Dynamic Model Learning: A Geometric Perspective

Lecture Notes in FEL3201/FEL3202

Håkan Hjalmarsson

*-marked sections and exercises are for FEL3202
## Contents

1 Background Theory 7
1.1 Vector Spaces 7
  1.1.1 Inner Product Spaces 7
  1.1.2 Subspaces and Orthogonal Projections 9
  1.1.3 Hilbert Spaces 10
  1.1.4 Orthonormal bases 11
1.2 Probability Theory 12
  1.2.1 Information Contents in Events, Bayes Rule and Independence 13
  1.2.2 Random Variables 13
  1.2.3 Expectation 14
  1.2.4 Information Contents in Random Variables and Conditional Probabilities 14
  1.2.5 Independent Random Variables 17
1.3 References 17

2 Introduction to Estimation 19
2.1 Introduction 19
2.2 Minimum Mean Square Error Estimation of Random Variables 19
  2.2.1 Introduction 19
  2.2.2 The Internal Structure of the Conditional Expectation 20
  2.2.3 A Hilbert Space Interpretation 22
  2.2.4 Linear Estimators 24
  2.2.5 Maximum A Posteriori Estimation 25
  2.2.6 Other Estimation Criteria* 25
  2.2.7 Exercises 25
2.3 Unbiased Parameter Estimation 26
  2.3.1 Introduction 26
  2.3.2 Efficient Estimators 27
    Efficient Estimators for Exponential Families 28
    Efficient Estimators for General Distributions 29
  2.3.3 The Maximum Likelihood Estimator 29
  2.3.4 Data compression 30
    Maximal data compression 30
    Finite dimensional data compression 31
  2.3.5 Uniform Minimum Variance Unbiased (UMVU) Estimators 32
    Generalizing the Rao-Blackwell theorem* 33
  2.3.6 Efficient Estimators, Sufficient Statistics and Maximum Likelihood Estimators 34
CONTENTS

2.3.7 Parameter Transformations .................................................. 34
   Maximum Likelihood Estimators ............................................. 34
   Efficient Estimators .............................................................. 34
2.3.8 Best Linear Unbiased Estimators (BLUE) ................................. 36
2.3.9 Exercises .............................................................................. 38
2.4 Using Estimation for Parameter Estimation ................................. 39
2.5 References .............................................................................. 40
2.A Schur Complement .................................................................. 41

3 Asymptotic Theory ...................................................................... 43
3.1 Introduction ............................................................................... 43
3.2 Limits of Random Variables ..................................................... 44
   3.2.1 Convergence in Mean ....................................................... 44
   3.2.2 Convergence in Probability .............................................. 45
   3.2.3 Convergence with Probability 1 ......................................... 45
   3.2.4 Convergence in Distribution ............................................ 46
   3.2.5 Relations Between Convergence Concepts .......................... 46
3.3 Large Sample Properties of Estimators ...................................... 48
   3.3.1 Asymptotics of Explicit Estimators ................................. 48
   Consistency .............................................................................. 48
   Convergence in Distribution .................................................... 49
   3.3.2 Criterion Based Estimators .............................................. 50
   Consistency .............................................................................. 50
   Convergence in Distribution .................................................... 51
   3.3.3 Asymptotic Efficiency ....................................................... 53
   3.3.4 Superefficient Estimators* ............................................... 54
   3.3.5 The Maximum Likelihood Estimator ............................... 54
   Consistency .............................................................................. 54
   Asymptotic Efficiency .............................................................. 55
3.4 Parameter Transformations ...................................................... 57
   3.4.1 Consistent Estimators ....................................................... 57
   3.4.2 The Extended Invariance Principle (EXIP) ....................... 57
   3.4.3 Beyond the EXIP .............................................................. 60
3.5 Using Estimation for Parameter Estimation ................................. 63
   3.5.1 Exercises ....................................................................... 64
3.6 References ............................................................................... 65
3.A Asymptotic Dominance of the Variance Term .............................. 66
   3.A.1 MSE Dominating the CRLB .............................................. 66
   3.A.2 Optimal MSE ................................................................. 68

4 Minimum Mean Square Error Parameter Estimation ....................... 71
4.1 The Bias-Variance Error Trade-Off ............................................ 71
4.2 Risk Estimation Methods ......................................................... 71
4.3 Probabilistic Models and the Bayes Estimator ............................. 71
4.4 Linear in the Parameters Models .............................................. 71

5 Linear in the Parameters Models .................................................. 73
6 Dynamical Models
   6.1 Model Structures and Probabilistic Models ........................................... 76
   6.2 Estimation Methods ................................................................................. 76
      6.2.1 Maximum Likelihood Estimation ......................................................... 76
      6.2.2 The Extended Invariance Principle ..................................................... 76
      6.2.3 Multi-Step Least-Squares Methods ...................................................... 76
      6.2.4 The Prediction Error Method .............................................................. 76
      6.2.5 Instrumental Variable Methods ........................................................... 76
      6.2.6 Indirect Inference .............................................................................. 76
   6.3 Linear Models ......................................................................................... 76
      6.3.1 Maximum Likelihood Estimation ......................................................... 76
      Multi-Step Least-Squares Methods .............................................................. 76
      6.3.2 The Prediction Error Method .............................................................. 76
      Multi-Step Least-Squares Methods .............................................................. 76
      Subspace Identification ............................................................................. 76
      Instrumental Variable Methods .................................................................. 76
      Bayesian Methods ..................................................................................... 76
      Time versus Frequency Domain Identification ......................................... 76
      Continuous Time Model Identification ....................................................... 76
   6.4 Nonlinear Models ................................................................................... 76
   6.5 Probabilistic Models .............................................................................. 76

7 Model Quality ............................................................................................ 77
   7.1 Variance Quantification .......................................................................... 77
      7.1.1 Fundamental Geometric Principles ..................................................... 77
      7.1.2 Fundamental Structural Results .......................................................... 77
      7.1.3 Variability of Estimated Frequency Response ....................................... 77
         A General Principle ................................................................................ 77
         Details for FIR Models .......................................................................... 77
         Details for ARMAX Models .................................................................. 77
         Details OE and BJ Models .................................................................... 77
      7.1.4 Variability of Nonlinear System Estimates ........................................... 77
      7.1.5 Bootstrap Methods ............................................................................ 77

8 Experiment Design ...................................................................................... 79
   8.1 Identifiability ......................................................................................... 79
   8.2 Persistence of Excitation ....................................................................... 79
   8.3 Input Signal Design ............................................................................... 79
      8.3.1 Common Input Signals ...................................................................... 79
         PRBS ................................................................................................. 79
         Sums of Sine-Waves and Crest Factor Correction .................................... 79
   8.4 Application Oriented Experiment Design .............................................. 79
   8.5 Adaptive Experiment Design .................................................................. 79
## CONTENTS

### Model Validation
- 9.1 Residual whiteness Tests .................................................. 81
- 9.2 Input to residual correlation tests ....................................... 81
- 9.3 Model Error Modelling ....................................................... 81

### Applications
- 10.1 Closed Loop Identification ............................................... 83
- 10.2 Network Models ............................................................. 83
- 10.3 Errors-in-Variables Models ............................................... 83
- 10.4 Block-structured Nonlinear Models ................................. 83
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 \( \mathcal{V} \) of elements \( v \), called vectors, on which two operations \( + \) and \( \cdot \), called vector addition and scalar multiplication, are defined such that for all vectors \( u, v, w \in \mathcal{V} \) and scalars \( c, d \) it holds

