STATISTICAL METHODS FOR SIGNAL PROCESSING

Alfred O. Hero

December 22, 2014

This set of notes is the primary source material for the course EECS564 “Estimation, filtering and detection” used over the period 1999-2014 at the University of Michigan Ann Arbor. The author can be reached at
Dept. EECS, University of Michigan, Ann Arbor, MI 48109-2122
Tel: 734-763-0564.
email hero@eecs.umich.edu;
http://www.eecs.umich.edu/~hero/.
## Contents

1 INTRODUCTION ........................................... 9  
1.1 STATISTICAL SIGNAL PROCESSING ....................... 9  
1.2 PERSPECTIVE ADOPTED IN THIS BOOK ............... 9  
1.2.1 PREREQUISITES ................................. 11  

2 NOTATION, MATRIX ALGEBRA, SIGNALS AND SYSTEMS .. 12  
2.1 NOTATION ........................................... 12  
2.2 VECTOR AND MATRIX BACKGROUND ................. 12  
2.2.1 ROW AND COLUMN VECTORS .................... 12  
2.2.2 VECTOR/VECTOR MULTIPLICATION ............... 13  
2.3 ORTHOGONAL VECTORS ................................ 13  
2.3.1 VECTOR/MATRIX MULTIPLICATION .............. 14  
2.3.2 THE LINEAR SPAN OF A SET OF VECTORS ....... 14  
2.3.3 RANK OF A MATRIX ............................. 14  
2.3.4 MATRIX INVERSION ............................. 14  
2.3.5 ORTHOGONAL AND UNITARY MATRICES ......... 15  
2.3.6 GRAMM-SCHMIDT ORTHOGONALIZATION AND ORTHONORMALIZATION ......... 15  
2.3.7 EIGENVALUES OF A SYMMETRIC MATRIX ...... 16  
2.3.8 MATRIX DIAGONALIZATION AND EIGENDECOMPOSITION ......... 16  
2.3.9 QUADRATIC FORMS AND NON-NEGATIVE DEFINITE MATRICES ......... 17  
2.4 POSITIVE DEFINITENESS OF SYMMETRIC PARTITIONED MATRICES ......... 17  
2.4.1 DETERMINANT OF A MATRIX ................... 18  
2.4.2 TRACE OF A MATRIX ............................ 18  
2.4.3 VECTOR DIFFERENTIATION ........ ............. 18  
2.5 SIGNALS AND SYSTEMS BACKGROUND ............... 19  
2.5.1 GEOMETRIC SERIES ............................ 19  
2.5.2 LAPLACE AND FOURIER TRANSFORMS OF FUNCTIONS OF A CONTINUOUS VARIABLE ............... 19  
2.5.3 Z-TRANSFORM AND DISCRETE-TIME FOURIER TRANSFORM (DTFT) ......... 19  
2.5.4 CONVOLUTION: CONTINUOUS TIME ............... 20  
2.5.5 CONVOLUTION: DISCRETE TIME ................. 20  
2.5.6 CORRELATION: DISCRETE TIME ................. 21  
2.5.7 RELATION BETWEEN CORRELATION AND CONVOLUTION ...... 21  
2.5.8 CONVOLUTION AS A MATRIX OPERATION ....... 21  
2.6 BACKGROUND REFERENCES ........................... 21  
2.7 EXERCISES ........................................... 22
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.1 SCALAR ESTIMATION CRITERIA FOR NON-RANDOM PARAMETERS</td>
<td>67</td>
</tr>
<tr>
<td>4.4.2 METHOD OF MOMENTS (MOM) SCALAR ESTIMATORS</td>
<td>70</td>
</tr>
<tr>
<td>4.4.3 MAXIMUM LIKELIHOOD (ML) SCALAR ESTIMATORS</td>
<td>74</td>
</tr>
<tr>
<td>4.4.4 SCALAR CRAMÉR-RAO BOUND (CRB) ON ESTIMATOR VARIANCE</td>
<td>77</td>
</tr>
<tr>
<td>4.5 ESTIMATION OF MULTIPLE NON-RANDOM PARAMETERS</td>
<td>84</td>
</tr>
<tr>
<td>4.5.1 MATRIX CRAMÉR-RAO BOUND (CRB) ON COVARIANCE MATRIX</td>
<td>85</td>
</tr>
<tr>
<td>4.5.2 METHODS OF MOMENTS (MOM) VECTOR ESTIMATION</td>
<td>88</td>
</tr>
<tr>
<td>4.5.3 MAXIMUM LIKELIHOOD (ML) VECTOR ESTIMATION</td>
<td>89</td>
</tr>
<tr>
<td>4.6 HANDLING NUISANCE PARAMETERS</td>
<td>96</td>
</tr>
<tr>
<td>4.7 BACKGROUND REFERENCES</td>
<td>99</td>
</tr>
<tr>
<td>4.8 EXERCISES</td>
<td>99</td>
</tr>
</tbody>
</table>

