), has no zeros or poles on or outside of the unit circle. The constant \( m \) is the mean of the process.

When higher order moments of \( v(t) \) come into play in the model, a simple modification of (5.6) and (5.7) is to add the assumption that \( \{e(t)\} \) is a sequence of iid random variables with some (possibly parametrized) pdf \( p_e \).

5.2 Dynamical Models

We refer to Chapters 4 and 5 in [11].

We point to two different model types

i) Models directly invertible with respect to the noise

\[
y(t) = \hat{y}(t|t - 1) + e(t) \tag{5.9}
\]

where \( \hat{y}(t|t - 1) \) is a function of past observations \( Z_{t-1} = \{y(t-1), u(t-1), \ldots, y(1), u(1)\} \) and where \( \{e(t)\} \) is assumed to be an iid sequence. Here the innovations \( e(t) \) can be computed in a simple way.
ii) Models not directly invertible with respect to the noise. For example
\[ x(t + 1) = f_t(x(t), u(t), w(t), \theta) \]
\[ y(t) = h_t(x(t), u(t), v(t), \theta) \]
where both \( \{w(t)\} \) and \( \{v(t)\} \) are modeled as stochastic processes.

5.3 Estimation Methods

5.3.1 Maximum Likelihood Estimation

In ML estimation we need the likelihood function \( p_Z(z; \theta) \) of our input-output observations \( z = \{y(1), u(1), \ldots, y(N), u(N)\} \). There are four main approaches to handle this

- Direct calculation
- Model inversion
- Marginalization
- Filtering

Direct calculation

In simple cases it is possible to directly express the likelihood function.

**Example 5.3.1.** Suppose that \( Y \in \mathbb{R}^N \) consist of a function \( \mu \) observed in correlated noise according to
\[
Y = \mu(\theta) + H(\theta)E, \quad E \in \mathcal{N}(0, \lambda_e I) \tag{5.10}
\]
Here \( E \in \mathbb{R}^M \) where \( M \) is arbitrary. Then we know from the properties of normal distributed random variables that
\[
Y \sim \mathcal{N}(\mu(\theta), \lambda_e H(\theta)H^T(\theta))
\]
from which the likelihood function follows.

In particular, stationary Gaussian processes can be directly handled in this way.

**Example 5.3.2.** Suppose that \( \{y(t)\} \) is a Gaussian stationary process satisfying
\[
y(t) + \theta y(t - 1) = e(t) \tag{5.11}
\]
where \( \{e(t)\} \) is the zero mean innovations sequence with variance \( \lambda_e \) and where \(-1 < \theta < 1\). A process of the type (5.11) is called an AR-1 process (AutoRegressive process of order 1).

To compute the covariance of \( Y = [y(1) \ldots y(N)]^T \), we observe that \( Y \) can be seen as the output at time \( t = N \) of the model
\[
z(t) = \begin{bmatrix} q^{-N+1} \\ \vdots \\ q^0 \end{bmatrix} y(t)
\]
Now
\[ y(t) = \frac{1}{1 + \theta q^{-1}} e(t) \]
and therefore \( y(t) \) has spectrum
\[ \Phi_y(e^{i\omega}) = \frac{1}{1 + \theta e^{-i\omega}} \lambda_e \]
Thus \( z(t) \) (and therefore \( Y \)) has covariance
\[ \Sigma = \frac{1}{2\pi} \int_{-\pi}^{\pi} \begin{bmatrix} e^{i(-N+1)\omega} & \cdots & e^{i\omega} \\ \vdots & \ddots & \vdots \\ e^{iN\omega} & \cdots & e^{i\omega} \end{bmatrix} \Phi_y(e^{i\omega}) d\omega \]
\[ = \frac{1}{2\pi} \int_{-\pi}^{\pi} \begin{bmatrix} e^{i(-N+1)\omega} & \cdots & e^{i\omega} \\ \vdots & \ddots & \vdots \\ e^{iN\omega} & \cdots & e^{i\omega} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ e^{iN\omega} \end{bmatrix} * \left| \frac{1}{1 + \theta e^{-i\omega}} \right|^2 \lambda_e d\omega \]
from which it can be shown that
\[ \Sigma = \frac{\lambda_e}{1 - \theta^2} \begin{bmatrix} 1 & -\theta & -\theta^2 & \cdots & (-\theta)^{N-1} \\ -\theta & 1 & -\theta & \cdots & (-\theta)^{N-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ (-\theta)^{N-1} & (-\theta)^{N-2} & \cdots & 1 \end{bmatrix} \]
Since the mean of \( y(t) \) is zero, we then have the distribution for \( Y \).

An alternative way to compute \( \Sigma \) for an AR-process is to multiply (5.11) with \( y(t) \) and take expectation. Since \( y(t) = e(t) + \hat{y}(t|t-1) \), \( E[e(t)y(t)] = \lambda_e \) and hence this operation gives
\[ R_y(0) + \theta R_y(1) = \lambda_e \]
Instead multiplying with \( y(t-1) \) gives
\[ R_y(1) + \theta R_y(0) = 0 \]
Collecting these equations gives
\[ \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix} \begin{bmatrix} R_y(0) \\ R_y(1) \end{bmatrix} = \lambda_e \]
which can be solved giving \( R_y(0) = \lambda_e / (1 - \theta^2) \) and \( R_y(1) = -\theta R_y(0) \). Multiplying (5.11) with \( y(t-\tau), \tau > 1 \) gives
\[ R_y(\tau) + \theta R_y(\tau - 1) = 0 \]
from which we get \( R_y(\tau) = (-\theta)^{|\tau|} R_y(0) \). Covariance formulas for some ARMA processes are given in Appendix 2C in [11].

In the examples the computation of the likelihood become more and more costly as \( N \) increases. In Example 5.3.2, to compute the likelihood we need to invert \( \Sigma \). Inverting an arbitrary matrix has a computational complexity of \( O(N^3) \) but using that \( \Sigma \) is Toeplitz, the complexity can be reduced to \( O(N^2) \).
5.3. ESTIMATION METHODS

Model Inversion

Suppose that $Y \in \mathbb{R}^N$ is given by $Y = f(E)$, $E \in \mathbb{R}^N$ where $f$ is one-to-one and where the pdf $p_E$ of $E$ is known. Then

$$p_Y(y) = p_E(f^{-1}(y)) \det \frac{df^{-1}(y)}{dy} \quad (5.12)$$

**Example 5.3.3** (Example 5.3.1 continued). From (5.10) it follows that

$$E = H^{-1}(\theta)(Y - \mu(\theta))$$

Thus using that $\det H^{-1}(\theta) = 1/\det H(\theta)$,

$$p_Y(y) = p_E(H^{-1}(\theta)(Y - \mu(\theta))) \frac{1}{\det H(\theta)}$$

which allows for any distribution of $E$.

The model (5.9) is invertible. We have

$$E = \begin{bmatrix} y(1) - \hat{y}(1|0) \\ \vdots \\ y(N) - \hat{y}(N|N-1) \end{bmatrix} \quad (5.13)$$

Notice that the Jacobian in (5.12) is lower triangular with ones in the diagonal (due to the linear term $y(t)$ in the expression for $e(t)$, for $t = 1, \ldots, N$. Thus when $\{e(t)\}$ are iid with pdf $p_e$, the likelihood is given by

$$p_Y(y) = \prod_{t=1}^{N} p_e(y(t) - \hat{y}(t|t-1))$$

**Example 5.3.4** (AR-1 model). Consider

$$y(t) + \theta y(t-1) = e(t), \quad t = 1, \ldots, N$$

where the pdf of $E = [e(1) \ldots e(N)]^T$ is assumed known. We have

$$y(1) = -\theta y(0) + e(1)$$
$$y(2) = -\theta y(1) + e(2) = -\theta(-\theta y(0) + e(1)) + e(2) = \theta^2 y(0) - \theta e(1) + e(2)$$
$$\vdots$$
$$y(N) = \ldots$$

giving

$$Y = T(\theta)E + L(\theta)y(0),$$
where $T(\theta)$ is a lower triangular Toeplitz matrix with first column $[1 \ -\theta \ \theta^2 \ \ldots \ (-\theta)^{N-1}]^T$. Now if $y(0)$ is assumed known or added as a parameter,

$$E = T^{-1}(Y - L(\theta)y(0))$$

so the above represents an invertible model. Since $\det T(\theta) = 1$,

$$p_Y(y) = f_E(T^{-1}(\theta)(y - L(\theta)y(0)))$$

**Example 5.3.5** (Example 5.3.4 continued). In Example 5.3.4 we can also use that

$$y(t) = \hat{y}(t|t-1) + e(t), \quad \hat{y}(t|t-1) = -\theta y(t-1)$$

and use (5.13), i.e.

$$E = \begin{bmatrix}
  y(1) + \theta y(0) \\
  y(2) + \theta y(1) \\
  \vdots \\
  y(N) + \theta y(N-1)
\end{bmatrix}$$

which is much simpler (and indeed is a proof that $T^{-1}(\theta)$ is a banded lower Toeplitz matrix with first column $[1 \ \theta \ 0 \ \ldots \ 0]^T$).

**Marginalization**

Suppose that $Y = f_W(E)$ where now also $W \in \mathbb{R}^M$ is a random variable with pdf $p_W$. Suppose that $f_w(\cdot)$ is invertible for every $w$ in the support of $W$. Then we can express the likelihood using the law of total probability

$$p_Y(y) = \int p_{Y,W}(y,w)dw = \int p_{Y|W}(y|w)p_W(w)dw = \int p_E(f_W^{-1}(Y)) \det \frac{df_W^{-1}(y)}{dy} p_W(w)dw$$

(5.15)

We thus need to marginalize out the unknown $W$ by computing an $M$-dimensional integral.

In Example 5.3.4 we ended up having to assume that $y(0)$ was known or should be estimated. This is a typical situation when using dynamical models. Assuming known initial conditions leads to what is known as the conditional ML method. An alternative is to make some stochastic assumption about the unknowns ($y(0)$ in the example). However, then the model is no longer invertible.