1. Closure: \( u + v \in \mathcal{V} \)
2. Commutativity: \( u + v = v + u \)
3. Associativity: \( (u + v) + w = u + (v + w) \)
4. Additive identity: \( \mathcal{V} \) contains an element, denoted by \( 0 \), such that \( 0 + v = v \), \( \forall v \in \mathcal{V} \)
5. Additive inverse: There exists a unique \( x(v) \in \mathcal{V} \) such that \( v + x(v) = 0 \). \( x(v) \) is called \(-v\).
6. For any scalar \( c \), \( c \cdot v \in \mathcal{V} \)
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, \( \mathcal{V} \) is said to be a real vector space, and when \( c \in \mathbb{C} \), \( \mathcal{V} \) 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 \( \langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{C} \), which, mimicking the properties of the scalar product, has to satisfy the following axioms for all \( u, v, w \in \mathcal{V} \) 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 i) \( \|v\| \geq 0 \), \( \forall v \in \mathcal{V} \) 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 \( \mathcal{V} \) is an inner product space iff the parallelogram law

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

holds for all \( u, v \in \mathcal{V} \). 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 \( u \in \mathcal{V} \) define its orthogonal projection on another vector \( v \in \mathcal{V} \) 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

\[ 0 \leq \frac{|\langle u, v \rangle|}{\|u\|\|v\|} \leq 1 \] (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 \( \{v_k\}_{k=1}^n, n < \infty \), 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 \( V \), \( \dim \left[ V \right] \), 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

\[
\langle u - \sum_{k=1}^n \alpha_k v_k, v_l \rangle = 0, \quad l = 1, \ldots, n
\] (1.7)

which in matrix form becomes

\[
\begin{bmatrix}
\langle v_1, v_1 \rangle & \ldots & \langle v_1, v_n \rangle \\
\vdots & \ddots & \vdots \\
\langle v_n, v_1 \rangle & \ldots & \langle v_n, v_n \rangle
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\vdots \\
\alpha_n
\end{bmatrix}
= 
\begin{bmatrix}
\langle u, v_1 \rangle \\
\vdots \\
\langle u, v_n \rangle
\end{bmatrix}
\] (1.8)

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

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

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 \( V = S \oplus S^\perp \).
CHAPTER 1. BACKGROUND THEORY

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 neighbourhood \( \{ u : \| u - v \| < \varepsilon \} \subset S, \varepsilon > 0 \). 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 preceding 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 \( H \) and let \( u \in H \) be given. Then there is a unique vector \( v \in S \) such that \( u - v \bot w \) for all \( w \in S \). The vector \( v \) solves

\[
\min_{v \in S} \|u - v\|
\]

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

$$\mathcal{L}_u = \{x = u + v, \ v \in S\}$$

Then the problem

$$\min_{x \in \mathcal{L}_u} \|x\|$$

has a unique solution $u_{\perp S}$. The solution is the unique $x \in \mathcal{L}_u$ satisfying

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

1.1.4 Orthonormal bases

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

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

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

and from Pythagoras theorem (1.5) it follows that

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

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

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

(1.10)

follows. A remarkable consequence of this inequality is that if one has an uncountable orthonormal set $\mathcal{M}$ in an inner product space $\mathcal{V}$, then for a given $u \in \mathcal{V}$, at most a countable set of Fourier coefficients $\langle u, e \rangle$, $e \in \mathcal{M}$ 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 $\mathcal{M}$ 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 $\mathcal{H}$. Then this series is convergent, to $v$ say, iff $\sum_{k=1}^{\infty} |\alpha_k|^2$ is convergent. This follows since $\sum_{k=1}^{\infty} \alpha_k e_k$ then is a Cauchy sequence.

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

In particular it holds that for any $u \in \mathcal{H}$, $\sum_{k=1}^{\infty} \langle u, e_k \rangle e_k$ 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 $\mathcal{H}$. 
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 a \( F \in S^2 \) 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

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

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

\[
F = \bigcup_{m=1}^{\infty} \cap_{n=m}^{\infty} F_N
\]
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 $\mathcal{F}$. 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 $B \in B$, where $X \in B = \{\omega : X(\omega) \in B\}$. We can see $P_X(B)$ as a probability measure on the measurable space $(\mathbb{R}, B)$. The distribution function is defined as $F_X(x) := P_X((\infty, x))$. If $P_X$ is absolutely continuous with respect to the Lesbegue 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

$$E[X] = \int_{\Omega} Xd\mathcal{P} := \int_{\Omega} X(\omega)d\mathcal{P}(d\omega)$$

when the integral is defined. We will not give a precise meaning of $\mathcal{P}(d\omega)$ and it would take us to 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

$$E[X] = \int_{-\infty}^{\infty} xdP_X(dx) = \int_{-\infty}^{\infty} xF_X(dx)$$

When the sample space is discrete it follows that

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

and when the pdf exists

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

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
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 \( Y \in B \) changes the probability that \( X \in B \) to the conditional probability distribution function

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

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

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

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 = x, 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

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( P_{X \mid Y}(X = x \mid Y = y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>( \frac{P(X = -1, Y = -1)}{P(Y = -1)} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1 )</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>( \frac{P(X = 1, Y = -1)}{P(Y = -1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{2}{3} )</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>( \frac{P(X = -1, Y = 0)}{P(Y = -1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = 1 )</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>( \frac{P(X = 1, Y = 2)}{P(Y = -1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = 1 )</td>
</tr>
</tbody>
</table>

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 \( P_{X \mid Y}(B_X \mid y) \) 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

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

(1.15)

As noted above, in addition to (1.15) we also need that $P_{X|Y}(B_X|B_Y)$ 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

$$
\int_{\Omega} X(\omega)P_{X|Y}(d\omega|y)
$$

We can also define the conditional expectation of $X$ given $Y$ as

$$
E[X|Y](\omega) = \int_{\Omega} X(\tilde{\omega})P_{X|Y}(X(\tilde{\omega})|Y(\omega))
$$

(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 $X$:

$$
E[E[X|Y]] = \int_{\Omega} E[X|Y](\omega)P(d\omega) = \int_{\Omega} \int_{\Omega} X(\tilde{\omega})P_{X|Y}(X(\tilde{\omega})|Y(\omega))P(d\omega)
$$

$$
= \int_{\Omega} X(\tilde{\omega}) \int_{\Omega} P_{X|Y}(X(\tilde{\omega})|Y(\omega))P(d\omega)
$$

$$
= \int_{\Omega} X(\tilde{\omega})P \{ \omega : X(\omega) \in (X(\tilde{\omega})|\Omega) = \int_{\Omega} X(\tilde{\omega})P(d\omega) = E[X]
$$

**Example 1.2.4** (Example 1.2.3 continued). The conditional mean is given by

$$
E[X|Y = -1] = -1 \times P_{X|Y}(X = -1|Y = 1) + 1 \times P_{X|Y}(X = 1|Y = 1) = -1 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{1}{3}
$$

$$
E[X|Y = 0] = -1 \times P_{X|Y}(X = -1|Y = 0) + 1 \times P_{X|Y}(X = 1|Y = 0) = -1 \times 1 + 0 = -1
$$

$$
E[X|Y = 2] = -1 \times P_{X|Y}(X = -1|Y = 2) + 1 \times P_{X|Y}(X = 1|Y = 2) = 0 + 1 \times 1 = 1
$$

**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

$$E[X|Y] = \int_{-\infty}^{\infty} xp_{X|Y}(x|Y)dx$$


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

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

1.3 References

Section 1.1

Chapters 2-3 in [4].