## 5 LINEAR ESTIMATION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 MIN MSE CONSTANT, LINEAR, AND AFFINE ESTIMATION</td>
<td>110</td>
</tr>
<tr>
<td>5.1.1 BEST CONSTANT ESTIMATOR OF A SCALAR RANDOM PARAMETER</td>
<td>111</td>
</tr>
<tr>
<td>5.2 BEST LINEAR ESTIMATOR OF A SCALAR RANDOM PARAMETER</td>
<td>111</td>
</tr>
<tr>
<td>5.3 BEST AFFINE ESTIMATOR OF A SCALAR R.V. θ</td>
<td>112</td>
</tr>
<tr>
<td>5.3.1 SUPERPOSITION PROPERTY OF LINEAR/AFFINE ESTIMATORS</td>
<td>114</td>
</tr>
<tr>
<td>5.4 GEOMETRIC INTERPRETATION: ORTHOGONALITY CONDITION AND PROJECTION THEOREM</td>
<td>114</td>
</tr>
<tr>
<td>5.4.1 LINEAR MINIMUM MSE ESTIMATION REVISITED</td>
<td>114</td>
</tr>
<tr>
<td>5.4.2 AFFINE MINIMUM MSE ESTIMATION</td>
<td>116</td>
</tr>
<tr>
<td>5.4.3 LMMSE ESTIMATOR IS MMSE ESTIMATOR FOR GAUSSIAN MODEL</td>
<td>118</td>
</tr>
<tr>
<td>5.5 BEST AFFINE ESTIMATION OF A VECTOR</td>
<td>119</td>
</tr>
<tr>
<td>5.6 NONSTATISTICAL LEAST SQUARES (LINEAR REGRESSION)</td>
<td>121</td>
</tr>
<tr>
<td>5.7 LINEAR MINIMUM WEIGHTED LEAST SQUARES ESTIMATION</td>
<td>126</td>
</tr>
<tr>
<td>5.7.1 PROJECTION OPERATOR FORM OF LMWLS PREDICTOR</td>
<td>127</td>
</tr>
<tr>
<td>5.8 LMWMS ESTIMATOR IS MLE AND UMVUE IN THE GAUSSIAN MODEL</td>
<td>130</td>
</tr>
<tr>
<td>5.9 BACKGROUND REFERENCES</td>
<td>132</td>
</tr>
<tr>
<td>5.10 APPENDIX: VECTOR SPACES</td>
<td>132</td>
</tr>
<tr>
<td>5.11 EXERCISES</td>
<td>136</td>
</tr>
</tbody>
</table>
6  OPTIMAL LINEAR FILTERING AND PREDICTION 141
6.1  WIENER-HOPF EQUATIONS OF OPTIMAL FILTERING .......... 141
6.2  NON-CAUSAL ESTIMATION ..................................... 143
6.3  CAUSAL ESTIMATION .......................................... 144
   6.3.1  SPECIAL CASE OF WHITE NOISE MEASUREMENTS ......... 145
   6.3.2  GENERAL CASE OF NON-WHITE MEASUREMENTS .......... 145
6.4  CAUSAL PREWHITENING VIA SPECTRAL FACTORIZATION .......... 148
6.5  CAUSAL WIENER FILTERING .................................... 149
6.6  CAUSAL FINITE MEMORY TIME VARYING ESTIMATION .......... 154
   6.6.1  SPECIAL CASE OF UNCORRELATED MEASUREMENTS ....... 155
   6.6.2  CORRELATED MEASUREMENTS: THE INNOVATIONS FILTER ... 156
   6.6.3  INNOVATIONS AND CHOLESKY DECOMPOSITION ............ 157
6.7  TIME VARYING ESTIMATION/PREDICTION VIA THE KALMAN FILTER 158
   6.7.1  DYNAMICAL MODEL ........................................ 159
   6.7.2  KALMAN FILTER: ALGORITHM DEFINITION ................. 160
   6.7.3  KALMAN FILTER: DERIVATIONS ............................ 160
6.8  KALMAN FILTERING: SPECIAL CASES ............................ 167
   6.8.1  KALMAN PREDICTION ...................................... 167
   6.8.2  KALMAN FILTERING ....................................... 167
6.9  STEADY STATE KALMAN FILTER AND WIENER FILTER ............. 168
6.10 SUMMARY OF STATISTICAL PROPERTIES OF THE INNOVATIONS .... 169
6.11 KALMAN FILTER FOR SPECIAL CASE OF GAUSSIAN STATE AND NOISE 170
6.12 BACKGROUND REFERENCES ...................................... 170
6.13 APPENDIX: POWER SPECTRAL DENSITIES .......................... 170
   6.13.1  ACF AND CCF ........................................... 170
   6.13.2  REAL VALUED WIDE SENSE STATIONARY SEQUENCES ....... 171
   6.13.3  Z-DOMAIN PSD AND CPSD ................................ 172
6.14 EXERCISES ..................................................... 172

7  FUNDAMENTALS OF DETECTION 183
7.1  THE GENERAL DETECTION PROBLEM ............................... 188
   7.1.1  SIMPLE VS COMPOSITE HYPOTHESES ........................ 189
   7.1.2  DECISION RULES AND TEST FUNCTIONS ..................... 190
   7.1.3  FALSE ALARM AND MISS ERRORS ........................... 191
7.2  BAYES APPROACH TO DETECTION ................................ 192
   7.2.1  ASSIGNING PRIOR PROBABILITIES .......................... 192
   7.2.2  MINIMIZATION OF AVERAGE RISK ......................... 193
7.2.3 OPTIMAL BAYES TEST MINIMIZES \( E[C] \) ........................................... 194
7.2.4 MINIMUM PROBABILITY OF ERROR TEST ........................................ 194
7.2.5 PERFORMANCE OF BAYES LIKELIHOOD RATIO TEST ..................... 195
7.2.6 MIN-MAX BAYES DETECTOR ............................................................. 195
7.2.7 EXAMPLES ..................................................................................... 197

7.3 CLASSIFICATION: TESTING MULTIPLE HYPOTHESES .................... 198
7.3.1 PRIOR CLASS PROBABILITIES ......................................................... 201
7.3.2 OPTIMAL CLASSIFIER MINIMIZES AVERAGE COST .................... 201
7.3.3 DEFICIENCIES OF BAYES APPROACH .......................................... 204

7.4 FREQUENTIST APPROACH TO DETECTION .................................... 204
7.4.1 CASE OF SIMPLE HYPOTHESES: \( \theta \in \{\theta_0, \theta_1\} \) ................ 205

7.5 ROC CURVES FOR THRESHOLD TESTS ............................................. 209
7.6 P-VALUES AND LEVELS OF SIGNIFICANCE .................................... 219
7.7 BACKGROUND AND REFERENCES ..................................................... 220
7.8 EXERCISES ..................................................................................... 221

8 DETECTION STRATEGIES FOR COMPOSITE HYPOTHESES .............. 225
8.1 UNIFORMLY MOST POWERFUL (UMP) TESTS .................................. 225
8.2 GENERAL CONDITION FOR UMP TESTS: MONOTONE LIKELIHOOD RA-

8.3 COMPOSITE HYPOTHESIS DETECTION STRATEGIES ....................... 241
8.3.1 BAYESIAN MINIMUM PROBABILITY OF ERROR APPROACH TO

8.3.2 MINIMAX TESTS ........................................................................... 242
8.3.3 LOCALLY MOST POWERFUL (LMP) SINGLE SIDED TEST ............. 245
8.3.4 MOST POWERFUL UNBIASED (MPU) TESTS ................................. 253
8.3.5 LOCALLY MOST POWERFUL UNBIASED DOUBLE SIDED TEST ... 254
8.3.6 CFAR DETECTION ........................................................................ 258
8.3.7 INVARIANT TESTS ....................................................................... 258

8.4 THE GENERALIZED LIKELIHOOD RATIO TEST ................................. 259
8.4.1 PROPERTIES OF GLRT ................................................................. 260
8.5 BACKGROUND REFERENCES ............................................................ 260
8.6 EXERCISES ..................................................................................... 261
9 COMPOSITE HYPOTHESES IN THE UNIVARIATE GAUSSIAN MODEL 269

9.1 TESTS ON THE MEAN: $\sigma^2$ KNOWN ................................................. 269
  9.1.1 CASE III: $H_0: \mu = \mu_o, H_1: \mu \neq \mu_o$ ................................. 269