**Example 5.3.6** (Example 5.3.4 continued). Suppose that $y(0)$ is independent of $E$ with pdf $p_Y(0)(y(0))$. Assuming that the elements of $\{e(t)\}$ are independent, we see from (5.14) that we cannot compute $e(1) = y(1) + \theta y(0)$. However, conditioning on $y(0)$ gives as in Example 5.3.4 that

$$p_{Y|Y(0)}(y|y(0)) = p_E\begin{bmatrix}
  y(1) + \theta y(0) \\
  y(2) + \theta y(1) \\
  \vdots \\
  y(N) + \theta y(N-1)
\end{bmatrix}$$
5.3. ESTIMATION METHODS

and hence

\[
p_Y(y) = \int p_E \left( \begin{bmatrix} y(1) + \theta y(0) \\ y(2) + \theta y(1) \\ \vdots \\ y(N) + \theta y(N-1) \end{bmatrix} \right) p_Y(0) dy(0) = \int p_E \left( \begin{bmatrix} y(1) + \theta y(0) \\ y(2) + \theta y(1) \\ \vdots \\ y(N) + \theta y(N-1) \end{bmatrix} \right) p_Y(y(0)) dy(0)
\]

Filtering

Except for in Exercise 5.3.5 we have not made use of any structure of a model. The methods have been quite general. The final approach allows to benefit from the recursive structure causal dynamical models possess. Given \( y \in \mathbb{R}^N \), we can write the likelihood as

\[
p_Y(y) = p_Y(N|Y_{N-1}) p_Y(N-1|Y_{N-2}) \cdots = \prod_{t=2}^N p_Y(t|Y_{t-1}) p(y(1))
\]  \hfill (5.16)

where \( y_t = \{y(1), \ldots, y(t)\} \), or the negative log-likelihood as

\[
-\log p_Y(y) = -\sum_{t=2}^N \log p_Y(t|Y_{t-1}) p(y(t)|y_{t-1}) - \log p(y(1))
\]  \hfill (5.17)

Example 5.3.7 (Exercise 5.3.5 continued). Assuming that \( \{e(t)\} \) is an iid sequence with pdf \( p_e \), the relation

\[ y(t) = -\theta y(t-1) + e(t) \]

implies that

\[
p_Y(t|Y_{t-1}) = p_e(y(t) + \theta y(t-1))
\]

so that

\[
p_Y(y) = \prod_{t=2}^N p_e(y(t) + \theta y(t-1)) p_Y(1)(y(1))
\]

Filtering for Nonlinear State-Space Models

In this section we will derive an expression for the predictive density \( p(y(t+1)|Y_t) \) for state-space models of the type

\[
x(t+1) = f(x(t), u(t)) + w(t), \quad x(0) \sim p_{X(0)}(x(0)) \\
y(t) = h(x(t)) + v(t)
\]  \hfill (5.18, 5.19)

where \( \{w(t)\} \) and \( \{v(t)\} \) are sequences of independent random variables with pdfs \( p_w \) and \( p_v \), respectively. For simplicity we will also assume that \( \{w(t)\} \) and \( \{v(t)\} \) are mutually independent. The input \( u(t) \) may be a function of \( Y_t = [y(1), \ldots, y(t)] \) (with \( Y_0 = y_0 = \emptyset \)) and some other external signals. This means that feedback is allowed. In order to avoid to clutter up the notation, we will therefore use the convention that knowing \( Y_t \) allows us to compute \( u(t) \). However, this does not
mean that we need to know the feedback mechanism, we simply need to know \(u(t)\) at time \(t\). The initial state \(x(0)\) is assumed to be independent of \(\{w(t)\}\) and \(\{v(t)\}\) with distribution \(p(x(0))\).

Assuming that \(p(x(t)|Y_t)\) is available, \(p(x(t + 1)|Y_t)\) is given as the integral

\[
p(x(t + 1)|Y_t) = \int p(x(t + 1), x(t)|Y_t)dx(t) = \int p(x(t + 1)|x(t), Y_t)p(x(t)|Y_t)dx(t)
\]

\[
= \int p_w(x(t + 1) - f_t(x(t), u(t)))p(x(t)|Y_t)dx(t)
\]

(5.20)

This is called the time update, or prediction, step. Given \(p(x(t + 1)|Y_t)\) and \(y(t + 1)\) we can compute \(p(x(t + 1)|Y_{t+1})\) from the measurement update given by

\[
p(x(t + 1)|Y_{t+1}) = \frac{p(x(t + 1), y(t + 1)|Y_t)}{p(y(t + 1)|Y_t)} = \frac{p(y(t + 1)|x(t + 1), Y_t)p(x(t + 1)|Y_t)}{p(y(t + 1)|Y_t)}
\]

\[
= \frac{p_v(y(t + 1) - h_{t+1}(x(t + 1)))p(x(t + 1)|Y_t)}{p(y(t + 1)|Y_t)}
\]

(5.21)

where

\[
p(y(t + 1)|Y_t) = \int p(y(t + 1)|x(t + 1), Y_t)p(x(t + 1)|Y_t)dx(t + 1)
\]

\[
= \int p_v(y(t + 1) - h_{t+1}(x(t + 1), u(t + 1))p(x(t + 1)|Y_t)dx(t + 1)
\]

(5.22)

These equations should be compared with straightforward marginalization where we would write the output explicitly in terms of \(W = [w(0) \ w(1) \ \ldots \ w(N - 1)]^T\) and \(V = [v(1) \ \ldots \ v(N)]^T\):

\[
Y = \begin{bmatrix}
h_1(x_1(0), w(0)) + v(1) \\
h_2(x_2(0), w(0), w(1)) + v(2) \\
\vdots \n\end{bmatrix} + h_N(x_N(0), w(0), w(1), \ldots, w(N)) + v(N)
\]

and then performing the \(N\) dimensional integration

\[
p_y(y) = \int \cdots \int p_v(y(1) - h_1(x_1(0), w(0)))dw(0) \cdots dw(N - 1) \cdots
\]

\[
p_v(y(N) - h_N(x_N(0), w(0), \ldots, w(N - 1)))
\]

\[
= \int p_v(y(1) - h_1(x_1(0), w(0)))) \cdots \left( \int p_v(y(N - 1) - h_{N-1}(x_{N-1}(0), w(0), \ldots, w(N - 2)))
\]

\[
\left( \int p_v(y(N) - h_N(x_N(0), w(0), \ldots, w(N - 1)))dw(N - 1) \right) dw(N - 2) \cdots dw(0)
\]

The Kalman Filter

For linear state-space models subject to Gaussian disturbances and measurement noise the integrals above can be computed explicitly. Details are given in Appendix 5.C. Let

\[
x(t + 1) = F_t x(t) + G_t u(t) + w(t)
\]

\[
y(t) = H_t x(t) + v(t)
\]

(5.23)
5.3. ESTIMATION METHODS

where

\[ x(0) \sim N(\bar{x}_0, \Sigma_0), \quad \begin{bmatrix} w(t) \\ v(t) \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q_t & S_t \\ S_t^T & R_t \end{bmatrix} \right), \quad R_t > 0, \quad Q_t \geq 0 \]

\[ \mathbb{E} \begin{bmatrix} w(t) \\ v(t) \end{bmatrix} \left[ \begin{bmatrix} w(t + \tau) \\ v(t + \tau) \end{bmatrix} \right]^T = 0, \quad \tau \neq 0, \quad \mathbb{E} \left[ \begin{bmatrix} w(t) \\ v(t) \end{bmatrix} e_o^T \right] = 0, \quad t \geq 0 \]

Then

\[ x(t + 1)|Y_t \sim N(\hat{x}(t + 1|t), \Sigma(t + 1|t)) \quad (5.24) \]
\[ y(t + 1)|Y_t \sim N(H_{t+1}\hat{x}(t + 1|t), H_{t+1}\Sigma(t + 1|t)H_{t+1} + R_{t+1}) \quad (5.25) \]

where

\begin{align*}
\hat{x}(t + 1|t) = \bar{F}_t \hat{x}(t|t - 1) + G_t u(t) + S_t R_t^{-1} y(t) + K_t (y(t) - H_t \hat{x}(t|t - 1)) \\
\Sigma(t + 1|t) = \bar{F}_t \Sigma(t|t - 1) \bar{F}_t^T - \bar{F}_t \Sigma(t|t - 1) H_t^T (R_t + H_t \Sigma(t|t - 1) H_t^T)^{-1} H_t \Sigma(t|t - 1) \bar{F}_t^T + \bar{Q}_t \\
K_t = \bar{F}_t \Sigma(t|t - 1) H_t^T (R_t + H_t \Sigma(t|t - 1) H_t^T)^{-1} \\
\end{align*}

(5.26) (5.27)

where \( \bar{F}_t = F_t - S_t R_t^{-1} H_t \) and \( \bar{Q}_t = Q_t - S_t R_t^{-1} S_t \), \( \hat{x}(1|0) = \bar{x}_0 \), and \( \Sigma(1|0) = \Sigma_0 \). The equation for \( \Sigma(t + 1|t) \) is known as a Riccati equation.

Observe that the one-step ahead predictor \( \hat{y}(t|t - 1) \) is linear in the observations \( Y_{t-1} \). This is consistent with Exercise 2.2.1 which shows that for Gaussian distributed random variables the optimal predictor is linear in the observations. In turn, from this it follows that in the non-Gaussian case the Kalman filter is the best linear predictor.

The Innovations Model

Introducing \( e(t) := y(t) - \hat{y}(t|t - 1) \), we can use \( \hat{x}(t|t - 1) \) as state-variable giving

\begin{align*}
\hat{x}(t + 1|t) = \bar{F}_t \hat{x}(t|t - 1) + G_t u(t) + S_t R_t^{-1} y(t) + K_t e(t), \quad \hat{x}(t + 1|t) \sim N(\bar{x}_0, \Sigma_0) \\
y(t) = H_t \hat{x}(t|t - 1) + e(t) \\
\mathbb{E} \left[ e(t) e^T(t) \right] = H_t \Sigma(t|t - 1) H_t + R_t \\
\end{align*}

(5.29) (5.30) (5.31)

From the orthogonality principle it follows that the \( e(t) \) are uncorrelated, which in the Gaussian case means that they are independent. The Kalman filter thus handles the non-invertibility of the model by forming a disturbance that is orthogonal to the past. For this reason \( \{e(t)\} \) is called the innovations sequence.