Section 1.2

Chapters 1-3 in [1].
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

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

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)

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

The MSE can be split in two terms

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

(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)

$$MMSE := MSE[\hat{x}^*] = \mathbb{E}[\|X - \mathbb{E}[X]\|^2] = \mathbb{E}[X^2] - \mathbb{E}^2[X]$$
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

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

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

\[
\text{MSE}[\hat{x}|Y = y] = \mathbb{E}[|X - \mathbb{E}[X|Y = y]|^2|Y = y] + \mathbb{E}[X|Y = y] - \hat{x}^2
\]  
(2.2)

which again is minimized by taking \( \hat{x} \) to be the mean of \( X \), but now the conditional mean \( \mathbb{E}[X|Y = y] \). This means that the optimal \( \hat{x} \) now will depend on the observation \( y \): \( \hat{x} = \hat{x}^*(y) = \mathbb{E}[X|Y = y] \). The minimum MSE is now a function of \( y \) and given by

\[
\text{MMSE}(y) = \text{MSE}[\hat{x}^*(y)|Y = y] = \mathbb{E}[|X - \mathbb{E}[X|Y = y]|^2|Y = y] = \mathbb{E}[X^2|Y = y] - \mathbb{E}^2[X|Y = y]
\]  
(2.3)

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

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

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

\[
\text{MSE}[\hat{x}] = \mathbb{E}[|X - \hat{x}|^2] = \mathbb{E}_Y[\mathbb{E}[|X - \hat{x}|^2|Y]] = \int \mathbb{E}[|X - \hat{x}|^2|Y = y] p_Y(y)dy
\]  
(2.4)

\[
= \int \text{MSE}[\hat{x}|Y = y] p_Y(y)dy
\]  
(2.5)

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

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

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 \( \hat{x}(Y) = \hat{x}^*(Y) = \mathbb{E}[X|Y] \) will for each \( Y = y \) minimize the integrand, and hence minimize the MSE. Taking the expectation of (2.3) gives

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

\[
\int_A Z(\omega) = \int_A X(\omega)p(\omega) \quad \forall A \in \sigma(X)
\]  
(2.7)
2.2. MINIMUM MEAN SQUARE ERROR ESTIMATION OF RANDOM VARIABLES

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 $\{ A \in \sigma(A), \int_A X(\omega) P(d\omega) \}$ 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

$$\int_A Z((Y(\omega))) P(d\omega) = \int_A X(\omega) P(d\omega) \quad \forall A \in \sigma(Y) \quad (2.8)$$

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

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

showing that $E[X|Y](\omega)$ 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

$$\int_{A_Y} Z d\mathbb{P} = \int_{A_Y} X d\mathbb{P} = \mathbb{P}(A_X \cap A_Y)$$

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

$$\int_{A_Y} Z d\mathbb{P} = z \mathbb{P}(A_Y)$$

we have

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

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

$$Z = \begin{cases} \mathbb{P}_{X|Y}(B_X|B_Y) & Y \in B_Y \\ \mathbb{P}_{X|Y}(B_X|B_Y^c) & Y \in B_Y^c \end{cases}$$
Consider now the conditional expectation (1.16)

\[ \mathbb{E} [X|Y] (\omega) = \int_\Omega X(\bar{\omega}) \mathcal{P}_{X|Y}(X(d\bar{\omega})|Y(\omega)) \]

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

The preceding two examples suggests that the conditional expectation \( \mathbb{E} [X|Y] \) 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

\[ \int_{Y \in B_Y} X(\omega) \mathcal{P}(d\omega) = \mathcal{P}(X \in B_X \cap Y \in B_Y) \]

while

\[ \int_{Y \in B_Y} \mathbb{E} [X|Y] (\omega) \mathcal{P}(d\omega) = \int_{Y \in B_Y} \int_\Omega X(\bar{\omega}) \mathcal{P}_{X|Y}(d\bar{\omega}|Y(\omega)) \mathcal{P}(d\omega) \]

\[ = \int_{Y \in B_Y} \mathcal{P}_{X|Y}(B_X|Y(\omega)) \mathcal{P}(d\omega) = \mathcal{P}(X \in B_X \cap Y \in B_Y) \]

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 \( \mathcal{P} \) 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.

\[ \mathbb{E} [f(Y)|Y] = f(Y) \]

More generally we see that when \( X = VW \) where \( V \in \sigma(Y) \) and \( W \) is another random variable,

\[ \mathbb{E} [X|Y] = \mathbb{E} [VW|Y] = V \mathbb{E} [W|Y], \quad \forall V \in \sigma(Y) \]

(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 \( (\Omega, \mathcal{F}, \mathcal{P}) \) be a probability space and let \( \mathcal{H} \) be the vector space of random variables \( X \) defined on this probability space with

\[ \mathbb{E} [X^2] < \infty \]

endowed with the inner product

\[ (X, Y) = \mathbb{E} [XY] \]

(2.10)
2.2. *MINIMUM MEAN SQUARE ERROR ESTIMATION OF RANDOM VARIABLES*

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 \(E[X|Y]\) as candidate. We have for any \(W \in S\)

\[
\langle X - E[X|Y], W \rangle = E[(X - E[X|Y])W]
\]

\[
\]

\[
= E[XW] - E[WX] = 0 \quad (2.13)
\]

where the third equality follows from (2.9). We have thus proved that the conditional expectation is the orthogonal projection of \(X\) on \(S\)

\[
X_{\|S} = E[X|Y] \quad (2.14)
\]

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 \(E[X|Y]\) is a vector of the conditional expectations \(E[X_i|Y]\) and each element of \(X - E[X|Y]\) 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 \(E[X_{\|S}] = E[X]\) (recall that the conditional mean has the same mean as the random variable itself), we can re-write this as

\[
\text{Cov}[X - X_{\|S}] = \text{Cov}[X] - \text{Cov}[X_{\|S}] \quad (2.17)
\]
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 $W = Y_1, \ldots, Y_n$

$$\langle X - LY, Y_k \rangle = 0 \quad k = 1, \ldots, n$$ (2.18)

which, using (1.9), can be written

$$L\langle Y, Y^T \rangle = \langle X, Y^T \rangle$$ (2.19)

i.e. the orthogonal projection is given by

$$X_{\parallel S_L} = \langle X, Y^T \rangle \langle Y, Y^T \rangle^{-1} Y$$ (2.20)

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

$$\|X - X_{\parallel S_L}\|^2 \leq \|X - X_{\parallel S}\|^2$$

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:

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

The OLE can be expressed as

$$X_{\parallel S_{Le}} = \text{Cov}[X, Y] \text{Cov}^{-1}[Y] (Y - \mathbb{E}[Y]) + \mathbb{E}[X]$$

which has covariance

$$\text{Cov}[X_{\parallel S_{Le}}] = \text{Cov}[X, Y] \text{Cov}^{-1}[Y] \text{Cov}[Y, X]$$

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 $X_{\parallel S_{Le}}$ 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

$$\langle x, y \rangle \langle y, y \rangle^{-1} \langle y, x \rangle$$ (2.21)

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. MINIMUM MEAN SQUARE ERROR ESTIMATION OF RANDOM VARIABLES

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

\[
\hat{X}_{MAP} = \arg \max_{x} p_{X|Y}(x|y)
\]

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

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

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 \( \mathcal{N}(2, 1) \) 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 \( \hat{X} \) of \( X := \hat{\theta}(Y) \) on a subspace \( \mathcal{S} \) 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). \( \hat{X} \) will thus give a new unbiased estimator which has smaller MSE (in matrix sense) than the original \( \hat{\theta} \). However, if \( \hat{X} \) depends on \( \theta \) this is not an estimator so some care has to be exercised when determining a suitable subspace to project on.