9.2 TESTS ON THE MEAN: $\sigma^2$ UNKNOWN ............................................. 271
  9.2.1 CASE I: $H_0: \mu = \mu_o, \sigma^2 > 0, H_1: \mu > \mu_o, \sigma^2 > 0$ .......... 271
  9.2.2 CASE II: $H_0: \mu \leq \mu_o, \sigma^2 > 0, H_1: \mu > \mu_o, \sigma^2 > 0$ ........ 274
  9.2.3 CASE III: $H_0: \mu = \mu_o, \sigma^2 > 0, H_1: \mu \neq \mu_o, \sigma^2 > 0$ ......... 274

9.3 TESTS ON VARIANCE: KNOWN MEAN .................................................. 274
  9.3.1 CASE I: $H_0: \sigma^2 = \sigma_o^2, H_1: \sigma^2 > \sigma_o^2$ .......................... 275
  9.3.2 CASE II: $H_0: \sigma^2 < \sigma_o^2, H_1: \sigma^2 > \sigma_o^2$ .......................... 276
  9.3.3 CASE III: $H_0: \sigma^2 = \sigma_o^2, H_1: \sigma^2 \neq \sigma_o^2$ ....................... 277

9.4 TESTS ON VARIANCE: UNKNOWN MEAN ................................................. 280
  9.4.1 CASE I: $H_0: \sigma^2 = \sigma_o^2, H_1: \sigma^2 > \sigma_o^2$ .......................... 280
  9.4.2 CASE II: $H_0: \sigma^2 < \sigma_o^2, \mu \in \mathbb{R}, H_1: \sigma^2 > \sigma_o^2, \mu \in \mathbb{R}$ 281
  9.4.3 CASE III: $H_0: \sigma^2 = \sigma_o^2, \mu \in \mathbb{R}, H_1: \sigma^2 \neq \sigma_o^2, \mu \in \mathbb{R}$ 281

9.5 TESTS ON MEANS OF TWO POPULATIONS: UNKNOWN COMMON VARIANCE .......... 281
  9.5.1 CASE I: $H_0: \mu_x = \mu_y, \sigma^2 > 0, H_1: \mu_x \neq \mu_y, \sigma^2 > 0$ ......... 281
  9.5.2 CASE II: $H_0: \mu_y \leq \mu_x, \sigma^2 > 0, H_1: \mu_y > \mu_x, \sigma^2 > 0$ ........ 285

9.6 TESTS ON EQUALITY OF VARIANCES OF TWO POPULATIONS ......................... 285
  9.6.1 CASE I: $H_0: \sigma^2_x = \sigma^2_y, H_1: \sigma^2_x \neq \sigma^2_y$ .................... 285
  9.6.2 CASE II: $H_0: \sigma^2_x = \sigma^2_y, H_1: \sigma^2_y > \sigma^2_x$ ...................... 286

9.7 TESTING FOR EQUAL MEANS AND VARIANCES OF TWO POPULATIONS 287

9.8 TESTS ON CORRELATION ................................................................. 288
  9.8.1 CASE I: $H_0: \rho = \rho_o, H_1: \rho \neq \rho_o$ .................................. 288
  9.8.2 CASE II: $H_0: \rho = 0, H_1: \rho > 0$ ............................................ 289

9.9 P-VALUES IN PRESENCE OF NUISANCE PARAMETERS .................................. 290

9.10 BACKGROUND REFERENCES .............................................................. 290

9.11 EXERCISES .................................................................................. 291

10 STATISTICAL CONFIDENCE INTERVALS ............................................... 292

10.1 DEFINITION OF A CONFIDENCE INTERVAL ........................................... 292

10.2 CONFIDENCE ON MEAN: KNOWN VAR ............................................... 293

10.3 CONFIDENCE ON MEAN: UNKNOWN VAR .............................................. 297

10.4 CONFIDENCE ON VARIANCE ............................................................... 298

10.5 CONFIDENCE ON DIFFERENCE OF TWO MEANS ....................................... 299

10.6 CONFIDENCE ON RATIO OF TWO VARIANCES ........................................ 299
11 SIGNAL DETECTION IN THE MULTIVARIATE GAUSSIAN MODEL 304
11.1 OFFLINE METHODS .............................................. 304
   11.1.1 GENERAL CHARACTERIZATION OF LRT DECISION REGIONS . 306
   11.1.2 CASE OF EQUAL COVARIANCES ............................ 309
   11.1.3 CASE OF EQUAL MEANS, UNEQUAL COVARIANCES ............ 324
11.2 APPLICATION: DETECTION OF RANDOM SIGNALS ..................... 330
11.3 DETECTION OF NON-ZERO MEAN NON-STATIONARY SIGNAL IN WHITE NOISE .......................................................... 339
11.4 ONLINE IMPLEMENTATIONS OF OPTIMAL DETECTORS .................. 339
   11.4.1 ONLINE DISCRIMINATION OF NON-STATIONARY SIGNALS ....... 340
   11.4.2 ONLINE DUAL KALMAN SIGNAL SELECTOR .................... 341
   11.4.3 ONLINE SIGNAL DETECTOR VIA CHOLESKY .................... 344
11.5 STEADY-STATE STATE-SPACE SIGNAL DETECTOR ...................... 346
11.6 BACKGROUND REFERENCES ........................................ 347
11.7 EXERCISES ...................................................... 349

12 COMPOSITE HYPOTHESES IN THE MULTIVARIATE GAUSSIAN MODEL 352
12.1 MULTIVARIATE GAUSSIAN MATRICES ................................ 353
12.2 DOUBLE SIDED TEST OF VECTOR MEAN ............................ 353
12.3 TEST OF EQUALITY OF TWO MEAN VECTORS ......................... 357
12.4 TEST OF INDEPENDENCE ........................................... 358
12.5 TEST OF WHITENESS .............................................. 359
12.6 CONFIDENCE REGIONS ON VECTOR MEAN ............................ 360
12.7 EXAMPLES ...................................................... 361
12.8 BACKGROUND REFERENCES ....................................... 364
12.9 EXERCISES ...................................................... 365

13 BIBLIOGRAPHY 366
1 INTRODUCTION

1.1 STATISTICAL SIGNAL PROCESSING

Many engineering applications require extraction of a signal or parameter of interest from degraded measurements. To accomplish this it is often useful to deploy fine-grained statistical models; diverse sensors which acquire extra spatial, temporal, or polarization information; or multi-dimensional signal representations, e.g. time-frequency or time-scale. When applied in combination these approaches can be used to develop highly sensitive signal estimation, detection, or tracking algorithms which can exploit small but persistent differences between signals, interferences, and noise. Conversely, these approaches can be used to develop algorithms to identify a channel or system producing a signal in additive noise and interference, even when the channel input is unknown but has known statistical properties.