The Stationary Kalman Filter

In the time-invariant case

\[ x(t + 1) = F x(t) + G u(t) + w(t) \\
y(t) = H x(t) + v(t) \quad (5.32) \]

with \( Q_t = Q \) etc, \( \Sigma(t + 1|t) \rightarrow \Sigma \) under certain conditions, where \( \Sigma \) must satisfy the Discrete Algebraic Riccati Equation (DARE)

\[ \Sigma = \bar{F} \Sigma \bar{F}^T - \bar{F} \Sigma H^T (R + H \Sigma H^T)^{-1} H \Sigma \bar{F}^T + \bar{Q} \quad (5.33) \]
where \( \bar{Q} = Q - SR^{-1}S^T \). In this situation the Kalman filter converges to the stationary Kalman filter

\[
\dot{x}(t+1|t) = \bar{F}\dot{x}(t|t-1) + Gu(t) + SR^{-1}y(t) + K(y(t) - H\dot{x}(t|t-1))
\]

(5.34)

\[
= (\bar{F} - KH)\dot{x}(t|t-1) + Gu(t) + (SR^{-1} + K)y(t)
\]

(5.35)

where

\[
\bar{F} = F - SR^{-1}H
\]

(5.36)

\[
K = \bar{F}\Sigma H^T(R + H\Sigma(t|t-1)H^T)^{-1}
\]

(5.37)

Often the time-varying filter is approximated by the stationary filter. As the DARE is a nonlinear equation it may have multiple solutions and it is therefore important to find the solution that corresponds to the limit of the time-varying filter. It may also happen that the stationary filter is not stable which renders this approximation useless. Conditions for this are given in the next theorem for which we need to introduce some systems theoretic concepts.

**Definition 5.3.1.** The pair \((F,H)\) is detectable if there exists \(K\) such that \(F - KH\) has all eigenvalues strictly inside the unit circle.

Detectability of a state-space model means that all unstable modes of the system can be observed in the output so that there exists a stabilizing output feedback \(u(t) - Ky(t)\).

**Definition 5.3.2.** A pair \((F,D)\) is unit circle controllable if

\[
xF = \lambda x, \quad |\lambda| = 1 \quad \Rightarrow xD \neq 0
\]

For stable models, unit circle controllability can be expressed in terms of the output spectrum

\[
\Phi_y(e^{i\omega}) = [H(e^{i\omega}I - F)^{-1} \quad I] \begin{bmatrix} Q & S^T \\ S & R \end{bmatrix} [H(e^{i\omega}I - F)^{-1} \quad I]^*,
\]

**Lemma 5.3.1** (Lemma 8.3.1 in [7]). Consider the time-invariant model (5.32) where \(F\) is stable and where

\[
\mathbb{E} \begin{bmatrix} w(t) \\ v(t) \end{bmatrix} \begin{bmatrix} w(t) \\ v(t) \end{bmatrix}^T = \begin{bmatrix} Q & S^T \\ S & R \end{bmatrix}
\]

Let \( \bar{Q} = Q - SR^{-1}S^T \) be factorized as \( \bar{Q} = DD^T \). Then \((\bar{F},D)\) unit circle controllable \(\iff\) \(\Phi_y(e^{i\omega}) > 0, \forall \omega\)

where \(\bar{F}\) is defined by (5.36).

Intuitively, if the model is not unit circle controllable, there are modes on the unit circle which are not excited by the disturbance and hence the stationary Kalman filter obtains no information about these modes and cannot account for these in its state estimate.

**Theorem 5.3.1** (Theorem 3.2 in [3]). Consider the model (5.23). Then detectability of \((F,H)\) is equivalent to that there exists a solution \(\Sigma_S \geq 0\) to the DARE (5.33) such that \(\bar{F} - KH\) has all eigenvalues inside or on the unit circle. There is only one solution with these properties.

Furthermore, detectability of \((F,H)\) and unit circle controllability of \((\bar{F},D)\), with \(\bar{F}\) defined by (5.36) and \(D\) by the factorization \(\bar{Q} = DD^T\), is equivalent to that there exists a solution \(\Sigma_S \geq 0\) to (5.33) such that \(\bar{F} - KH\) has all eigenvalues inside the unit circle. There is only one solution with these properties.
An important conclusion of Theorem 5.3.1 is that for stable models strict positivity of the output spectrum is necessary for the stationary Kalman filter to be stable.

Conditions for when $\Sigma(t|t-1)$ converges to $\Sigma_S$ defined in Theorem 5.3.1 are given next.

**Theorem 5.3.2** (Theorems 4.1 and 4.2 in [3]). Suppose that $(F,H)$ is detectable. Then $\Sigma_0 \geq \Sigma^*$ implies

$$\Sigma(t|t-1) \to \Sigma_S$$

Furthermore, if $\Sigma_0 > 0$, then detectability of $(F,H)$ and a strictly positive output spectrum is equivalent to that

$$\Sigma(t|t-1) \to \Sigma_S$$

with the convergence being exponentially fast.

**Spectral Factorization**

The innovations representation of the stationary Kalman filter

$$\dot{x}(t+1|t) = \dot{F}\dot{x}(t|t-1) + Gu(t) + SR^{-1}y(t) + Ke(t) \quad (5.38)$$

$$y(t) = H\dot{x}(t|t-1) + e(t) \quad (5.39)$$

can be written

$$\dot{x}(t+1|t) = \dot{F}\dot{x}(t|t-1) + Gu(t) + SR^{-1}(H\dot{x}(t|t-1) + e(t)) + Ke(t)$$

$$= F\dot{x}(t|t-1) + Gu(t) + (K + SR^{-1})e(t) \quad (5.40)$$

$$y(t) = H\dot{x}(t|t-1) + e(t) \quad (5.41)$$

which in transfer function form can be expressed as

$$y(t) = G(q)u(t) + H(q)e(t) \quad (5.42)$$

where

$$G(q) = H(qI - F)^{-1}G \quad (5.43)$$

$$H(q) = I + H(qI - F)^{-1}(K + SR^{-1})^{-1} \quad (5.44)$$

We notice that $H$ is monic, i.e. $H(\infty) = I$. Furthermore, using the matrix inversion lemma, Lemma 5.A.1,

$$H^{-1}(q) = (I + H(qI - F)^{-1}(K + SR^{-1}))^{-1} = I - H((K + SR^{-1})H + qI - F)^{-1}(K + SR^{-1})$$

$$= I - H(qI - (\bar{F} - KH))^{-1}(K + SR^{-1}) \quad (5.45)$$

Since $\bar{F} - KH$ is stable under appropriate conditions, see Theorem 5.3.1, $H^{-1}(q)$ is stable. This means that all zeros of $H(q)$ are strictly inside the unit circle, i.e. $H$ is minimum phase.

Suppose now that $F$ is stable. Then if (5.32) is initiated with $\Sigma_0 = \Sigma_S$, then the stochastic term in $y(t)$ (caused by $w(t)$, $v(t)$ and $x(0)$) is stationary with representation $H(q)e(t) = \sum_{k=0}^{\infty} h_k e_{t-k}$, with
\( \{e(t)\} \) being an iid sequence with zero mean and covariance matrix corresponding to the stationary covariance matrix in (5.25), i.e.

\[
\mathbb{E} [e(t)e^T(t)] = R + H\Sigma_SH^T
\]

We the Kalman filter provides a factorization of the spectrum of the stochastic term in \( y(t) \) as

\[
H(e^{i\omega})(R + H\Sigma_SH^T)H^*(e^{i\omega})
\]

where \( H \) is monic, stable and inversely stable. This is known as spectral factorization. When the input \( u(t) \equiv 0 \), the representation

\[
y(t) = H(q)e(t)
\]

is the Wold decomposition when \( \{y(t)\} \) is stationary, i.e. when either the model was initiated at time \( t = -\infty \) or when the initial state \( x(0) \) corresponds to stationary conditions, i.e. its covariance matrix satisfies

\[
\Sigma_0 = \bar{F}\Sigma_0\bar{F}^T + SR^{-1}H\Sigma_0 + \tilde{Q}
\]

We summarize our findings.

**Theorem 5.3.3.** Suppose that for the state-space model (5.23), \((F, H)\) is detectable and that \((\bar{F}, D)\) is unit circle controllable. Then there exists a \( \Sigma_S \) such that the corresponding stationary Kalman filter is stable where the transfer function (5.44) from the innovations sequence to the output is monic and inversely stable.

If in addition \( F \) is stable, then the condition of unit circle controllability can be replaced by strict positivity of the output spectrum. In this case (5.44) is stable.

**Relation to the Direct Method**

In this section we assume that \( y \) is scalar. Assuming that the input is zero, the Kalman filter computes row vectors \( \tilde{L}_{t-1} \in \mathbb{R}^{t-1} \) such that

\[
\hat{y}(t|t-1) = \tilde{L}_{t-1}Y_{t-1}
\]

Using (5.17), the negative log likelihood function is thus

\[
V_1 = \frac{1}{2} \sum_{t=1}^{N} (y(t) - \tilde{L}_{t-1}Y_{t-1})^T (H_t\Sigma(t|t-1)H_t^T + R_t)^{-1} (y(t) - \tilde{L}_{t-1}Y_{t-1}) + \frac{1}{2} \sum_{t=1}^{N} \log \det (H_t\Sigma(t|t-1)H_t^T + R_t)
\]

Let us now return to the direct method. For Gaussian data, i.e.