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

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

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

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

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

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

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

We will call \( \langle S(Y; \theta), S^{T}(Y; \theta) \rangle \) the Fisher information matrix (FIM) \( I_{F}(\theta) \). Before proceeding we notice that

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

(2.22)

Notice now that

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

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

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
2.3. **UNBIASED PARAMETER ESTIMATION**

function is independent of which unbiased estimator we want to estimate. This “estimator” is given by

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

(2.23)

Here Pythagoras relation (2.17) becomes

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

where

\[
\text{Cov} \left[ \hat{\theta}(Y) \|_{S_{Le}} \right] = E \left[ I_F^{-1}(\theta) S(Y; \theta) S^T(Y; \theta) I_F^{-1}(\theta) \right] = I_F^{-1}(\theta)
\]  

(2.24)

and since a covariance matrix is positive semi-definite we have that

\[
\text{Cov} \left[ \hat{\theta}(Y) \right] \geq I_F^{-1}(\theta)
\]  

(2.25)

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

\[
\text{Cov} \left[ f(Y) \right] \geq \frac{\partial}{\partial \theta} E \left[ f(Y) \right] I_F^{-1}(\theta) \frac{\partial}{\partial \theta} E \left[ f(Y) \right]^T
\]  

(2.26)

We can use this result to derive a CRLB for biased estimators. Suppose that \( \hat{\theta}(Y) \) has bias \( b(\theta) \), i.e.

\[
E \left[ \hat{\theta}(Y) \right] = \theta + b(\theta)
\]

then (2.26) gives directly that

\[
\text{Cov} \left[ \hat{\theta}(Y) \right] \geq (I + b'(\theta)) I_F^{-1}(\theta) (I + b'(\theta))^T
\]  

(2.27)

### 2.3.2 Efficient Estimators

An unbiased estimator is efficient if it attains the CRLB. Equality in (2.25) will hold only if

\[
\hat{\theta}(Y) = \hat{\theta}(Y) \|_{S_{Le}} = I_F^{-1}(\theta) S(Y; \theta) + \theta \quad \text{a.e.}
\]  

(2.28)

but for this to be an estimator it cannot depend on \( \theta \).

For a given pdf \( p(Y; \theta) \) one can always check if (2.28) 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

$$p(y; \theta) = e^{\varphi(T(y)) - B(\theta)} h(y)$$  \hspace{1cm} (2.29)

where $\varphi(\cdot) : \mathbb{R}^s \to \mathbb{R}^s$ is 1-1 and twice differentiable. Introducing $\eta = \varphi(\theta)$ we can write the pdf on the canonical form

$$e^{\eta T(y) - A'(\eta)} h(y)$$  \hspace{1cm} (2.30)

where $A'(\eta) = B(\varphi^{-1}(\eta))$. For such a family it holds that $T(Y)$ has distribution

$$p(t; \eta) = e^{\eta t - A'(\eta) k(t)}$$  \hspace{1cm} (2.31)

for some positive function $k(t)$. This distribution has mean and covariance matrix

$$E[T(Y)] = A'(\eta) = A'(\varphi(\theta)) = (\varphi'(\theta))^T B'(\theta), \quad \text{Cov}[T(Y)] = A''(\eta) = A''(\varphi(\theta))$$  \hspace{1cm} (2.32)

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 (2.32), can be written as

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

giving the Fisher Information Matrix

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

Hence an efficient estimator has to have the form

$$\hat{\theta}(Y)_{\parallel S_{Le}} = I_F^{-1}(\theta) S(Y; \theta) + \theta$$  \hspace{1cm} (2.33)

$$= \left( (\varphi'(\theta))^T \text{Cov}[T(Y)] \varphi'(\theta) \right)^{-1} (\varphi'(\theta))^T (T(Y) - E[T(Y)]) + \theta$$

$$= (\varphi'(\theta))^{-1} \text{Cov}^{-1}[T(Y)] (T(Y) - E[T(Y)]) + \theta \hspace{1cm} (2.34)$$

In order for this quantity to be independent of $\theta$ we see that it is necessary that

$$\varphi'(\theta) = \text{Cov}^{-1}[T(Y)]$$  \hspace{1cm} (2.35)

giving

$$\hat{\theta}(Y)_{\parallel S_{Le}} = T(Y) - E[T(Y)] + \theta$$

from which we see that a second necessary condition is that

$$E[T(Y)] = \theta \hspace{1cm} (2.36)$$
2.3. UNBIASED PARAMETER ESTIMATION

However if (2.36) holds, differentiating the first equation in (2.32) gives

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

which is (2.35), and hence (2.36) is sufficient (and necessary) for (2.33) to be independent of \( \theta \). Under (2.36), the FIM thus simplifies to

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

We have thus established that for the exponential family (2.29) satisfying (2.36), the estimator

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

(2.37)
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 (2.29) and (2.36) 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 (2.30) where \( \eta = \varphi(\theta) \). Then there exists an efficient estimator for \( \theta \) if

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

2.3.3 The Maximum Likelihood Estimator

For an exponential family (2.29) satisfying (2.35) and (2.36), the score function is given by

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

and hence the efficient estimator (2.37) 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,

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

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

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

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.
2.3.4 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 lose 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.

**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

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

which does not depend on $\theta$. \qed

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.
2.3. UNBIASED PARAMETER ESTIMATION

Finite dimensional data compression

Example 2.3.1. Suppose that $Y_1, \ldots, Y_N$ are iid $N(m, \sigma^2)$. Then the joint pdf is given by

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

from which we see that

$$T(Y) = \left[ \frac{\sum_{k=1}^{N} y_k}{\sum_{k=1}^{N} y_k^2} \right]$$

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 [5]). 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 (2.29) 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 (2.29) with $s \leq r$.

Example 2.3.2. 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 (2.29). 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.
2.3.5 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 $p$ does not depend on $\theta$. We realized that this projection will reduce the MSE also if the estimator is unbiased and has 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.

**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 $p$. \hfill $\square$

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
Thus conditioning on a sufficient statistic strictly reduces the risk when

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

(2.38)

For $\hat{\theta}_{\text{UMVU}}$ to be UMVU (2.38) 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 $E_{Y \mid T(Y)} [\Delta(Y) | T(Y)]$ is a function of $T$ and satisfies

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

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

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

$$\langle \Delta(Y), \mathbb{E} [\hat{\theta}(Y) | T(Y)] \rangle = \mathbb{E} [\Delta(Y) \mathbb{E} [\hat{\theta}(Y) | T(Y)]]$$

$$= \mathbb{E} [\mathbb{E} [\Delta(Y) \mathbb{E} [\hat{\theta}(Y) | T(Y)]] | T(Y)]$$

$$= \mathbb{E} [\mathbb{E} [\hat{\theta}(Y) | T(Y)] \mathbb{E} [\Delta(Y) | T(Y)]]$$

$$= \mathbb{E} [\mathbb{E} [\hat{\theta}(Y) | T(Y)] 0] = 0$$

Thus $E_{Y \mid T(Y)} [\hat{\theta}(Y) | T(Y)]$ 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.

**Generalizing the Rao-Blackwell theorem**

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

$$\Phi (\mathbb{E} [Y]) \leq \mathbb{E} [\Phi(Y)]$$

(2.39)

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 (2.39) 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

$$\Phi (\mathbb{E} [X]) \leq \mathbb{E} [\Phi(X)]$$

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

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

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.
2.3.6 Efficient Estimators, Sufficient Statistics and Maximum Likelihood Estimators

It is known that for an exponential family (2.29), \( T(Y) \) 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 2.3.2 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 (2.35) and (2.36). 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 2.3.3 that an efficient estimator is a ML estimator.

\[
\text{Efficient} \Rightarrow \text{Sufficient statistic} \Rightarrow \text{ML estimator} \quad (2.40)
\]

2.3.7 Parameter Transformations

Maximum Likelihood Estimators

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 \( q : \mathbb{R}^n \rightarrow \mathbb{R}^m \) 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

\[
\hat{\alpha}_{ML} = q(\hat{\theta}_{ML}) \quad (2.41)
\]

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 (2.41) 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 \( D_\alpha \).