Broadly stated, statistical signal processing is concerned with the reliable estimation, detection and classification of signals which are subject to random fluctuations. Statistical signal processing has its roots in probability theory, mathematical statistics and, more recently, systems theory and statistical communications theory. The practice of statistical signal processing involves: (1) description of a mathematical and statistical model for measured data, including models for sensor, signal, and noise; (2) careful statistical analysis of the fundamental limitations of the data including deriving benchmarks on performance, e.g. the Cramér-Rao, Ziv-Zakai, Barankin, Rate Distortion, Chernov, or other lower bounds on average estimator/detector error; (3) development of mathematically optimal or suboptimal estimation/detection algorithms; (4) asymptotic analysis of error performance establishing that the proposed algorithm comes close to reaching a benchmark derived in (2); (5) simulations or experiments which compare algorithm performance to the lower bound and to other competing algorithms. Depending on the specific application, the algorithm may also have to be adaptive to changing signal and noise environments. This requires incorporating flexible statistical models, implementing low-complexity real-time estimation and filtering algorithms, and on-line performance monitoring.

1.2 PERSPECTIVE ADOPTED IN THIS BOOK

This book is at the interface between mathematical statistics and signal processing. The idea for the book arose in 1986 when I was preparing notes for the engineering course on detection, estimation and filtering at the University of Michigan. There were then no textbooks available which provided a firm background on relevant aspects of mathematical statistics and multivariate analysis. These fields of statistics formed the backbone of this engineering field in the 1940’s 50’s and 60’s when statistical communication theory was first being developed. However, more recent textbooks have downplayed the important role of statistics in signal processing in order to accommodate coverage of technological issues of implementation and data acquisition for specific engineering applications such as radar, sonar, and communications. The result is that students finishing the course would have a good notion of how to solve focussed problems in these applications but would find it difficult either to extend the theory to a moderately different problem or to apply the considerable power and generality of mathematical statistics to other applications areas.

The technological viewpoint currently in vogue is certainly a useful one; it provides an essential engineering backdrop to the subject which helps motivate the engineering students. However, the disadvantage is that such a viewpoint can produce a disjointed presentation of the component
parts of statistical signal processing making it difficult to appreciate the commonalities between
detection, classification, estimation, filtering, pattern recognition, confidence intervals and other
useful tools. These commonalities are difficult to appreciate without adopting a proper statistical
perspective. This book strives to provide this perspective by more thoroughly covering elements of
mathematical statistics than other statistical signal processing textbooks. In particular we cover
point estimation, interval estimation, hypothesis testing, time series, and multivariate analysis.
In adopting a strong statistical perspective the book provides a unique viewpoint on the subject
which permits unification of many areas of statistical signal processing which are otherwise difficult
to treat in a single textbook.

The book is organized into chapters listed in the attached table of contents. After a quick review
of matrix algebra, systems theory, and probability, the book opens with chapters on fundamentals
of mathematical statistics, point estimation, hypothesis testing, and interval estimation in the
standard context of independent identically distributed observations. Specific topics in these
chapters include: least squares techniques; likelihood ratio tests of hypotheses; e.g. testing for
whiteness, independence, in single and multi-channel populations of measurements. These chapters
provide the conceptual backbone for the rest of the book. Each subtopic is introduced with a set
of one or two examples for illustration. Many of the topics here can be found in other graduate
textbooks on the subject, e.g. those by Van Trees, Kay, and Srinath et al. However, the coverage
here is broader with more depth and mathematical detail which is necessary for the sequel of the
textbook. For example in the section on hypothesis testing and interval estimation the full theory
of sampling distributions is used to derive the form and null distribution of the standard statistical
tests of shift in mean, variance and correlation in a Normal sample.

The second part of the text extends the theory in the previous chapters to non i.i.d. sampled
Gaussian waveforms. This group contains applications of detection and estimation theory to single
and multiple channels. As before, special emphasis is placed on the sampling distributions of
the decision statistics. This group starts with offline methods; least squares and Wiener filtering;
and culminates in a compact introduction of on-line Kalman filtering methods. A feature not found
in other treatments is the separation principle of detection and estimation which is made explicit
via Kalman and Wiener filter implementations of the generalized likelihood ratio test for model
selection, reducing to a whiteness test of each the innovations produced by a bank of Kalman
filters. The book then turns to a set of concrete applications areas arising in radar, communications,
auditory and radar signal processing, imaging, and other areas of signal processing. Topics
include: testing for independence; parametric and non-parametric testing of a sample distribution;
extensions to complex valued and continuous time observations; optimal coherent and incoherent
receivers for digital and analog communications;

A future revision will contain chapters on performance analysis, including asymptotic analysis
and upper/lower bounds on estimators and detector performance; non-parametric and semipara-
tmetric methods of estimation; iterative implementation of estimators and detectors (Monte Carlo
Markov Chain simulation and the EM algorithm); classification, clustering, and sequential de-
sign of experiments. It may also have chapters on applications areas including: testing of binary
Markov sequences and applications to internet traffic monitoring; spatio-temporal signal process-
ing with multi-sensor sensor arrays; CFAR (constant false alarm rate) detection strategies for
Electro-optical (EO) and Synthetic Aperture Radar (SAR) imaging; and channel equalization.
1.2.1 PREREQUISITES

Readers are expected to possess a background in basic probability and random processes at the level of Stark&Woods [78], Ross [68] or Papoulis [63], exposure to undergraduate vector and matrix algebra at the level of Noble and Daniel [61] or Shilov [74], and basic undergraduate course on signals and systems at the level of Oppenheim and Willsky [62]. These notes have evolved as they have been used to teach a first year graduate level course (42 hours) in the Department of Electrical Engineering and Computer Science at the University of Michigan from 1997 to 2010 and a one week short course (40 hours) given at EG&G in Las Vegas in 1998.

The author would like to thank Hyung Soo Kim, Robby Gupta, and Mustafa Demirci for their help with drafting the figures for these notes. He would also like to thank the numerous students at UM whose comments led to an improvement of the presentation. Special thanks goes to Laura Balzano and Clayton Scott of the University of Michigan, Raviv Raich of Oregon State University and Aaron Lanterman of Georgia Tech who provided detailed comments and suggestions for improvement of earlier versions of these notes. **End of chapter**
2 NOTATION, MATRIX ALGEBRA, SIGNALS AND SYSTEMS

Keywords: vector and matrix operations, matrix inverse identities, linear systems, transforms, convolution, correlation.