\[
Y \sim \mathcal{N}(0, \Sigma_Y)
\]

the negative log-likelihood is

\[
V_2 = \frac{1}{2} Y^T \Sigma_Y^{-1} Y + \frac{1}{2} \log \det \Sigma_Y
\]
Now, both $V_1$ and $V_2$ are the likelihoods of the same quantity and should therefore be equal. To reveal the connection let us factor $\Sigma_{YY}$ into what is known as a LDL decomposition

$$\Sigma_{YY} = LDL^T$$

where $D$ is diagonal (with elements $\{d_t\}$ and $L$ is lower triangular with ones in the diagonal. For a positive semi-definite matrix such a decomposition always exists and is unique. We can thus write the likelihood $V_2$ as

$$V_2 = \frac{1}{2} Y^T L^{-T} D^{-1} L^{-1} Y + \frac{1}{2} \log \det LDL^T$$

Now

$$\det LDL^T = \det L \det D \det L^T = \det D = \prod_{t=1}^N d_t$$

since $\det L = 1$ due to the triangularity and the ones in the diagonal. Furthermore, with $L$ being lower triangular $L^{-1}$ is also lower triangular and it is easy to verify that $L^{-1}$ also has ones in the diagonal. Row $t$ of $L^{-1}$ thus has the structure

$$[-L_t \ 1 \ 0_{1 \times N-t}]$$

so that

$$L^{-1} Y = \begin{bmatrix} y(1) \\ y(2) - L_2 Y_1 \\ \vdots \\ y(N) - L_N Y_{N-1} \end{bmatrix}$$

Using this gives

$$V_2 = \frac{1}{2} Y^T L^{-T} D^{-1} L^{-1} Y + \frac{1}{2} \log \det LDL^T$$

$$= \frac{1}{2} \sum_{t=1}^N (y(t) - L_t Y_{t-1})^T d_t^{-1} (y(t) - L_t Y_{t-1}) + \sum_{t=1}^N \log \det d_t$$

Comparing with $V_1$ we see that it must hold that $L_t = \hat{L}_t$ and that $d_t = H_t \Sigma(t|t-1) H_t^T + R_t$. We see that what the Kalman filter does is to invert the covariance matrix of $Y$ in a recursive manner. Now the expensive part of one step in the Kalman filter is the computation of the inverse of $H_t \Sigma(t|t-1) H_t^T + R_t$ in the update of $\Sigma(t|t-1)$. The computational complexity of this is proportional to $n_y^3$, where $n_y$ being the dimension of the output. Thus in terms of the sample size $N$, the computational complexity of the Kalman filter is $O(N^3)$ as already mentioned.

### The Non-Gaussian Case

Under non-Gaussian assumptions, the Kalman filter no longer provides the conditional distributions of $x(t)|Y_{t-1}$ and $y(t)|Y_{t-1}$ and neither are $\hat{x}(t|t-1)$ and $\hat{y}(t|t-1)$ := $H\hat{x}(t|t-1)$ the conditional
means of these distributions. Instead, $\hat{x}(t|t-1)$ and $\hat{y}(t|t-1)$ will be the best linear estimators based on past observations and $\Sigma(t|t-1)$ is still the covariance of the estimation error $x(t) - \hat{x}(t|t-1)$.

There is one exception to this and that is when the noise model is invertible and the initial state is known. Then $\hat{x}(t|t-1)$ and $\hat{y}(t|t-1)$ will still be the conditional means but the distribution of the noise has to be used to compute the conditional pdfs.

**Example 5.3.8.** Consider again the AR-1 model

$$y(t) + \theta y(t-1) = e(t)$$

which we can write on state-space form as

$$x(t+1) = y(t)$$
$$y(t) = -\theta x(t) + e(t)$$

which has $Q = 0$. Assuming that we know $y(0)$ exactly, $\Sigma_0 = 0$ and the Kalman filter will be stationary with $\Sigma(t|t-1) = 0$, $K_t = 0$, i.e.

$$\hat{x}(t|t-1) = y(t-1)$$
$$\hat{y}(t|t-1) = -\theta \hat{x}(t|t-1)$$

so that $y(t) - \hat{y}(t|t-1) = e(t)$.

The stationary Kalman filter will still correspond to the Wold decomposition but now $\{e(t)\}$ will only be an orthogonal (uncorrelated) sequence and not an independent sequence.

**Unstable Systems**

A remarkable fact with the Kalman filter is that the conditions for convergence to $\Sigma_S$ do not require the system to be stable. Thus, the covariance of the estimation error will stay bounded under these conditions regardless of how the actual state behaves. This will hold even if the time-varying Kalman filter is approximated with the stationary Kalman filter if the system operates under stabilizing feedback. This makes the Kalman filter very useful in control applications.

Returning to the transfer function form (5.42) of the stationary Kalman filter, we notice from (5.43)–(5.44) that $H(q)$ shares the poles with $G(q)$. Now, input-output models, minimal transfer function models can be used and hence, in terms of noise models, the stationary Kalman filter extends the class of rational noise models (5.7) which we, following the Wold decomposition, had constrained to be stable, to noise models which share unstable poles with the input-output model $G(q)$.

**5.3.2 The Prediction Error Method**

Already in Section 2.5 we outlined the idea of using estimation for parameter estimation. We will now pursue this idea for dynamical systems. The basic idea is to form a predictor $y(t|t-1; \theta)$ of the output using the past observations and the model and if this predictor is carefully constructed the prediction error variance

$$\mathbb{E} [(y(t) - y(t|t-1; \theta))^2]$$
should be minimized if the model \( \theta \) is taken as the true parameter. Now, since we cannot compute the expectation above as it depends on the true unknown parameter \( \theta_0 \) we form the sample estimate based on our measurements

\[
V_N(\theta) := \frac{1}{N} \sum_{t=1}^{N} (y(t) - y(t|t-1; \theta))^2
\]

and minimize this with respect to \( \theta \) instead. As we will see later, the better predictor we use the more accurate will our estimate be. Thus ideally we would like to use the conditional mean \( \mathbb{E}[y(t)|Y_{t-1}] \) as predictor. We also remark that other loss functions than the quadratic proposed above can be used, although this loss function is by far the most frequently used.

### 5.3.3 Relation to Maximum Likelihood Estimation

There is a strong connection between the prediction error approach and ML. Recall the expression (5.17) for the negative log-likelihood. Now \( \hat{y}(t|t-1) \) is a function of the past and, introducing \( e(t) = y(t) - \hat{y}(t|t-1) \), therefore

\[
p_{Y(t)|Y_{t-1}}(y(t)|y_{t-1}) = p_{e(t)|Y_{t-1}}(y(t) - \hat{y}(t|t-1)|y_{t-1})
\]

giving the negative log-likelihood

\[
-\log p_Y(y) = -\sum_{t=2}^{N} p_{e(t)|Y_{t-1}}(y(t) - \hat{y}(t|t-1)|y_{t-1}) - \log p(y(1))
\]

(5.47)

This is now not exactly a prediction error method as the terms in the loss function depends on the data not only through the prediction error \( y(t) - \hat{y}(t|t-1) \). However, the model (5.9) where the innovations sequence \( \{e(t)\} \) is assumed iid with pdf \( p_e \) gives \( p_{e(t)|Y_{t-1}}(y(t) - \hat{y}(t|t-1)|y_{t-1}) = p_e(y(t) - \hat{y}(t|t-1)) \) and hence

\[
-\log p_Y(y) = -\sum_{t=2}^{N} p_e(y(t) - \hat{y}(t|t-1)) - \log p(y(1))
\]

(5.48)

Thus, ML corresponds to using \( -\log p_e \) as loss function.

Below we will discuss predictors for different models.

### Linear Models

Suppose that

\[
v(t) = H(q)e(t) = \sum_{k=0}^{\infty} h_k e(t-k)
\]

(5.49)

where \( \{e(t)\} \) is a sequence of zero mean independent random variables with pdfs \( \{f_E(t)\}_{t=-\infty}^{\infty} \) and where \( H(z) \) is stable and that

\[
\frac{1}{H(z)} = \sum_{k=0}^{\infty} \tilde{h}_k z^{-k}
\]
is analytic in $|z| \geq 1$. Then we can define the filter $H^{-1}(q) = \sum_{k=0}^{\infty} \tilde{h}_k z^{-k}$ which will be stable, i.e. $\sum_{k=0}^{\infty} |\tilde{h}_k| < \infty$. We can then compute $e(t)$ through (see Lemma 3.1 in [11])

$$e(t) = H^{-1}(q)v(t) = \sum_{k=0}^{\infty} \tilde{h}_k v(t - k)$$

and hence

$$v(t) = H(q)e(t) = e(t) + (H(q) - 1)e(t) + e(t) + (H(q) - 1)H^{-1}(q)v(t) = e(t) + (1 - H^{-1}(q))v(t)$$

Now, noticing that $\tilde{h}_0 = 1$ (why?)

$$\hat{v}(t|t - 1) := (1 - H^{-1}(q))v(t) = -\sum_{k=1}^{\infty} \tilde{h}_k v(t - k)$$

and hence $\hat{v}(t|t - 1)$ depends only on the past. Furthermore, as $e(t)$ is independent of $e(t - 1), e(t - 2), \ldots$ it is independent of $v(t - 1), v(t - 2), \ldots$ and hence

Suppose now that we have the model

$$y(t) = G(q)u(t) + v(t), \quad v(t) = H(q)e(t)$$

where we assume that $\{u(t)\}$ is known. Then

$$\hat{v}(t|t - 1) = (1 - H^{-1}(q))v(t) = (1 - H^{-1}(q))(y(t) - G(q)u(t))$$

and therefore the conditional mean of $y(t)$ given the past is given by

$$\hat{y}(t|t - 1) = G(q)u(t) + \hat{v}(t|t - 1)$$

$$= G(q)u(t) + (1 - H^{-1}(q))(y(t) - G(q)u(t))$$

$$= H^{-1}(q)G(q)u(t) + (1 - H^{-1}(q))y(t)$$

This predictor assumes that the past from $t = -\infty$ is available. To account for non-stationary and initial conditions, the Kalman filter has to be used. As already noticed, except for the Gaussian case, the Kalman filter no longer provides the conditional mean predictor but the best linear predictor, except when the model is invertible.

One may also use predictors that are not optimal in any respect. For example the output-error model

$$\hat{y}(t) = G(q)u(t),$$

which is the mean of $y(t)$, can be used as a predictor.

**Nonlinear Models**

For non-invertible non-linear models, the conditional mean can be obtained from non-linear filtering. For the nonlinear state space model (5.18) the equations (5.20) and (5.21) can be used. These are computationally intractable and so called particle filter methods have been developed to cope with this. We will come back to this later in the course.
5.3. ESTIMATION METHODS

However, one may also use simpler predictors. We may, as we did in the linear case, for example use the mean of the model.