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 (2.29), with \( \theta \) replaced by \( \alpha \) and with

\[
\varphi(\alpha) = \Sigma^{-1} \alpha, \quad B(\alpha) = \frac{1}{2} \alpha^T \Sigma^{-1} \alpha
\]
2.3. UNBIASED PARAMETER ESTIMATION

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

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

(2.42)

Using that \( \varphi^{-1}(\eta) = \Sigma \eta \) we can express this as (2.30) with

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

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

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

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

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

where \( \Phi \) has full column rank. We then have

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

Introducing

\[
\hat{\theta}(x) = (\Phi^T \Sigma^{-1} \Phi)^{-1} \Phi^T \Sigma^{-1} x \in \mathbb{R}^m
\]

(2.43)

we can write this as

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

This pdf has the same form as (2.42) and hence \( \hat{\theta}(\hat{\alpha}(Y)) \) is a complete sufficient statistic and

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

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

We can also obtain the mean explicitly from

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

In conclusion,

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

(2.44)

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.
2.3.8 Best Linear Unbiased Estimators (BLUE)

Let \( Y \in \mathbb{R}^N \) be a random vector with mean and covariance
\[
E[Y] = \Phi \theta, \quad \theta \in \mathbb{R}^n, \quad n \leq N, \quad \text{Cov}[Y] = \Sigma > 0
\]  

(2.45)

The class of linear unbiased estimators of \( \theta \) is given by
\[
\{LY : L \in \mathbb{R}^{n \times N}, L\Phi = I\}
\]  

(2.46)

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 (2.46), and not all estimators. If we can find a “score” function which has zero mean and the identity as covariance with all estimators in (2.46), 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 \( Y \):
\[
S(Y) = \bar{S}Y + m, \quad \bar{S} \in \mathbb{R}^{n \times N}
\]

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

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

Taking \( \bar{S} = \Phi^T\Sigma^{-1} \) gives
\[
\langle \hat{\theta}, S(Y)^T \rangle = I
\]

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

(2.47)

giving the projection on the subspace generated by \( S(Y) \) and a constant (2.23) as
\[
\hat{\theta}_{||S_{le}} = (\Phi^T\Sigma^{-1}\Phi)^{-1}\Phi\Sigma^{-1}(Y - E[Y]) + \theta
\]
\[
= (\Phi^T\Sigma^{-1}\Phi)^{-1}\Phi\Sigma^{-1}(Y - \Phi\theta) + \theta
\]
\[
= (\Phi^T\Sigma^{-1}\Phi)^{-1}\Phi\Sigma^{-1}Y
\]  

(2.48)

from which we see that this is a bona fide estimator, and hence the BLUE. The quantity (2.47) 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 (2.45) is given by

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

(2.49)

**Remark 2.3.2.** Returning to the Section 2.3.7 and the paragraph on efficient estimators, we see that \( \hat{\theta}(\hat{\alpha}(Y)) \), defined in (2.44), 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 (2.45), 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

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

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

(2.50)

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

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

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

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

(2.51)

This estimate is unbiased regardless of \( W \) and belongs therefore to the class (2.46). 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 [7]. In particular, knowing the solution, one can use Schur complement (Section 2.A). We have

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

(2.52)
Since $\Phi$ is assumed to be full column rank and $\Sigma > 0$, $\Phi^T \Sigma^{-1} \Phi > 0$ and hence Schur complement gives that (2.52) 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 (2.49).

2.3.9 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 $m$ and variance $\sigma^2$ and let $\theta = [m \sigma^2]^T$.

(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 2.3.2 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.
2.4. **USING ESTIMATION FOR PARAMETER ESTIMATION**

(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

\[
\frac{1}{2} N(0, 1) + \frac{1}{2} N\left(\theta, \left(e^{-1/\theta^2}\right)\right)
\]

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:

2.4 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 \( \hat{Y}^N = \mathbb{E}[Y^N] \) given by

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

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

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

with equality only if \( \hat{Y}^N(\theta) = \mathbb{E}[Y^N] \), 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 \( \frac{1}{N} \mathbb{E}\|Y^N - \hat{Y}(\theta)\|^2 \):

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

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 (2.50) 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 $Y_1^N \in \mathbb{R}^M$ and $Y_2^N \in \mathbb{R}^{N-M}$ and then consider the OLE of $Y_1^N$ given $Y_2^N \in \mathbb{R}^{N-M}$ and then pursue the approach above. Here a myraid 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.

2.5 References

Section 2.3.1

Chapter 1, Sections 2.1,2.5-2.6 in [5].

Section 2.3.8: Section 4.3 in [7] (Chapter 6 in [3]).
2.A. Schur Complement

Consider the symmetric block-matrix

\[ X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \]

If \( A > 0 \) then

\[ X > 0 \iff C - B^T A^{-1} B > 0 \]
\[ X \geq 0 \iff C - B^T A^{-1} B \geq 0 \]

If \( C > 0 \) then

\[ X > 0 \iff A - B C^{-1} B^T > 0 \]
\[ X \geq 0 \iff A - B C^{-1} B^T \geq 0 \]
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 $N(\theta, 1)$ 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

$$\text{MSE} \left[ \hat{\theta}_N(\lambda_0) \right] \approx \frac{1}{N^2} \frac{1}{\theta^4} + \frac{1}{N} \left(1 - \frac{1}{N\theta^2}\right)^2 \approx \frac{1}{N^2} \frac{1}{\theta^4} + \frac{1}{N} - \frac{2}{N^2 \theta^2} \approx \frac{1}{N}$$

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.A 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

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

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

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

This is the same as

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

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

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

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
\[ \int_{\Omega} \| \hat{\theta}_N(\omega) - \theta \|^2 dP(d\omega) \to 0, \quad \text{as } N \to \infty. \]

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
\[
0 \leftarrow \int_A \| \hat{\theta}_N(\omega) - \theta \|^2 dP(d\omega) + \int_{A^c} \| \hat{\theta}_N(\omega) - \theta \|^2 dP(d\omega) \geq \varepsilon \int_{A^c} dP(d\omega) = \varepsilon P(\| \hat{\theta}_N - \theta \|^2 > \varepsilon)
\]

Thus we see that for any \( \varepsilon > 0 \) it holds that
\[ P(\| \hat{\theta}_N - \theta \|^2 \geq \varepsilon) \to 0, \quad \text{as } N \to \infty. \]

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 \( (\Omega, \mathcal{F}, P) \) 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
\[ \hat{\theta}_N(\omega) \to \theta, \quad \text{as } N \to \infty \] (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 [1]). 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 $E[|Y_k|^{1+\delta}]$ is bounded for some $0 < \delta \leq 1$. The strong law of large numbers can also hold for dependent sequences.