Before launching into statistical signal processing we need to set the stage by defining our notation. We then briefly review some elementary concepts in linear algebra and signals and systems. At the end of the chapter you will find some useful references for this review material.

2.1 NOTATION

We attempt to stick with widespread notational conventions in this text. However inevitably exceptions must sometimes be made for clarity.

In general upper case letters, e.g. $X, Y, Z$, from the end of the alphabet denote random variables, i.e. functions on a sample space, and their lower case versions, e.g. $x$, denote realizations, i.e. evaluations of these functions at a sample point, of these random variables. We reserve lower case letters from the beginning of the alphabet, e.g. $a, b, c$, for constants and lower case letters in the middle of the alphabet, e.g. $i, j, k, l, m, n$, for integer variables. Script and calligraphic characters, e.g. $\mathcal{S}, \mathcal{I}, \mathcal{O}, \mathcal{X}$, are used to denote sets of values. Exceptions are calligraphic upper case letters that denote standard probability distributions, e.g. Gaussian, Cauchy, and Student-t distributions $\mathcal{N}(x), \mathcal{C}(v), \mathcal{T}(t)$, respectively, and script notation for power spectral density $\mathcal{P}_X$. Vector valued quantities, e.g. $\mathbf{x}, \mathbf{X}$, are denoted with an underscore and matrices, e.g. $\mathbf{A}$, are bold upper case letters from the beginning of the alphabet. An exception is the matrix $\mathbf{R}$ that we use for the covariance matrix of a random vector. The elements of an $m \times n$ matrix $\mathbf{A}$ are denoted generically $a_{ij}^{m,n}$ and we also write $\mathbf{A} = (a_{ij})_{i,j=1}^{m,n}$ when we need to spell out the entries explicitly.

The letter $f$ is reserved for a probability density function and $p$ is reserved for a probability mass function. Finally in many cases we deal with functions of two or more variables, e.g. the density function $f(x; \theta)$ of a random variable $X$ parameterized by a parameter $\theta$. We use subscripts to emphasize that we are fixing one of the variables, e.g. $f_\theta(x)$ denotes the density function over $x$ in a sample space $\mathcal{X} \subset \mathbb{R}$ for a fixed $\theta$ in a parameter space $\Theta$. However, when dealing with multivariate densities for clarity we will prefer to explicitly subscript with the appropriate ordering of the random variables, e.g. $f_{X,Y}(x, y; \theta)$ or $f_{X|Y}(x|y; \theta)$.

2.2 VECTOR AND MATRIX BACKGROUND

2.2.1 ROW AND COLUMN VECTORS

A vector is an ordered list of $n$ values:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

which resides in $\mathbb{R}^n$.

Convention: in this course $\mathbf{x}$ is (almost) always a column vector. Its transpose is the row vector...
\[ \mathbf{x}^T = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \]

When the elements \( x_i = u + jv \) are complex (\( u, v \) real valued, \( j = \sqrt{-1} \)) the Hermitian transpose is defined as

\[ \mathbf{x}^H = \begin{bmatrix} x_1^* & \cdots & x_n^* \end{bmatrix} \]

where \( x_i^* = u - jv \) is the complex conjugate of \( x_i \).

Some common vectors we will see are the vector of all ones and the \( j \)-th elementary vector, which is the \( j \)-th column of the identity matrix:

\[ \mathbf{1} = [1, \ldots, 1]^T, \quad \mathbf{e}_j = [0, \ldots, 0, 1, 0, \ldots, 0]^T \]

### 2.2.2 VECTOR/VECTOR MULTIPLICATION

For 2 vectors \( \mathbf{x} \) and \( \mathbf{y} \) with the same number \( n \) of entries, their “inner product” is the scalar

\[ \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{n} x_i y_i \]

The 2-norm \( \|\mathbf{x}\|_2 \) of a vector \( \mathbf{x} \) is its length and it is defined as (we drop the norm subscript when there is no risk of confusion)

\[ \|x\| = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{n} x_i^2} \]

For 2 vectors \( \mathbf{x} \) and \( \mathbf{y} \) of possibly different lengths \( n, m \) their “outer product” is the \( n \times m \) matrix

\[ \mathbf{xy}^T = (x_i y_j)_{i,j=1}^{n,m} \]

\[ = \begin{bmatrix} xy_1, \ldots, xy_m \end{bmatrix} \]

\[ = \begin{bmatrix} x_1y_1 & \cdots & x_1y_m \\ \vdots & \ddots & \vdots \\ x_ny_1 & \cdots & x_ny_m \end{bmatrix} \]

### 2.3 ORTHOGONAL VECTORS

If \( \mathbf{x}^T \mathbf{y} = 0 \) then \( \mathbf{x} \) and \( \mathbf{y} \) are said to be orthogonal. If in addition the lengths of \( \mathbf{x} \) and \( \mathbf{y} \) are equal to one, \( \|\mathbf{x}\| = 1 \) and \( \|\mathbf{y}\| = 1 \), then \( \mathbf{x} \) and \( \mathbf{y} \) are said to be orthonormal vectors.
2.3.1 VECTOR/MATRIX MULTIPLICATION

Let $A$ be an $m \times n$ matrix with columns $a_1, \ldots, a_n$ and $x$ be any $n$-element vector.

The (compatible) product $A x$ is a (column) vector composed of linear combinations of the columns of $A$

$$A x = \sum_{j=1}^{n} x_j a_j$$

For $y$ an $m$-element vector the product $y^T A$ is a (row) vector composed of linear combinations of the rows of $A$

$$y^T A = \sum_{i=1}^{m} y_i a_i$$

2.3.2 THE LINEAR SPAN OF A SET OF VECTORS

Let $x_1, \ldots, x_n$ be a set of $p$ dimensional (column) vectors and construct the $p \times n$ matrix

$$X = [x_1, \ldots, x_n].$$

Let $a = [a_1, \ldots, a_n]^T$ be a vector of coefficients. Then $y = \sum_{i=1}^{n} a_i x_i = X a$ is another $p$ dimensional vector that is a linear combination of the columns of $X$. The linear span of the vectors $x_1, \ldots, x_n$, equivalently, the column space or range of $X$, is defined as the subspace of $\mathbb{R}^p$ that contains all such linear combinations:

$$\text{span}\{x_1, \ldots, x_n\} = \{y : y = X a, \ a \in \mathbb{R}^n\}.$$ 

In other words, when we allow $a$ to sweep over its entire domain $\mathbb{R}^n$, $y$ sweeps over the linear span of $x_1, \ldots, x_n$.

2.3.3 RANK OF A MATRIX

The (column) rank of a matrix $A$ is equal to the number of its columns that are linearly independent. The dimension of the column space of a rank $p$ matrix $A$ is equal to $p$.