We may also use the optimal linear (in $Y$) predictor. We can compute all prediction errors directly using (5.46). Suppose that the model has mean $\mu(\theta)$ and covariance matrix $\Sigma_{YY}(\theta)$ for our observation vector $Y$. Then from the $LDL^T$ factorization $\Sigma_{YY}(\theta) = LDL^T$ we get

$$L^{-1}(Y - \mu(\theta)) = \begin{bmatrix}
y(1) - \mu_1(\theta) \\
y(2) - L_2Y_1 - \mu_2(\theta) \\
\vdots \\
y(N) - L_NY_{N-1} - \mu_N(\theta)
\end{bmatrix}$$

(5.50)

5.3.4 The Extended Invariance Principle
5.3.5 Multi-Step Least-Squares Methods
5.3.6 Instrumental Variable Methods
5.3.7 Indirect Inference
5.3.8 Exercises

The exercises below are based on the following simple MA-1 (Moving Average of order 1) model

$$y(t) = e(t) + ce(t - 1)$$

(5.51)

where $\{e(t)\}$ is a zero mean sequence with $\mathbb{E}[e^2(t)] = \lambda_t$

5.3.1. The Kalman filter.

(a) Write (5.51) on state-space form using $x(t) = e(t - 1)$ as state.

(b) Suppose that $\{e(t)\}$ is Gaussian distributed. What is the optimal predictors of $x(t + 1)$ and $y(t + 1)$ given $Y_t = [y(1) \ldots y(t)]^T$?

(c) Suppose that $\lambda_t = \lambda$ for $t = 0, 1, \ldots$. Which solutions exist to the algebraic Riccati equation (ARE)? Express the stationary Kalman filter on transfer function form

$$\hat{y}(t|t - 1) = L(q)y(t)$$

When is this filter stable?

(d) Suppose that $\lambda_t = \lambda$ for $t = 0, 1, \ldots$. Analyze the behavior of the Kalman filter you derived in (b) as $t \to \infty$. Distinguish the three cases i) $|c| < 1$, ii) $|c| > 1$ and iii) $|c| = 1$. For which cases will you be able to asymptotically (as $t \to \infty$) be able to estimate the state exactly? Verify this via a simulation.

(e) Suppose that $\lambda_t = \lambda$ for $t = 0, 1, \ldots$. Suppose that $|c| < 1$ and that $y(s), s < t$ is available. Use transfer function formalism to derive the optimal one-step ahead predictor of $y(t)$. Compare with your analysis in (d).

(f) * Suppose that $\lambda_t = \lambda$ for $t = 0, 1, \ldots$. Suppose that $|c| > 1$ and that $y(s), s < t$ is available. Use transfer function formalism to derive the optimal one-step ahead predictor of $y(t)$. Compare with your analysis in (d).
CHAPTER 5. DYNAMICAL MODELS

(g) * Suppose that \( \lambda_t = \lambda \) for \( t = 0, 1, \ldots \). Suppose that \(|c| = 1\) and that \( y(s), s < t \) is available. Use the Kalman filter you derived in (b) to derive the optimal one-step ahead predictor of \( y(t) \).

5.3.2. Estimation using stationary data. Suppose that \( \lambda_t = \lambda \) for \( t = 0, 1, \ldots \) and that \( \{e(t)\} \) is Gaussian distributed.

(a) What is the spectrum of \( \{y(t)\} \)?

(b) Perform spectral factorization of the spectrum you obtained in (a).

(c) Use the result from (b) to discuss if \( c \) and the noise variance \( \lambda \) can be uniquely determined from data. Also discuss the case when \( \lambda \) is known.

(d) What is the distribution of \( Y_N = [y(1) \ldots y(N)]^T \)?

(e) Suppose that the noise variance \( \lambda \) is known. Use your result from (d) to implement the ML estimator for \( c \) in Matlab. How large sample size \( N \) can you handle?

(f) Implement the ML estimator using the Kalman filter you derived in Exercise 5.3.1b. Verify that you get exactly the same estimate as in (e). How large sample sizes \( N \) can you handle now?

(g) Argue that the Kalman filter implementation of the ML estimator corresponds to a prediction error method using the optimal time-varying predictor.

(h) Implement an approximation of the ML-estimator using the stationary Kalman filter (Exercise 5.3.1e) with zero initial conditions for the predictor instead of the optimal predictor. Make a Monte Carlo study, comparing the MSE of this estimator and the exact ML-estimator for the case \( c = 0.8 \) and \( \lambda = 1 \). Use \( N = 5, 10, 20, 30, 40, 50, 100, 200, 500, 1000, 10000 \). Comment on your results.

5.3.3. Estimation using non-stationary data. Identifiability under Gaussian assumptions. Suppose that \( \lambda_t = \lambda \) for \( t = 1, \ldots, N \) but that \( \lambda_0 = 0 \) and that \( \{e(t)\} \) is Gaussian distributed.

(a) What is the distribution of \( Y_N = [y(1) \ldots y(N)]^T \)?

(b) Implement the ML-estimator starting from the expression you have obtained in (a).

(c) Show that in this case there is a one-to-one correspondence between \( Y_N \) and \([e(0) \ldots e(N-1)]^T \). Use this to make an alternative derivation of the ML-estimator.

(d) Implement the ML-estimator using the Kalman filter you have implemented in Exercise 5.3.1b. The Kalman filter equations depend on the noise variance \( \lambda \) but do you need to know \( \lambda \) in this case? Show that the Kalman filter is unstable when \(|c| > 1\). Does this mean that the filter is useless or incorrect?

(e) * In the stationary case we saw that it is not possible to estimate both \( c \) and \( \lambda \) consistently. Assume now that we have a computer with infinite precision. Would it then be possible to estimate \( \lambda \) and \( c \) consistently as \( N \to \infty \) in the non-stationary setting we consider in this exercise? If so, how?

(f) * Show that \( \lambda \) and \( c \) are uniquely determined by the distribution. Compare with the stationary case studied in Exercise 5.3.2.

5.3.4. * Estimation using non-stationary data. Comparison between ML and PEM. Suppose that \( \lambda_t = \lambda \) for \( t = 0, 1, \ldots, N - 1 \) but that \( \lambda_N = 0 \) and that \( \{e(t)\} \) is Gaussian distributed.
5.3. ESTIMATION METHODS

(a) What is the distribution of \( Y_N = [y(1) \ldots y(N)]^T \)?

(b) Suppose that \( \lambda \) is known. Implement the ML-estimator of \( c \) using your result in (a).

(c) In Exercise 5.3.2g we saw that in the stationary case the ML estimator can be interpreted as a prediction error method where the optimal time-varying predictor is used.

In the non-stationary case we are considering in this exercise, what is the PEM estimator using the optimal predictor? Is it the same as the ML-estimator you have derived above?

(d) Perform a Monte Carlo study comparing the ML-estimator and PEM using the optimal time-varying predictor. Use \( c = 0.8 \) and \( \lambda = 1 \) and \( N = 5, 10, 20, 30, 40, 50, 100, 200, 500 \).

Comment on your results. In particular consider that the only difference with the stationary case is that here \( e(N) = 0 \) and not \( e(N) \sim N(0, \lambda) \).

(e) A drawback with your implementation of the ML-estimator is that an \( N \times N \) matrix needs to be inverted. Show that the Kalman filter can be used to avoid this. Implement this version of the ML-estimator and verify that it gives exactly the same estimates as your previous implementation.

5.3.5. * Estimation under non-Gaussian assumptions: Non-stationary case. Suppose that \( \lambda_t = \lambda \) for \( t = 1, \ldots, N \) but that \( \lambda_0 = 0 \) and that \( \{e(t)\} \) is Laplace distributed, i.e.

\[
p_e(x) = \frac{1}{\sqrt{2\lambda}} e^{-\sqrt{2\lambda}|x|}
\]

(a) Assume \( \lambda \) known. Use the Kalman filter to implement the ML-estimator for \( c \).

(b) Perform a Monte Carlo study comparing the ML-estimator with the ML-estimator based on Gaussian assumptions. Use \( c = 0.9, \lambda = 1 \) and \( N = 5, 10, 20, 30, 40, 50, 100, 200, 500, 1000 \).

5.3.6. * Estimation under non-Gaussian assumptions: Stationary case. Suppose that \( \lambda_t = \lambda \) for \( t = 0, \ldots, N \) and that \( \{e(t)\} \) is Laplace distributed, i.e.

\[
p_e(x) = \frac{1}{\sqrt{2\lambda}} e^{-\sqrt{2\lambda}|x|}
\]

(a) Determine the probability density function for \( y(1) \).

(b) Use (5.17), Exercise 5.3.6a and Exercise 5.3.1d to construct a cost function that approximates the negative log-likelihood function. Which is the range of \( c \) for which the method is expected to work?

(c) Perform a Monte Carlo study of the algorithm for \( c = 0.8, \lambda = 1 \) and \( N = 2000 \) where you compare the MSE of the method with the ML-estimator for the Gaussian case you developed in Exercise 5.3.2f. Does your new estimator have better accuracy than the one based on Gaussian assumptions?

Repeat the simulation study but now with Gaussian noise.

5.3.9 Computer exercises

You do not need to hand in any solutions to these exercises - just make sure you understand the results you obtain.

The purpose of these exercises is both to understand some important concepts and to familiarize yourself with the Matlab System Identification Toolbox.
5.3.1. In `computerexercise1.mat` the data set \( z \) contains data from the system \( G, H \) (also supplied). The system is of ARX-type.

(a) Plot the Bodediagrams of \( G \) and \( H \).

(b) It is good practice to examine the data first:

i. Plot the data in the time domain: `plot(z)`. What is your impression of the input? Does it seem to be persistently exciting?

ii. To further examine the input estimate a time-series model for it: `mu=arx(iddata(u),4)`. Plot the spectrum of the estimated model: `spectrum(mu)`. Which type of spectrum does the input seem to have?

iii. Discuss how you can use system identification to estimate if the data has been generated in open or closed loop. Base your approach only on the assumption that (as in this example) the true system contains at least one delay. Try your approach. Do you see any indication of feedback?

(c) Estimate an ARX-model with the true orders and delay: `marx=arx(z,[3 3 1])`. Use `bodeplot(G,marx)` to plot the resulting model and to compare it with the true system in the frequency domain. How is the fit?