**Lemma 3.2.2** ([6]). 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 $N > M > 0$

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

Then for any $\delta > 0$

$$\frac{1}{N} \sum_{k=1}^{N} Y_k \leq CN^{\frac{\delta}{2}\sigma-1} \log^{2+\delta} N \quad \text{w.p.} 1$$

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

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

Whenever $X_N$ and $X$ have distribution functions $F_N(x) = P(\{\omega : X_N(\omega) < x\})$ 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

$$X_N \xrightarrow{\text{dist}} X, \quad \text{and} \quad X_N \xrightarrow{\text{dist}} F$$

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 $P$

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

### 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 $P$ be the Lebesgue 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 \(A_N^c\) it holds that
\[
\|\hat{\theta}_N - \theta\| = N
\]
Then clearly for every \(\omega, \hat{\theta}_N \to \theta\) and convergence w.p.1 holds. However
\[
\mathbb{E}\left[\|\hat{\theta}_N - \theta\|^2\right] \geq N^2 \mathbb{P}(A_N^c) = N^2 \frac{1}{N} \to \infty \text{ as } N \to \infty
\]
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 \(X_N \to 0\) in probability and that there exists a random variable \(Y, \mathbb{E}[|Y|^2] < \infty\), such that \(|X_N| \leq Y\) a.e. Then \(X_N \to 0\) 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 \(\alpha_N \to \infty\) the rotation never stops which means that any given point \(\omega\) on the circle will belong to \(A_N^c\) 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
\[
X_N \xrightarrow{\text{dist}} X, \ Y_N \xrightarrow{\text{dist}} y, \ Z_N \xrightarrow{\text{dist}} z
\]
where \(y\) and \(z\) are constants. Then
\[
X_N Y_N + Z_N \xrightarrow{\text{dist}} Xy + z
\]
Furthermore, \(Z_N \xrightarrow{p} z\).

**Corollary 3.2.1.** Suppose that \(Y_N \xrightarrow{p} Y\) and \(y_N \to y\). Then
\[
\mathbb{P}(Y_N \leq y_N) \to F_Y(y)
\]
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
\[
\sqrt{N}(X_N - \theta) \xrightarrow{\text{dist}} N(0, P)
\]
Let \(q(x)\) be differentiable with \(q'(\theta) \neq 0\). Then
\[
\sqrt{N}(q(X_N) - q(\theta)) \xrightarrow{\text{dist}} N\left(0, q'(\theta)P\left(q'(\theta)\right)^T\right)
\]
Lemma 3.2.5 (Convergence of moments). Suppose that
\[ X_N \xrightarrow{dist} X \]
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 \( r, 0 < r < p \)
\[ \lim_{N \to \infty} \mathbb{E} [ X_N^p ] = \mathbb{E} [ X^p ] \]

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
\[ R_N := \frac{1}{N} \sum_{k=1}^{N} \varphi_k \varphi_k^T \to R > 0 \] (3.2)
and where \( \{E_k\} \) is iid with zero mean and variance \( \lambda < \infty \). The least squares estimator is given by

\[
\hat{\theta}_N = R_N^{-1} \frac{1}{N} \sum_{k=1}^{N} \varphi_k Y_k = \theta_o + R_N^{-1} \frac{1}{N} \sum_{k=1}^{N} \varphi_k E_k \tag{3.3}
\]

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 [1]) 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 [1]) 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 \( \mathbb{E} \left[ |E_k|^{2+\delta} \right] \leq M < \infty \) 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 2.4 that estimators can be solutions to optimization problems
\[ \hat{\theta}_N = \arg \min_{\theta \in D}\theta V_N(\theta) \]  
(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
\[ \sup_{\theta \in D_0} |V_N(\theta) - V(\theta)| \xrightarrow{w.p.1} 0 \]
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
\[ \hat{\theta}_N \xrightarrow{w.p.1} D_c = \{\theta : V(\theta) \leq V(\xi) \forall \xi \in D_0\} \]
(3.6)

Under the conditions of Theorem 3.3.1, consistency hinges on if \( D_c \) is a singleton containing the true parameter \( \theta_o \). Typically, the estimator is constructed such that the limit \( V(\theta) \) is minimized by the true parameter \( \theta_o \). 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 \( Y_k^2 \) 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 \( \theta^* \) is a solution to (3.4) then so is \( -\theta^* \) 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
\[ V_N(\theta) = \frac{1}{N} \sum_{k=1}^{N} |Y_k - \theta|^2 \]
(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 \( \theta_o \) 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

\[
\hat{\theta}_N - \theta_* = - \left( V''(\xi_N) \right)^{-1} V'_N(\theta_*)
\]

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

\[
\hat{\theta}_N - \theta_* \approx - \left( V''(\theta_*) \right)^{-1} V'_N(\theta_*)
\] (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

\[
V'_N(\theta) = \frac{2}{N} \sum_{k=1}^{N} (Y_k - \theta)\theta
\]

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

\[
\sqrt{N}V'_N(\theta_o) = 2\theta \frac{1}{\sqrt{N}} \sum_{k=1}^{N} (Y_k - \theta)
\] (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

\[
\sqrt{N}V'_N(\theta_o) \overset{\text{dist}}{\to} N(0, 4\theta_o^6)
\]

As in the example, often

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

**Lemma 3.3.1** (Lemma 6.1.14 in [5]). Suppose that \( Y_k \xrightarrow{\text{dist}} Y \) 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 \( H \),
\[
\sqrt{N}(\hat{\theta}_N - \theta_o) \xrightarrow{\text{dist}} H
\]
(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
  \[
  \hat{\theta}_N \xrightarrow{P} \theta_o
  \]
  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 \( \theta_o \). However, if \( \sup_{N} N^{\frac{1}{2}(1+\delta)} \mathbb{E}[|\hat{\theta}_N - \theta_o|^{1+\delta}] < \infty \) for some \( \delta > 0 \), Lemma 3.2.5 gives the stronger

  \[
  \lim_{N \to \infty} \sqrt{N} \mathbb{E}[|\hat{\theta}_N - \theta_o|] = 0
  \]
  As we noted earlier, for an estimator efficient in the limit, \( \lim_{N \to \infty} \mathbb{E}[\hat{\theta}_N] = \theta_o \).

- The limit distribution \( H \) does not necessarily have to have zero mean, see, e.g., Example 6.2.3 in [5].

- 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 \( \Sigma_H \),

\[
\Sigma_H \leq \liminf_N \text{NMSE}[\hat{\theta}_N]
\]
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 [5]). Let $Y_1, Y_2, \ldots$ be iid with pdf $p(y; \theta)$, $\theta \in D_0 \subset \mathbb{R}$ 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 $\theta_o \in D_0$, 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

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

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

$$\sqrt{N}(\hat{\lambda}_N - g(\theta)) \mathop{\xrightarrow{dist}} N(0, \Lambda(\theta))$$

then

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

(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 Lebesgue 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 [5], 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
\[
\sup \text{NMSE} \left[ \hat{\theta}_N \right] \to \infty
\]
In particular, for each \( N \) setting the true \( \theta \) to \( \theta_N := N^{-1/4} \) will result in that \( \text{NMSE} \left[ \hat{\theta}_N \right] \to \infty \) as \( N \to \infty \).

3.3.5 The Maximum Likelihood Estimator

As noted the ML estimator is efficient only for the exponential family (2.29) satisfying (2.35) and (2.36). 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)
\[
V_N(\theta) = \frac{1}{N} S(Y^N; \theta) := -\frac{1}{N} \sum_{k=1}^{N} \log p(Y_k; \theta) 
\]
(3.16)
Since \( \{Y_k\} \) are independent and assuming that
\[
\mathbb{E} [\log p(Y_1; \theta)] < \infty, \quad \mathbb{E} [\log p(Y_1; \theta)] < \infty 
\]
(3.17)
we can use the strong law of large numbers (Theorem 5.4.2 in [1]) and obtain
\[
V_N(\theta) \xrightarrow{w.p.1} V(\theta) := -\mathbb{E} [\log p(Y_1; \theta)] 
\]
(3.18)
which is known as the Shannon entropy for $Y_1$ when $\theta = \theta_o$. Let us introduce

$$D_N(\theta|\theta_o) = V_N^{ML}(\theta) - V_N^{ML}(\theta_o) = -\frac{1}{N} \sum_{k=1}^{N} \log \left( \frac{p(Y_k; \theta)}{p(Y_k; \theta_o)} \right)$$  \hspace{1cm} (3.19)

$$D_N(\theta|\theta_o) \xrightarrow{w.p.1} D(\theta|\theta_o) := -\mathbb{E} \left[ \log \left( \frac{p(Y_1; \theta)}{p(Y_1; \theta_o)} \right) \right]$$  \hspace{1cm} (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 that the maximum-likelihood estimate $\hat{\theta}_{ML,N} \xrightarrow{w.p.1} D_c$. There are a number of counter-examples such as Example 6.4.1 in [5]. 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 [5]. However, 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

$$V_N'(\theta) = \frac{1}{N} S(Y^N; \theta) = \frac{1}{N} \sum_{k=1}^{N} S(Y_k; \theta)$$  \hspace{1cm} (3.21)