If $A$ has full rank then

$$0 = A x = \sum_{i} x_i a_{si} \iff x = 0.$$ 

If in addition $A$ is square then it is said to be non-singular.

2.3.4 MATRIX INVERSION

If $A$ is non-singular square matrix then it has an inverse $A^{-1}$ that satisfies the relation $A A^{-1} = I$.

In the special case of a $2 \times 2$ matrix the matrix inverse is given by (Cramèr’s formula)

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad \text{if} \quad ad \neq bc$$
Sometimes when a matrix has special structure its inverse has a simple form. The books by Graybill [24] and Golub and VanLoan [22] give many interesting and useful examples. Some results which we will need in this text are: the Sherman-Morrison-Woodbury identity

\[
[A + UV^T]^{-1} = A^{-1} - A^{-1}U[I + V^TA^{-1}U]^{-1}V^TA^{-1},
\]

where \(A, U, V\) are compatible matrices, \([A + UV^T]^{-1}\) and \(A^{-1}\) exist; and the partitioned matrix inverse identity

\[
\begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}^{-1} = \begin{bmatrix}
A_{11}^{-1} - A_{12}A_{22}^{-1}A_{21} & -A_{11}^{-1}A_{12}A_{22}^{-1}A_{21}^{-1}
\end{bmatrix} = \begin{bmatrix}
A_{22} - A_{21}A_{11}^{-1}A_{12}^{-1}
\end{bmatrix},
\]

assuming that all the indicated inverses exist.

\subsection{ORTHOGONAL AND UNITARY MATRICES}

A real square matrix \(A\) is said to be orthogonal if all of its columns are orthonormal, i.e.,

\[
A^TA = I.
\]

The generalization of orthogonality to complex matrices \(A\) is the property of being unitary,

\[
A^H A = I.
\]

The relation (3) implies that if \(A\) is an orthogonal matrix it is invertible and has a very simple inverse

\[
A^{-1} = A^T.
\]

\subsection{GRAMM-SCHMIDT ORTHOGONALIZATION AND ORTHONORMALIZATION}

Let \(x_1, \ldots, x_n\) be a set of \(n\) linearly independent \(p\) dimensional column vectors \((n \leq p)\) whose linear span is the subspace \(H\). Gramm-Schmidt orthogonalization is an algorithm that can be applied to this set of vectors to obtain a set of \(n\) orthogonal vectors \(y_1, \ldots, y_n\) that spans the same subspace. This algorithm proceeds as follows.

**Step 1:** select \(y_1\) as an arbitrary starting point in \(H\). For example, choose any coefficient vector \(a_1 = [a_{11}, \ldots, a_{1n}]^T\) and define \(y_1 = Xa_1\) where \(X = [x_1, \ldots, x_n]\).

**Step 2:** construct the other \(n - 1\) vectors \(y_2, \ldots, y_n\) by the following recursive procedure:

\[
\text{For } j = 2, \ldots, n: \quad y_j = x_j - \sum_{i=1}^{j-1} K_i y_i \text{ where } K_j = X_j^T y_{j-1} / y_{j-1}^T y_{j-1}.
\]

The above Gramm-Schmidt procedure can be expressed in compact matrix form [69]

\[
Y = HX,
\]

where \(Y = [y_1, \ldots, y_n]\) and \(H\) is called the Gramm-Schmidt matrix.

If after each step \(j = 1, \ldots, n\) of the procedure one maps normalizes the length of \(y_j\), i.e., \(y_j \leftarrow \hat{y}_j = y_j / \|y_j\|\), the algorithm produces an orthonormal set of vectors. This is called Gram-Schmidt
orthonormalization and produces an matrix $\mathbf{Y}$ with orthonormal columns and identical column span as that of $\mathbf{X}$. The Gramm-Schmidt orthonormalization procedure is often used to generate an orthonormal basis $y_1, \ldots, y_p$ for $\mathbb{R}^p$ starting from an arbitrarily selected initial vector $y_1$. The matrix formed from such a basis will have the structure

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

and

$$\mathbf{Y}^T \mathbf{Y} = \mathbf{I}.$$  

In the above $y_2, \ldots, y_p$ are orthonormal vectors that are said to accomplish completion of the basis with respect to the initial vector $y_1$.

### 2.3.7 EIGENVALUES OF A SYMMETRIC MATRIX

If $\mathbf{R}$ is arbitrary $n \times n$ symmetric matrix, that is, $\mathbf{R}^T = \mathbf{R}$, then there exist a set of $n$ orthonormal eigenvectors $v_i$,

$$v_i^T v_j = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

and a set of associated eigenvectors $\lambda_i$ such that:

$$\mathbf{R} v_i = \lambda_i v_i, \quad i = 1, \ldots, n.$$ 

These eigenvalues and eigenvectors satisfy:

$$v_i^T \mathbf{R} v_i = \lambda_i$$

$$v_i^T \mathbf{R} v_j = 0, \quad i \neq j.$$ 

### 2.3.8 MATRIX DIAGONALIZATION AND EIGENDECOMPOSITION

Let $\mathbf{U} = [v_1, \ldots, v_n]$ be the $n \times n$ matrix formed from the eigenvectors of a symmetric matrix $\mathbf{R}$. If $\mathbf{R}$ is real symmetric $\mathbf{U}$ is a real orthogonal matrix while if $\mathbf{R}$ is complex Hermitian symmetric $\mathbf{U}$ is a complex unitary matrix:

$$\mathbf{U}^T \mathbf{U} = \mathbf{I}, \quad \text{(U an orthogonal matrix)}$$

$$\mathbf{U}^H \mathbf{U} = \mathbf{I}, \quad \text{(U a unitary matrix)}.$$ 

where as before $H$ denotes Hermitian transpose. As the Hermitian transpose of a real matrix is equal to its ordinary transpose, we will use the more general notation $\mathbf{A}^H$ for any (real or complex) matrix $\mathbf{A}$.

The matrix $\mathbf{U}$ can be used to diagonalize $\mathbf{R}$

$$\mathbf{U}^H \mathbf{R} \mathbf{U} = \mathbf{\Lambda},$$

(4)
In cases of both real and Hermitian symmetric $R$ the matrix $A$ is diagonal and real valued
\[
A = \text{diag}(\lambda_i) = \begin{bmatrix}
\lambda_1 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & \lambda_n
\end{bmatrix},
\]
where $\lambda_i$’s are the eigenvalues of $R$.