(d) All information about the estimated model is contained in `marx`. For example the covariance matrix of the parameter estimate is retrieved as `marx.covar`. In itself it is not very informative. However, the toolbox uses it to (approximately) compute the uncertainty of other quantities by way of Gauss approximation formula (Theorem 3.2.1). The \( 3\sigma \) (3 standard-deviations) uncertainty of the model in the frequency domain is added to the Bodediagram by `bodeplot(G,marx,'sd',3)`.

i. Does the uncertainty interval cover the true system?

ii. The asymptotic black-box theory gives the approximate expression

\[
\text{Cov} \hat{G}_N(e^{i\omega}) \approx \frac{n}{N} \Phi_v(\omega)\Phi_u(\omega)
\]

This expression is valid for high model orders \( n \) but can be used in general as a coarse approximation of the uncertainty. Use \( H \) and your estimate of the input spectrum to assess whether this expression is reasonable for `marx`.

(e) Use `moe=oe(z,[3 3 1])` to estimate an output error model with the correct orders for \( G \). Compare the model with \( G \) in the frequency domain, including the model uncertainty.

i. Does the model seem to be biased? Explain this using the expression

\[
\mathbb{E}[e^2(t,\theta)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} |G_o(e^{i\omega}) - G(e^{i\omega},\theta)|^2 \frac{\Phi_u(\omega)}{|H(e^{i\omega},\theta)|^2} d\omega + |H_o(e^{i\omega}) - H(e^{i\omega},\theta)|^2 |H(e^{i\omega},\theta)|^2 d\omega + \lambda_o \quad (5.52)
\]

which holds for the system/model

\[
y(t) = G_o(q)u(t) + H_o(q)e_o(t), \quad \mathbb{E}[e_o(t)]^2 = \lambda_o \\
y(t) = G(q,\theta)u(t) + H(q,\theta)e(t)
\]
when the data has been generated in open loop. The expression (5.52) follows from
\[
\varepsilon(t, \theta) = H^{-1}(q, \theta)(y(t) - G(q, \theta)u(t))
\]
\[
= \frac{G_o(q) - G(q, \theta)}{H(q, \theta)} u(t) + \frac{H_o(q)}{H(q)} e_o(t)
\]
\[
= \frac{G_o(q) - G(q, \theta)}{H(q, \theta)} u(t) + \frac{H_o(q) - H(q)}{H(q)} e_o(t) + e_o(t)
\]

ii. How is the uncertainty compared to the ARX-model. Explain the difference.

(f) Use `mfir=arx(z,[10 0 1])` to estimate a finite impulse response model of order 10.

i. Is this model structure able to model $G$?

ii. Compare the estimated model with $G$ in the frequency domain. The model appears to at least try to model the peak in the frequency response. Right? Explain this using (5.52).

iii. Why is the fit not better at the peak?

iv. Add $3\sigma$-uncertainty bounds to the Bode diagram. Do the uncertainty bounds capture the model error? The explanation to this is that the toolbox only computes the variance error and does that based on the assumption that the model structure captures the true system.

(g) Suppose that we are only interested in the frequency dip around 0.25 rad/s. A natural idea is to bandpass filter the data around this frequency and then do the estimation with the prefiltered data.

i. Verify that $L$ could be a suitable filter by comparing it with the true system in the frequency domain. To only plot the magnitude use

```matlab
1 h=bodeplot(G,L);
2 setoptions(h,'PhaseVisible','off');
```

ii. Filter both $u$ and $y$ with $L$ and create a new data set using `iddata` and then estimate a new 10th order FIR model using this data.

iii. Compare the estimated model with $G$ in the frequency domain. Explain what you see.

iv. Formulate a model
\[
y(t) = \bar{G}(q, \theta)u(t) + \bar{H}(q, \theta)e(t)
\]

for which the PEM gives exactly the model you have obtained using prefiling. Conclude that prefiling can be given a modeling interpretation.

v. Estimate an ARX model using $n_a = 3$, $n_b = 3$ and $n_k = 1$ (as before) using the prefiltered data. How is the model fit? Why?

vi. Which model structure is required to recover the true system from the prefiltered data? Try this model structure. Here you may need the command `pem(data,[na nb nc nd nf nk])` to estimate a general polynomial model from the `iddata` object `data`. How does it work? Why?
(h) Estimate a first order ARX-model: \( \text{marxloworder=arx(z,[1 1 1])} \) and compare it with the true system in the frequency domain. Use (5.52) to explain why the fit is good at high frequencies.

5.3.2. The file computerexercise2.mat contains another data set \( z \) from the same system as in the previous exercise.

(a) Look at the data, examine if there seems to be feedback, and if so deduce which type of controller that is being used, and then estimate an ARX-model with \( n_u = n_b = 3 \) and \( n_k = 1 \) as well as an output error model with \( n_f = n_b = 3 \) and \( n_k = 1 \). Compare the models and discuss what the differences compared to the previous exercise are and what they depend on.

(b) Examine which of the subspace methods CVA, SSARX and MOSEP works on this data. The function call is

```matlab
1 mcva=n4sid(z,3,'DisturbanceModel','Estimate',...  
2 n4sidOptions('N4Weight','CVA'))
```

etc.

5.3.3. The order of persistence of excitation of a signal \( \{u(t)\} \) can be checked using the rank condition in Lemma 13.1. Show that an alternative way is to estimate FIR models of increasing order, comparing the loss functions. Compare these two approaches for some signals consisting of a sum of sinusoids.

5.3.4. In Chapter 4 we discussed biased estimators. In this exercise we will compare parametric models with biased high order FIR models estimated using Empirical Bayes. The approach is a generalization of the series of examples appearing in Section 4.5, starting with Example 4.5.1

Our starting point is the FIR model

\[
y(t) = \sum_{k=1}^{n} b_k u(t - k) + e(t), \quad t = 1, \ldots, N
\]

which we can write in vector form as

\[
Y = \Phi \theta + E,
\]

where \( \theta \) is a vector consisting of the impulse response coefficients, \( \theta = [b_1 \ldots b_n]^T \), and \( \Phi \) is lower Toeplitz built from the input sequence.

(a) **Regularization vs MAP.** The toolbox has functionality for regularization where instead of minimizing the prediction error criterion \( V_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} e^2(t, \theta) \),

\[
V_N(\theta) + J(\theta),
\]

where \( J \) does not depend on data, is minimized with respect to \( \theta \). Show that this can be interpreted as Maximum A Posteriori (MAP) estimation of \( \theta \) under the model

\[
-\log p_Y|\theta(y|\theta) \propto V_N(\theta), \quad -\log p_\theta(\theta) \propto J(\theta)
\]
5.3. ESTIMATION METHODS

(b) MAP vs posterior mean. The regularized cost-function used in the toolbox is given by

\[ V_N(\theta) + \frac{\lambda}{N} \theta^T R \theta \]

where \( \lambda > 0 \) and where \( R \) is a positive definite matrix.

Conclude from the preceding problem that this cost function corresponds to Gaussian observations \( E \sim \mathcal{N}(0, \lambda_e I) \) and Gaussian prior

\[ \theta \sim \mathcal{N}(0, P) \]

where

\[ P = \frac{\lambda_e}{\lambda} R^{-1} \]

(c) * Show that under the assumptions of the preceding problem, the MAP estimator coincides with the Bayes estimator (i.e. the posterior mean).

(d) The data set \( z \) in computerexercise3.mat contains input-output data from the first order system

\[ y(t) = \frac{q^{-1}}{1 - 0.9q^{-1}} u(t) + e(t) \]

Estimate a FIR model of order 100 and plot and compare its impulse response with the true impulse response. As you can see the estimates are very noisy.

You can now try to decrease the variance of the estimates using regularization in the following way:

```plaintext
1. optridge = arxOptions;
2. optridge.Regularization.Lambda = 100;
3. optridge-Regularization.R = eye(nfir);
4. mfirridge = arx(z, [0 nfir 1], optridge);
```

This means that the regularization term is

\[ \frac{100}{N} \theta^T \theta = \frac{100}{N} ||\theta||^2 \]

i.e. the squared magnitude of all impulse response coefficients are penalized. Plot and compare the estimated impulse response with the true one. What do you observe? Try to manually tune \( \lambda \) so as to get as good fit as possible to the true impulse response (don’t spend too much time on this). To measure the quality of the estimate you can use

\[ \text{norm}(G-mfirridge) \]

Compare your result with a first order output error model. Would you say that the result of the regularization/Bayes estimator is satisfactory?

(e) Choice of kernel. We saw in Section 4.8 that the variance \( P(\eta) \) of the prior (aka kernel) is used as a model of \( \theta^T \eta \) to estimate the pointwise risk. Thus the more accurately \( P(\eta) \) can model the true \( \theta_o \theta_o^T \), the better pointwise risk estimate is obtained and the smaller is the resulting MSE.
Consider a first order system

\[ G(q) = \frac{bq^{-1}}{1-pq^{-1}} = b \sum_{k=1}^{\infty} p^{k-1} q^{-k} \]

Then the impulse response coefficients satisfy

\[ \theta \theta^T = b^2 \begin{bmatrix} \rho^0 & \ldots & \rho^{n-1} \\ \rho^1 & \ldots & \rho^n \\ \vdots & \ddots & \vdots \\ \rho^{n-1} & \ldots & \rho^{2n-2} \end{bmatrix} \]

This could thus be a suitable structure of a kernel. However, since it is rank one it will force the estimate in this direction - the prior dictates the direction of the vector of impulse response coefficients and experience has shown that this results in poor estimates. An alternative kernel but with similar structure is the following

\[ P_{dc} := \lambda \begin{bmatrix} 1 & r^1 & r^2 & \ldots & r^{n-1} \\ r^1 & r^1 & r^2 & \ldots & r^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r^{n-1} & r^{n-1} & r^{n-1} & \ldots & r^{n-1} \end{bmatrix} \]

This kernel is full rank and known as the DC-kernel.