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

$$V_N''(\theta) = \frac{1}{N} \frac{\partial}{\partial \theta} S(Y^N; \theta) = \frac{1}{N} \sum_{k=1}^{N} \frac{\partial}{\partial \theta} S(Y_k; \theta) \xrightarrow{w.p.1} V(\theta) := \mathbb{E} \left[ \frac{\partial}{\partial \theta} S(Y_k; \theta) \right]$$  \hspace{1cm} (3.22)
but according to (2.22) 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 that when the ML estimator is consistent it is typically also asymptotically efficient. In fact,

\[
NE [V'_N(\theta_o)(V'_N(\theta_o))^T] = I_F(\theta)
\]

so \(\tilde{P}(\theta_o)\) in (3.10) is given by

\[
\tilde{P}(\theta_o) = I_F(\theta_o)
\]

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 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 \(I(\theta_o)\) 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 \(\tilde{\theta}_N \approx \theta_o\) for \(N\) sufficiently large and \(S(Y^N; \tilde{\theta}_N) \approx S(Y^N; \theta_o)\) as well as \(I(\tilde{\theta}_N) \approx I(\theta_o)\) leading to the estimator

\[
\tilde{\theta}_N = (NI_F(\tilde{\theta}_N))^{-1} S(Y^N; \tilde{\theta}_N) + \tilde{\theta}_N
\]

Alternatively making use of \(\frac{\partial}{\partial \theta} S(Y^N; \tilde{\theta}_N) \approx -NI(\theta_o)\) leads to the estimator

\[
\tilde{\theta}_N = -\left(\frac{\partial}{\partial \theta} S(Y^N; \tilde{\theta}_N)\right)^{-1}(\theta_o)S(Y^N; \tilde{\theta}_N) + \tilde{\theta}_N
\]

However, it is not sufficient that \(\tilde{\theta}_N\) 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

\[
P(|X_N| \geq C(\varepsilon)) < \varepsilon, \quad N \geq N(\varepsilon) < \infty
\]
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 $\bar{\theta}_N$ is $\sqrt{N}$-consistent, that is that $\bar{\theta}_N - \theta_o$ is $O_p(1/\sqrt{N})$. For example any estimator for which

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

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 $\bar{\theta}_N$. We will return to how to numerically compute estimates later.

### 3.4 Parameter Transformations

#### 3.4.1 Consistent Estimators

In Section 2.3.7 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 $\bar{\theta}_N$ is a consistent estimator of $\theta$. Then, $q(\bar{\theta})$ is a consistent estimator of $\alpha$. Furthermore, suppose that

$$
\sqrt{N}(\bar{\theta}_N - \theta) \overset{dist}{\to} N(0, P)
$$

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

$$
\sqrt{N}(q(\bar{\theta}_N) - \alpha) \overset{dist}{\to} N(0, q'(\theta)P(q'(\theta))^T)
$$

(3.28)

In particular, if $\hat{\theta}_N$ is asymptotically efficient, then it follows from (3.28) and Corollary 3.3.1 that $q(\hat{\theta}_N)$ is asymptotically efficient for $\alpha$. 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 2.3.7 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 $D_\alpha \in \mathbb{R}^m$, $m \geq n$, but that we now have an estimate $\hat{\alpha}_N$ of $\alpha \in D_\alpha \supset D_\alpha$. Suppose that $\hat{\alpha}_N$ is consistent and, denoting the true parameter by $\alpha_o$, has asymptotic distribution

$$
\sqrt{N}(\hat{\alpha}_N - \alpha_o) \overset{dist}{\to} N(0, \Sigma(\alpha_o))
$$

(3.29)

In Remark 2.3.2 we highlighted that when $q$ was a linear mapping, 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

$$
\hat{\theta}_N = \arg \min_{\theta \in \Theta} V_N(\theta), \text{ where } V_N(\theta) = \frac{1}{N}(\hat{\alpha}_N - q(\theta))^T \Sigma_N^{-1}(\hat{\alpha}_N - q(\theta))
$$

(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 $N\Sigma_N \to \Sigma(\alpha_o)$, 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

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

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 (2.49) 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

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

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

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

(3.32)

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:

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

(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

$$\langle X, X^T \rangle - \langle X, Y \rangle \langle Y, Y^T \rangle \langle Y, X^T \rangle$$

(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
3.4. PARAMETER TRANSFORMATIONS

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 $\tilde{V}_N(\alpha)$

$$\hat{\alpha}_N = \arg \min_{\alpha \in \tilde{D}_\alpha} \tilde{V}_N(\alpha) \quad (3.35)$$

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

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

where

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

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

$$\tilde{\Sigma}_\theta := \left( Q^T \tilde{V}'_N(\alpha_o) \tilde{P}(\alpha_o) \tilde{V}''_N(\alpha_o) \right)^{-1} \quad (3.38)$$

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

$$\bar{\Sigma}_\theta := (\bar{V}'_N(\theta))^{-1} \bar{P}(\theta_o)(\bar{V}''_N(\theta))^{-1} \quad (3.39)$$

where

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

Now

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

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

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

and hence

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

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

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

$$= Q^T \tilde{V}'' \left( \tilde{P}^{-1} - Q \left( Q^T \tilde{P} Q \right)^{-1} Q^T \right) \tilde{V}' Q$$
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.A): The difference is positive semidefinite iff

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

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 $J_N(\theta) := V_N(q(\theta))$ gives worse accuracy asymptotically than going through the detour of first estimating $\alpha$ and then $\theta$ through (3.30). Only in the case that $V''(\alpha_0) = \tilde{P}(\alpha_0)$ 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 $\tilde{P}(\alpha_0)$, defined in (3.37), and $V''(\theta_0)$ 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 hence the direct approach yields the Fisher Information Matrix as asymptotic covariance matrix. However, when $\tilde{V}_N$ is not the ML-criterion (or an asymptotic equivalent), then $\tilde{P}(\theta_0)$ and $V''(\theta_0)$ 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 $D_\alpha \subset \hat{D}_\alpha$. 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

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

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

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** [8].

### 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., [8]. However, computationally it may be even more advantageous to let $\hat{D}_\alpha$ outside $D_\alpha$. We illustrate with a simple example.
Example 3.4.1. Let the model be

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

(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

\[ Y^N = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = m_N(\theta) + E, \quad \text{where} \]

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

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

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

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)

\[ D_\alpha = \left\{ g := \{g_k\}_{k=0}^\infty : \ g_k = \theta^k, \ k = 0, 1, \ldots, \ \theta \in D_\theta \right\} \]

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\(^1\) 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

\[ \hat{\alpha}_{N,n} = \left( \Phi_{N,n}^T \Phi_{N,n} \right)^{-1} \Phi_{N,n}^T Y^N \]

(3.44)

\(^1\)We will call such models Finite Impulse Response (FIR) models.
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:

\[
\begin{bmatrix}
1 \\
-g_0 \\
-g_1 \\
\vdots
\end{bmatrix} \theta = \begin{bmatrix}
g_0 \\
g_1 \\
g_2 \\
\vdots
\end{bmatrix}
\]

(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

\[
\begin{bmatrix}
1 \\
-\hat{\alpha}_0 \\
-\hat{\alpha}_1 \\
\vdots \\
-\hat{\alpha}_{n-1}
\end{bmatrix} \theta = \begin{bmatrix}
\hat{\alpha}_0 \\
\hat{\alpha}_1 \\
\vdots \\
\hat{\alpha}_{n-2}
\end{bmatrix},
\]

(3.46)

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).
3.5. USING ESTIMATION FOR PARAMETER ESTIMATION

While the two step procedure in Example 3.4.1 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

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

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 \( D_\alpha(N) := D_{\alpha,n(N)} \) 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 [2].