The expression (4) implies that
\[
R = U\Lambda U^H,
\]
which is called the eigendecomposition of $R$. As $A$ is diagonal, an equivalent summation form for this eigendecomposition is
\[
R = \sum_{i=1}^{n} \lambda_i u_i u_i^H. 
\]

### 2.3.9 Quadratic Forms and Non-Negative Definite Matrices

For a square symmetric matrix $R$ and a compatible vector $x$, a quadratic form is the scalar defined by $x^T R x$. The matrix $R$ is non-negative definite (nnd) if for any $x$
\[
x^T R x \geq 0.
\]

$R$ is positive definite (pd) if it is nnd and "=" in (6) implies that $x = 0$, or more explicitly $R$ is pd if
\[
x^T R x > 0, \quad x \neq 0.
\]

Examples of nnd (pd) matrices:
* $R = B^T B$ for arbitrary (pd) matrix $B$
* $R$ symmetric with only non-negative (positive) eigenvalues

**Rayleigh Theorem:** If $A$ is a nnd $n \times n$ matrix with eigenvalues $\{\lambda_i\}_{i=1}^{n}$ the quadratic form
\[
\min(\lambda_i) \leq \frac{u^T A u}{u^T u} \leq \max(\lambda_i)
\]
where the lower bound is attained when $u$ is the eigenvector of $A$ associated with the minimum eigenvalue of $A$ and the upper bound is attained by the eigenvector associated with the maximum eigenvalue of $A$.

### 2.4 Positive Definiteness of Symmetric Partitioned Matrices

If $A$ is a symmetric matrix with partition representation (2) then it is easily shown that
\[
A = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix} = \begin{bmatrix}
I & -A_{12}A_{22}^{-1} \\
0 & I
\end{bmatrix}^{-1} \begin{bmatrix}
A_{11} - A_{12}A_{22}^{-1}A_{21} & O^T \\
O & A_{22}
\end{bmatrix} \begin{bmatrix}
I & O^T \\
-A_{22}^{-1}A_{21} & I
\end{bmatrix}^{-1}, \tag{8}
\]
as long as $A_{22}^{-1}$ exists. Here $O$ denotes a block of zeros. This implies: if $A$ is positive definite the matrices $A_{11} - A_{12}A_{22}^{-1}A_{21}$ and $A_{22}$ are pd. By using an analogous identity we can conclude that $A_{22} - A_{21}A_{11}^{-1}A_{12}$ and $A_{11}$ are also pd.
2.4.1 DETERMINANT OF A MATRIX

If \( A \) is any square matrix its determinant is

\[
|A| = \prod_i \lambda_i
\]

Note: a square matrix is non-singular iff its determinant is non-zero.

If \( A \) is partitioned as in (2) and \( A_{11}^{-1} \) and \( A_{22}^{-1} \) exist then

\[
|A| = |A_{11}| |A_{22}| - A_{21} A_{11}^{-1} A_{12} = |A_{22}| |A_{11} - A_{12} A_{22}^{-1} A_{21}|
\]  
(9)

This follows from the decomposition (8).

2.4.2 TRACE OF A MATRIX

For any square matrix \( A = (a_{ij}) \) the trace of \( A \) is defined as

\[
\text{trace}(A) = \sum_i a_{ii} = \sum_i \lambda_i
\]

One has an important identity: for compatible matrices \( A \) and \( B \)

\[
\text{trace}(AB) = \text{trace}(BA).
\]

This has the following implication for quadratic forms:

\[
x^T R x = \text{trace}(xx^T R).
\]

2.4.3 VECTOR DIFFERENTIATION

Differentiation of functions of a vector variable often arise in signal processing and estimation theory. If \( h = [h_1, \ldots, h_n]^T \) is an \( n \times 1 \) vector and \( g(h) \) is a scalar function then the gradient of \( g(h) \), denoted \( \nabla g(h) \) or \( \nabla_h g(h) \) when necessary for conciseness, is defined as the (column) vector of partials

\[
\nabla g = \left[ \frac{\partial g}{\partial h_1}, \ldots, \frac{\partial g}{\partial h_n} \right]^T.
\]

In particular, if \( c \) is a constant

\[
\nabla_h c = 0.
\]

if \( \bar{x} = [x_1, \ldots, x_n]^T \)

\[
\nabla_h (\bar{x}^T \bar{x}) = \nabla_h (\bar{x}^T h) = \bar{x},
\]

and if \( B \) is an \( n \times n \) matrix

\[
\nabla_h (h - \bar{x})^T B(h - \bar{x}) = 2B(h - \bar{x}).
\]

For a vector valued function \( g(h) = [g_1(h), \ldots, g_m(h)]^T \) the gradient of \( g(h) \) is an \( m \times n \) matrix. In particular, for a scalar function \( g(h) \), the two applications of the gradient \( \nabla(\nabla g)^T \) gives the \( n \times n \) Hessian matrix of \( g \), denoted as \( \nabla^2 g \). This yields useful and natural identities such as:

\[
\nabla_h^2 (h - \bar{x})^T B(h - \bar{x}) = 2B.
\]

For a more detailed discussion of vector differentiation the reader is referred to Kay [40].
2.5 SIGNALS AND SYSTEMS BACKGROUND

Here we review some of the principal results that will be useful for dealing with signals and systems encountered in this book.

2.5.1 GEOMETRIC SERIES

One of the most useful formulas in discrete time signal and systems engineering is:

\[
\sum_{i=0}^{n} a^n = \frac{1 - a^{n+1}}{1 - a}, \quad \text{if } a \neq 1; \quad \sum_{i=0}^{\infty} a^n = \frac{1}{1 - a}, \quad \text{if } |a| < 1.
\]

2.5.2 LAPLACE AND FOURIER TRANSFORMS OF FUNCTIONS OF A CONTINUOUS VARIABLE

If \( h(t), -\infty < t < \infty \), a square integrable function of a continuous variable \( t \) (usually time) then its Laplace and Fourier transforms are defined as follows.

The Laplace transform of \( h \) is

\[
\mathcal{L}\{h\} = H(s) = \int_{-\infty}^{\infty} h(t)e^{-st} \, dt
\]

where \( s = \sigma + j\omega \in \mathbb{C} \) is a complex variable.

The Fourier transform of \( h \) is

\[
\mathcal{F}\{h\} = H(\omega) = \int_{-\infty}^{\infty} h(t)e^{-j\omega t} \, dt
\]

Note: \( \mathcal{F}\{h\} = \mathcal{L}\{h\}|_{s=j\omega} \).