Let us construct such a kernel for our first order system. We notice that

\[ \{P_{dc}\}_{i,j} = \lambda r^{\max(i,j)-1} \]

We match the \((1, 1)\) element of \(P_{dc}\) with the square of the first impulse response coefficient, resulting in the following code:

```matlab
1 g=impulse(G,nfir);  
2 g=g(2:end);  
3 r=0.9;  
4 lambdadc=g(1);  
5 ix=ones(nfir,1)*(1:nfir);  
6 Pdc=lambdadc*r.^(max(ix,ix')-1);  
7 Rdc=inv(lambdadc/lambdae*Pdc);  
8 optdc = arxOptions;  
9 optdc.Regularization.Lambda =lambdadc;  
10 optdc.Regularization.R = Rdc;  
11 mfirdc = arx(z, [0 nfir 1], optdc);  
```

Compare the resulting estimate with the previous estimates.

(f) In the toolbox, the hyperparameters \(\lambda\) and \(r\) can be estimated using marginalized ML, cf. Section 4.5.9. This gives the Empirical Bayes estimate.

Redo the regularized model estimation using

```matlab
1 [LEB,REB] = arxRegul(z,[0 100 ... 0],arxRegulOptions('RegularizationKernel','DC'));  
```
instead.
How does this estimate compare with the previous ones?

5.3.5. Biased estimation of high order systems. In the previous problem, EB performs worse than a first order output-error model. A case where EB can be highly competitive is when the true order is high, especially when most of the poles and zeros are close to the origin and therefore do not contribute little to the dynamics. In computerexercise4.mat the data set \( z \) has been generated by a 30th order system \( G \). Estimate a 100th order FIR model using Empirical Bayes as well as a full (30th) order output error model. Compare their impulse responses and Bode diagrams with the true system \( G \). Also determine the order of an output error model (use the same order for numerator and denominator) which minimizes the MSE of the frequency response (use \( \text{norm}(G\text{-model}) \)). Compare this model with the Empirical Bayes estimate. Conclusions?

Multi-Step Least-Squares Methods

5.3.10 The Prediction Error Method

Multi-Step Least-Squares Methods

Subspace Identification

Instrumental Variable Methods

Bayesian Methods

Time versus Frequency Domain Identification

Continuous Time Model Identification

5.4 Nonlinear Models

5.5 Probabilistic Models

5.A Matrix Inversion Lemma

Next follows a useful result on matrix inversion.

**Lemma 5.A.1.** Suppose that \( A \in \mathbb{R}^{n \times n} \) and \( C \in \mathbb{R}^{m \times m} \) are invertible. Then for \( B \in \mathbb{R}^{n \times m} \), \( D \in \mathbb{R}^{m \times n} \)

\[
(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}
\]  
(5.54)

5.B Completing the Square

**Lemma 5.B.1.** Suppose that \( P_y > 0 \) and \( P_x > 0 \). Then

\[
(y - Ax)^T P_y^{-1}(y - Ax) + (x - z)^T P_x^{-1}(x - z) = (x - \hat{x})^T S^{-1}(x - \hat{x}) + (y - Az)^T T^{-1}(y - Az)
\]
where
\[ T = (P_y + AP_x A^T) \]
\[ S = (P_x^{-1} + A^T P_y^{-1} A)^{-1} = P_x - P_x A^T T^{-1} AP_x \]
\[ \dot{x} = z + L(y - Az) \]
\[ L = P_x A^T T^{-1} \]

Proof. We have
\[ (y - Ax)^T P_y^{-1} (y - Ax) + (x - z)^T P_x^{-1} (x - z) \]
\[ = x^T (P_x^{-1} + A^T P_y^{-1} A)x - 2x^T (A^T P_y^{-1} y + P_x^{-1} z) + y^T P_y^{-1} y + z^T P_x^{-1} z \]
\[ = (x - Sw)^T S^{-1} (x - Sw) - w^T Sw + y^T P_y^{-1} y + z^T P_x^{-1} z \]  
(5.55)

where
\[ w = A^T P_y^{-1} y + P_x^{-1} z \]  
(5.57)

Using the Miatrix Inversion Lemma (5.54),
\[ T^{-1} = (P_y + AP_x A^T)^{-1} = P_y^{-1} - P_y^{-1} AS A^T P_y \]  
(5.58)

and furthermore, using Exercise 4.9.1,
\[ P_x^{-1} - P_x^{-1} SP_x^{-1} = P_x^{-1} S (S^{-1} - P_x^{-1}) = P_x^{-1} (P_x^{-1} + A^T P_y^{-1} A)^{-1} A^T P_y^{-1} A \]
\[ = (I + A^T P_y^{-1} A P_x)^{-1} A^T P_y^{-1} A \]
\[ = A^T P_y^{-1} (I + A P_x A^T P_y^{-1})^{-1} A \]
\[ = A^T (P_y + A P_x A^T)^{-1} A = A^T T^{-1} A \]  
(5.59)

Also
\[ P_x^{-1} S A^T P_y^{-1} = P_x^{-1} (P_x^{-1} + A^T P_y^{-1} A)^{-1} A^T P_y^{-1} = (I + A^T P_y^{-1} A P_x)^{-1} A^T P_y^{-1} \]
\[ = A^T (I + P_y^{-1} A P_x A^T)^{-1} P_y^{-1} = A^T (P_y + A P_x A^T)^{-1} = A^T T^{-1} \]  
(5.60)

Using (5.58)–(5.60) gives
\[ - w^T Sw + y^T P_y^{-1} y + z^T P_x^{-1} z = \]
\[ y^T (P_y^{-1} - P_y^{-1} AS A^T P_y^{-1}) y + z^T (P_x^{-1} - P_x^{-1} S P_x^{-1}) z - 2z^T (P_x^{-1} S A^T P_y^{-1}) y \]
\[ y^T T^{-1} y + z^T A^T T^{-1} A z - 2z^T A^T T^{-1} y = (y - Az)^T T^{-1} (y - Az) \]  
(5.61)

Next
\[ Sw = (P_x - P_x A^T T^{-1} AP_x) (A^T P_y^{-1} Ay + P_x^{-1} z) \]
\[ = P_x A^T T^{-1} (T - AP_x A^T) P_y^{-1} y + z - P_x A^T T^{-1} Az \]
\[ = P_x A^T T^{-1} y + z - P_x A^T T^{-1} Az \]  
(5.62)

Inserting (5.61) and (5.62) in (5.55) now gives the result. \[\square\]
5.C Derivation of the Kalman Filter

5.B.1 Expectation of a Normal PDF

Lemma 5.B.2. Let \( N(y; m, P) \) denote the pdf of a normal distribution with mean \( m \in \mathbb{R}^n \) and covariance \( P \). Let \( X \sim N(m, P_x) \), \( X \in \mathbb{R}^m \). Then

\[
\mathbb{E}[N(y; AX, P_y)] = N(y; Am, AP_x A^T) \tag{5.63}
\]

Proof. Completing the square using Lemma (5.5.1) gives

\[
\mathbb{E}[N(y; AX, P_y)] = \int \frac{1}{(2\pi)^{n/2}\sqrt{\det P_y}} e^{-\frac{1}{2}(y-AX)^T P_y^{-1}(y-AX)} \mathbb{E}\left[\frac{1}{(2\pi)^{m/2}\sqrt{\det P_x}} e^{-\frac{1}{2}(x-m)^T P_x^{-1}(x-m)}\right] dx
\]

\[
= \int \frac{1}{(2\pi)^{(n+m)/2}\sqrt{\det P_y P_x}} e^{-\frac{1}{2}((y-AX)^T P_y^{-1}(y-AX) + (x-m)^T P_x^{-1}(x-m))} dx
\]

\[
= \int \frac{1}{(2\pi)^{(n+m)/2}\sqrt{\det P_y P_x}} e^{-\frac{1}{2}((y-Am)^T T^{-1}(y-Am) + (x-\hat{x})^T S^{-1}(x-\hat{x}))} dx
\]

where \( T, S \) and \( \hat{x} \) are defined in Lemma (5.5.1). We can further write this as

\[
\mathbb{E}[N(y; AX, P_y)] = \frac{1}{(2\pi)^{(n)/2}\sqrt{\det P_y P_x S}} e^{-\frac{1}{2}(y-Am)^T T^{-1}(y-Am)} \int N(x; \hat{x}, S) dx
\]

\[
= \frac{1}{(2\pi)^{(n)/2}\sqrt{\det P_y P_x S^{-1}}} e^{-\frac{1}{2}(y-Am)^T T^{-1}(y-Am)} \tag{5.65}
\]

but

\[
\det P_y P_x S^{-1} = \det P_y P_x (P_x^{-1} + A^T P_y^{-1} A)^{-1} = \det P_y (I + A^T P_y^{-1} A P_x)
\]

\[
= \det P_y \det (I + A^T P_y^{-1} A P_x) = \det P_y \det (I + A P_x A^T P_y^{-1})
\]

\[
= \det(P_y (I + A P_x A^T P_y^{-1})) = \det(P_y + A P_x A^T) = \det T \tag{5.66}
\]

Inserting this in (5.65) gives the result. \( \square \)

5.C Derivation of the Kalman Filter

In this appendix we will derive the Kalman filter for the state-space model

\[
x(t + 1) = F_t x(t) + G_t u(t) + w(t)
\]

\[
y(t) = H_t x(t) + v(t)
\]

where

\[
x(0) \sim N(\bar{x}_0, \Sigma_0), \quad \begin{bmatrix} w(t) \\ v(t) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q_t & S_t \\ S_t^T & R_t \end{bmatrix}\right), \quad R_t > 0, \quad Q_t \geq 0
\]

\[
\mathbb{E}\left[\begin{bmatrix} w(t) \\ v(t) \end{bmatrix} \begin{bmatrix} w(t + \tau) \\ v(t + \tau) \end{bmatrix}^T\right] = 0, \quad \tau \neq 0, \quad \mathbb{E}\left[\begin{bmatrix} w(t) \\ v(t) \end{bmatrix} x(0)^T\right] = 0, \quad t \geq 0
\]

The input \( u(t) \) may be a function of \( Y_t = [y(1), \ldots, y(t)] \), with \( Y_0 = y(0) = 0 \), and some other external signals. This means that feedback is allowed. In order to avoid to clutter up the notation, we
will therefore use the convention that knowing $Y_t$ allows us to compute $u(t)$. However, this does not mean that we need to know the feedback mechanism, we simply need to know $u(t)$ at time $t$.