3.5 Using Estimation for Parameter Estimation

In Section 2.4 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

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

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 \( V_N(\theta, I_N) \stackrel{w.p.1}{\to} V(\theta) \) 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 3.3.2, 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

\[
W_N = \Sigma^{-1}_N(\theta) := \text{Cov}^{-1} Y
\]
belongs to this class.

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

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

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

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

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

where

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

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

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

We thus will assume that the pdfs of $Y$ and $\{W_N\}$ are such that these limits exists. 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 ?? we may suspect that $W_N = \Sigma_N^{-1}$ will be optimal. This choice gives

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

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

$$\Phi_N^T \Sigma_N^{-1} \Phi_N - \Phi_N^T W_N \Phi_N \left( \Phi_N^T W_N \Sigma_N W_N \Phi_N \right)^{-1} \Phi_N^T W_N \Phi_N$$

This difference is positive definite iff

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

$$\Leftrightarrow$$

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

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.

### 3.5.1 Exercises

3.5.1. Let $Y_1, \ldots, Y_N$ be iid $N(\theta, a\theta^2)$ 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 $a\theta^2$ (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 [5], Sections 4.1–4.2, 4.5, 5.1*-5.4* in [1].

Section 3.3
Chapter 6 in [5].
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 (2.27) gives

$$\text{MSE} \left[ \hat{\theta} \right] = b^2(\theta) + \text{Cov} \left[ \hat{\theta} \right] \geq b^2(\theta) + \frac{(1 + b'(\theta))^2}{NI_F(\theta)} \quad (3.48)$$

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

$$NI_F(\theta)b^2(\theta) + 2b'(\theta) + (b'(\theta))^2 \leq 0$$

or

$$b'(\theta) \leq -\frac{NI_F(\theta)}{2}b^2(\theta) - \frac{1}{2}(b'(\theta))^2 \quad (3.50)$$

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

\[
\frac{l/2}{\sqrt{N\gamma}} = \frac{1}{\sqrt{N\gamma}}
\]

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 \( \mathcal{O}(1/N) \) 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

\[
\tilde{b}(\theta) = \frac{b(a)}{1 + \frac{N\gamma}{2} b(a)(\theta - a)}
\]

when \( b(a) > 0 \), and

\[
\tilde{b}(\theta) = \frac{b(c)}{1 + \frac{N\gamma}{2} b(c)(\theta - a)}
\]

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

\[
\theta - a = \frac{(N\gamma)^{\frac{1}{2} + \frac{\delta}{2}} b(a) - 1}{\frac{N\gamma}{2} b(a)} \leq 2(N\gamma)^{-\frac{1}{2} + \frac{\delta}{2}}
\]

Thus if \( 0 < \delta < 1/2 \) and \( b(a) > 0 \),

\[
\tilde{b}^2(\theta) \leq \frac{1}{(N\gamma)^{1+\delta}} \quad \forall \theta \geq a + a_N, \quad 0 \leq a_N \leq 2(N\gamma)^{-\frac{1}{2} + \frac{\delta}{2}} \to 0, \quad \text{as} \quad N \to \infty
\]
Similarly we can show that if $0 < \delta < 1/2$ and $b(b) < 0$,
\[
\tilde{b}^2(\theta) \leq \frac{1}{(N\gamma)^{1+\delta}} \quad \forall \theta \leq c - c_N, \quad 0 \leq c_N \leq 2(N\gamma)^{-\frac{1}{2} + \frac{\delta}{2}} \to 0, \quad \text{as } N \to \infty
\]

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 (2.27) 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
\[
b(\theta) = C_1 e^{-\sqrt{N}\theta} + C_2 e^{\sqrt{N}\theta}
\]
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

\[
b(\theta) = \frac{1}{e^{\sqrt{N}(b-a)} - e^{-\sqrt{N}(b-a)}}
\begin{bmatrix}
  e^{\sqrt{N}\theta} & e^{\sqrt{N}\theta} \\
  -e^{-\sqrt{N}\theta} & e^{-\sqrt{N}\theta}
\end{bmatrix}
\begin{bmatrix}
  e^{\sqrt{N}b} & -e^{\sqrt{N}a} \\
  -e^{-\sqrt{N}b} & e^{-\sqrt{N}a}
\end{bmatrix}
\begin{bmatrix}
  b(a) \\
  b(b)
\end{bmatrix}
\]

\[
= \frac{1}{e^{\sqrt{N}(b-a)} - e^{-\sqrt{N}(b-a)}}
\begin{bmatrix}
  e^{\sqrt{N}(b-a)} & e^{\sqrt{N}(b-a)} \\
  -e^{-\sqrt{N}(b-a)} & e^{-\sqrt{N}(b-a)}
\end{bmatrix}
\begin{bmatrix}
  e^{\sqrt{N}(\theta-a)} & e^{\sqrt{N}(\theta-a)} \\
  -e^{-\sqrt{N}(\theta-a)} & e^{-\sqrt{N}(\theta-a)}
\end{bmatrix}
\begin{bmatrix}
  b(a) \\
  b(b)
\end{bmatrix}
\]

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

Minimum Mean Square Error Parameter Estimation

4.1 The Bias-Variance Error Trade-Off
4.2 Risk Estimation Methods
4.3 Probabilistic Models and the Bayes Estimator
4.4 Linear in the Parameters Models
Chapter 5

Linear in the Parameters Models
Chapter 6

Dynamical Models

6.1 Model Structures and Probabilistic Models

6.2 Estimation Methods

6.2.1 Maximum Likelihood Estimation

6.2.2 The Extended Invariance Principle

6.2.3 Multi-Step Least-Squares Methods

6.2.4 The Prediction Error Method

6.2.5 Instrumental Variable Methods

6.2.6 Indirect Inference

6.3 Linear Models

6.3.1 Maximum Likelihood Estimation

Multi-Step Least-Squares Methods

6.3.2 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

6.4 Nonlinear Models

6.5 Probabilistic Models
Chapter 7

Model Quality

7.1 Variance Quantification

7.1.1 Fundamental Geometric Principles
7.1.2 Fundamental Structural Results
7.1.3 Variability of Estimated Frequency Response
   A General Principle
   Details for FIR Models
   Details for ARMAX Models
   Details OE and BJ Models
7.1.4 Variability of Nonlinear System Estimates
7.1.5 Bootstrap Methods
Chapter 8

Experiment Design

8.1 Identifiability
8.2 Persistence of Excitation
8.3 Input Signal Design
8.3.1 Common Input Signals
PRBS
Sums of Sine-Waves and Crest Factor Correction
8.4 Application Oriented Experiment Design
8.5 Adaptive Experiment Design
Chapter 9

Model Validation

9.1 Residual whiteness Tests
9.2 Input to residual correlation tests
9.3 Model Error Modelling
Chapter 10

Applications

10.1 Closed Loop Identification
10.2 Network Models
10.3 Errors-in-Variables Models
10.4 Block-structured Nonlinear Models
Bibliography