As a preparatory step we explicitly incorporate the cross-correlation between $w$ and $v$. To this end introduce

$$\bar{w}(t) = w(t) - S_t R_t^{-1} v(t)$$

Obviously $E[\bar{w}(t)v^T(t)] = 0$ so that $\bar{w}$ and $v$ are independent. Furthermore,

$$\tilde{Q}_t := E[\bar{w}(t)\bar{w}^T(t)] = Q_t + S_t R_t^{-1} R_t^{-1} S_t^T - 2 S_t R_t^{-1} S_t^T = Q_t - S_t R_t^{-1} S_t^T$$

and we can write

$$x(t + 1) = F_t x(t) + w(t) G_t u(t) - S_t R_t^{-1} (H_t x(t) + v(t) - y(t))$$

$$= \bar{F}_t x(t) + \bar{w}(t) + G_t u(t) + S_t R_t^{-1} y(t)$$

where $\bar{F}_t = F_t - S_t R_t^{-1} H_t$.

There are several ways to derive the Kalman filter, for example using orthogonal projections, but here we will use the filtering equations

$$p(x(t + 1)|Y_t) = \int p(x(t + 1)|x(t)) p(x(t)|Y_t) dx(t)$$

(5.68)

$$p(x(t + 1)|Y_{t+1}) = \frac{p(y(t + 1)|x(t + 1)) p(x(t + 1)|Y_t)}{p(y(t + 1)|Y_t)}$$

(5.69)

$$p(y(t)|Y_{t-1}) = \int p(y(t)|x(t)) p(x(t)|Y_{t-1}) dx(t)$$

(5.70)

To begin with we notice that

$$p(y(t)|x(t)) = p_y(y(t) - H_t x(t)) = N(y(t); H_t x(t), R_t)$$

(5.71)

$$p(x(t + 1)|x(t), Y_t) = p_x(x(t + 1) - \bar{F}_t x(t) - G_t u(t) - S_t R_t^{-1} y(t))$$

$$= N(x(t + 1), \bar{F}_t x(t) + G_t u(t) + S_t R_t^{-1} y(t), \tilde{Q}_t)$$

(5.72)

We will derive the filter using induction. We begin by noticing that

$$x(0)|Y_0 = x(0) \sim N(\hat{x}(0|0), \Sigma(0|0)), \quad \hat{x}(0|0) := \bar{x}_0, \Sigma(0|0) = \Sigma_0$$

(5.73)

so let us assume that

$$x(t)|Y_t \sim N(\hat{x}(t|t), \Sigma(t|t))$$

(5.74)

for some $\hat{x}(t|t)$ and $\Sigma(t|t)$.

Let us now consider the time update (5.68), which we can write as

$$p(x(t + 1)|Y_t) = E\left[ N(x(t + 1) - G_t u(t) - S_t R_t^{-1} y(t); \tilde{F}_t x(t), \tilde{Q}_t) \right], \quad \text{where } X \sim N(\hat{x}(t|t), \Sigma(t|t))$$

Lemma 5.B.2 gives that

$$p(x(t + 1)|Y_t) = N(x(t + 1) - G_t u(t) - S_t R_t^{-1} y(t); \bar{F}_t \hat{x}(t|t), \tilde{Q}_t + \tilde{F}_t \Sigma(t|t) \tilde{F}_t^T)$$

$$= N(x(t + 1); \bar{F}_t \hat{x}(t|t) + G_t u(t) + S_t R_t^{-1} y(t), \tilde{Q}_t + \tilde{F}_t \Sigma(t|t) \tilde{F}_t^T)$$

$$= N(x(t + 1); \hat{x}(t + 1|t), \Sigma(t + 1|t))$$
where

\[
\dot{x}(t+1|t) := \bar{F}_t \dot{x}(t|t) + G_t u(t) + S_t R_t^{-1} y(t)
\]

\[
\Sigma(t+1|t) := \bar{Q}_t + \bar{F}_t \Sigma(t|t) \bar{F}_t^T
\]

Thus

\[x(t+1) | Y_t \sim N(\dot{x}(t+1|t), \Sigma(t+1|t))\]

Next, \( p(x(t+1)|Y_{t+1}) \) is given by the measurement update (5.69). We notice that \( x(t+1) \) depends only on the factors \( p(y(t+1)|x(t+1)) p(x(t+1)|Y_t) \), and then only through the exponential factor

\[
e^{-\frac{1}{2}((y(t+1)-H_{t+1}x(t+1))^T R_{t+1}^{-1} (y(t+1)-H_{t+1}x(t+1)) + (x(t+1) - \hat{x}(t+1|t))^T \Sigma^{-1}(t+1|t)(x(t+1) - \hat{x}(t+1|t)))}
\]

and by completing the square by way of Lemma 5.B.1 (identifying \( A \leftrightarrow H_{t+1}, z \leftrightarrow \hat{x}(t+1|t) \), \( P_x = \Sigma(t+1|t), P_y \leftrightarrow R_{t+1} \)) we can write this as

\[
e^{-\frac{1}{2}((x(t+1) - \hat{x}(t+1|t))^T \Sigma^{-1}(t+1|t)(x(t+1) - \hat{x}(t+1|t)))}
\times e^{-\frac{1}{2}((y(t+1)-H_{t+1}\hat{x}(t+1|t))^T (R_{t+1} + H_{t+1}\Sigma(t+1|t) H_{t+1}^T)^{-1} (y(t+1)-H_{t+1}\hat{x}(t+1|t)))}
\]

where

\[
\dot{x}(t+1|t+1) = \dot{x}(t+1|t) + L_{t+1} (y(t+1) - H_{t+1}\hat{x}(t+1|t))
\]

\[
L_{t+1} = \Sigma(t+1|t) H_{t+1}^T (R_{t+1} + H_{t+1}\Sigma(t+1|t) H_{t+1}^T)^{-1}
\]

\[
\Sigma(t+1|t+1) = \Sigma(t+1|t) - \Sigma(t+1|t) H_{t+1}^T (R_{t+1} + H_{t+1}\Sigma(t+1|t) H_{t+1}^T)^{-1} H_{t+1} \Sigma(t+1|t)
\]

Since \( p(x(t+1)|Y_{t+1}) \) is a density and \( x(t+1) \) appears only in the exponent, where it appears in quadratic form, it must hold that

\[x(t+1) | Y_{t+1} = N(\dot{x}(t+1|t+1), \Sigma(t+1|t+1)) \quad (5.75)\]

We have thus shown that (5.74) implies (5.75) and since (5.74) holds for \( t = 0 \) by way of (5.73), induction now gives the result.

Finally, we derive the one-step output predictor using (5.70), which can be written

\[
p(y(t)|Y_{t-1}) = \mathbb{E}[N(y(t); H_t X, R_t)], \quad \text{where} \quad X \sim N(\dot{x}(t|t-1), \Sigma(t|t-1))
\]

Lemma 5.B.2 gives that

\[
p(y(t)|Y_{t-1}) = N(y(t); H_t \dot{x}(t|t-1), R_t + H_t \Sigma(t|t-1) H_t^T)
\]

so that

\[y(t)|Y_{t-1} = N(H_t \dot{x}(t|t-1), R_t + H_t \Sigma(t|t-1) H_t^T)\]
Chapter 6

Model Quality

6.1 Variance Quantification

6.1.1 Fundamental Geometric Principles

6.1.2 Fundamental Structural Results

6.1.3 Variability of Estimated Frequency Response

A General Principle

Details for FIR Models

Details for ARMAX Models

Details OE and BJ Models

6.1.4 Variability of Nonlinear System Estimates

6.1.5 Bootstrap Methods

6.1.6 Exercises

6.1.6.1 In the lectures we have used a geometric approach to analyze the asymptotic covariance matrix $P$. Now, $P$ is a just a matrix with some structure, more specifically

$$P = \lambda \langle \Psi, \Psi \rangle^{-1}, \quad \langle \Psi, \Psi \rangle = E\psi(t)\psi^T(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Psi(e^{i\omega})\Psi^*(e^{i\omega})d\omega$$

so of course all the analysis can be done using matrix algebra.

Use matrix algebra to prove that the upper left $n \times n$ block of $P$ does not decrease when the number of estimated parameters is increased from $n$ to $n + m$, $m > 0$. It is here assumed that then the first $n$ rows of $\psi(t)$ (and $\Psi(q)$) remain the same. What is the a necessary and sufficient condition for there to be no increase?

6.1.6.2 Let

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t), \quad y(t) \in \mathbb{R}^{n_y}$$

where $\{e(t)\}$ is assumed to be a sequence of zero mean independent normal distributed random variables with unknown covariance matrix $\mathbb{E}[e(t)e^T(t)] = \Gamma$. Derive the ML-criterion using
transfer function calculations - i.e. neglect any finite sample effects such as unknown intial conditions etc. Then show that it is possible to eliminate $\Gamma$ from the negative log-likelihood function and that the resulting criterion for $\theta$ corresponds to

$$V_N(\theta) = \log \det \sum_{t=1}^{N} \varepsilon(t, \theta)\varepsilon^T(t, \theta)$$

where $\{\varepsilon(t, \theta)\}$ are the prediction errors.

### 6.1.7 Computer exercises

6.1.7.1. Identify a model for the data set computerexercise5.mat. It is suggested to follow the procedure in the Help for the Matlab System Identification Toolbox found under

- System Identification Toolbox
  - Linear Model Identification
    * Linear Model Identification Basics
      - Model Structure Selection: Determining Model Order and Input Delay
Chapter 7

Experiment Design

7.1 Identifiability
7.2 Informative Experiments
7.3 Persistence of Excitation
7.4 Input Signal Design
7.4.1 Common Input Signals
    PRBS
    Sums of Sine-Waves and Crest Factor Correction
7.5 Application Oriented Experiment Design
7.6 Adaptive Experiment Design
Chapter 8

Model Validation

8.1 Residual whiteness Tests
8.2 Input to residual correlation tests
8.3 Model Error Modelling
Chapter 9

Applications

9.1 Closed Loop Identification
9.2 Network Models
9.3 Errors-in-Variables Models
9.4 Block-structured Nonlinear Models
Bibliography