Example: if \( h(t) = e^{-at}u(t) \), for \( a > 0 \), then the Laplace transform is

\[
H(s) = \int_{0}^{\infty} e^{-at}e^{-st} \, dt = \int_{0}^{\infty} e^{-(a+s)t} \, dt = \frac{-1}{a+s}e^{-a\omega} \bigg|_{0}^{\infty} = \frac{1}{a+s}
\]

2.5.3 Z-TRANSFORM AND DISCRETE-TIME FOURIER TRANSFORM (DTFT)

If \( h_k, k = \ldots, -1, 0, 1, \ldots \), is a square summable function of a discrete variable then its Z-transform and discrete-time Fourier transform (DTFT) are defined as follows.

The Z-transform is

\[
\mathcal{Z}\{h\} = H(z) = \sum_{k=-\infty}^{\infty} h_kz^{-k}
\]

The DTFT is

\[
\mathcal{F}\{h\} = H(\omega) = \sum_{k=-\infty}^{\infty} h_k e^{-j\omega k}
\]

Note: \( H(\omega) \) really means \( H(e^{j\omega}) \) and is an abuse of notation.
The DTFT is always periodic in $\omega$ with period $2\pi$.

Example: if $h_k = a^{|k|}$, then for $|az^{-1}| < 1$ and $|az| < 1$, the Z-transform is

$$H(z) = \sum_{k=-\infty}^{\infty} a^{|k|}z^{-k} = \sum_{k=-\infty}^{-1} a^{-k}z^{-k} + \sum_{k=0}^{\infty} a^k z^{-k} \]

$$H(z) = \sum_{k=1}^{\infty} (az)^k + \sum_{k=0}^{\infty} (az^{-1})^k = \frac{az}{1-az} + \frac{1}{1-az^{-1}}$$

Likewise the DTFT is (for $|a| < 1$):

$$H(\omega) = H(z)|_{z = e^{j\omega}} = \frac{1 - a^2}{1 - 2a \cos \omega + a^2}$$

### 2.5.4 CONVOLUTION: CONTINUOUS TIME

If $h(t)$ and $x(t)$ are square integrable functions of a continuous variable $t$ then the convolution of $x$ and $h$ is defined as

$$(h * x)(t) = \int_{-\infty}^{\infty} h(t - \tau)x(\tau) \, d\tau$$

Note: The convolution of $h$ and $x$ is a waveform indexed by time $t$. $(h * x)(t)$ is this waveform evaluated at time $t$ and is frequently denoted $h(t) * x(t)$.

Example: $h(t) = e^{-at}u(t)$, for $a > 0$, (the filter) and $x(t) = e^{-bt}u(t)$, for $b > 0$, (the filter input) then

$$(h * x)(t) = \int_{-\infty}^{\infty} e^{-a(t-\tau)}e^{-br}u(t-\tau)u(\tau) \, d\tau = \left(\int_{0}^{t} e^{-a(t-\tau)}e^{-br} \, d\tau\right) u(t)$$

$$= e^{-at} \left(\int_{0}^{t} e^{-(b-a)\tau} \, d\tau\right) u(t) = e^{-at} \left(-\frac{1}{b-a} e^{-(b-a)t}\bigg|_{0}^{t}\right) u(t) = \frac{e^{-at} - e^{-bt}}{b-a} u(t)$$

### 2.5.5 CONVOLUTION: DISCRETE TIME

If $h_k$ and $x_k$ are square integrable sequences then

$$h_n * x_n = \sum_{j=-\infty}^{\infty} h_j x_{n-j} = \sum_{j=-\infty}^{\infty} h_{n-j} x_j$$

$h_k$ is called a “causal” filter if it is zero for negative indices:

$$h_k = 0, \quad k < 0$$
2.5.6 CORRELATION: DISCRETE TIME

For time sequences \( \{x_k\}_{k=1}^{n} \) and \( \{y_k\}_{k=1}^{n} \) their temporal correlation is

\[
z_n = \sum_{j=1}^{n} x_j y_j^*
\]

2.5.7 RELATION BETWEEN CORRELATION AND CONVOLUTION

The temporal correlation is directly related to the convolution of \( x_k \) with a filter impulse response \( h_k \) where the output of the filter is sampled at time \( k = n \):

\[
z_n = \sum_{j=1}^{n} x_j y_j^* = \sum_{j=-\infty}^{\infty} x_k h_{n-k} = h_n \ast x_n,
\]

where the filter impulse response is equal to the shifted and time reversed signal \( y_k \),

\[
h_k = \begin{cases} y_{n-k}, & k = 1, \ldots, n \\ 0, & \text{o.w.} \end{cases}
\]

The filter \( h_k \) is called the matched filter and is used for optimal detection of a known signal \( \{y_k\} \) in white Gaussian noise.

2.5.8 CONVOLUTION AS A MATRIX OPERATION

Let \( h_k \) be a causal filter impulse response and let \( x_k \) be an input starting at time \( k = 1 \). Arranging \( n \) outputs \( z_k \) in a vector \( z \) it is easy to see that

\[
\begin{bmatrix} z_n \\ \vdots \\ z_1 \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} h_{n-j} x_j \\ \vdots \\ \sum_{j=1}^{n} h_{1-j} x_j \end{bmatrix} = \begin{bmatrix} h_0 & h_1 & \cdots & h_{n-1} \\ 0 & h_0 & \cdots & h_{n-2} \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & h_0 \end{bmatrix} \begin{bmatrix} x_n \\ \vdots \\ x_1 \end{bmatrix}
\]

2.6 BACKGROUND REFERENCES

There are many useful textbooks that cover areas of this chapter. I learned elementary linear algebra from Noble and Daniel [61]. A more advanced book that is focused on computational linear algebra is Golub and Van Loan [22] which covers many fast and numerically stable algorithms arising in signal processing. Another nice book on linear algebra with emphasis on statistical applications is Graybill [24] that contains lots of useful identities for multivariate Gaussian models. For background on signals and systems Oppenheim and Willsky [62] and Proakis and Manolakis [65] are good elementary textbooks. The encyclopedic book by Moon and Stirling [57] is a good general resource for mathematical methods in signal processing.
2.7 EXERCISES

2.1 Let \( \mathbf{a}, \mathbf{b} \) be \( n \times 1 \) vectors and let \( \mathbf{C} \) be an invertible \( n \times n \) matrix. Assuming \( \alpha \) is not equal to \(-1/((\mathbf{a}^T \mathbf{C}^{-1} \mathbf{b}))\) show the following identity:

\[
|\mathbf{C} + \alpha \mathbf{a} \mathbf{b}^T|^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1} \mathbf{a} \mathbf{b}^T \mathbf{C}^{-1} \alpha / (1 + \alpha \mathbf{a}^T \mathbf{C}^{-1} \mathbf{b}).
\]

2.2 A discrete time LTI filter \( h(k) \) is causal when \( h(k) = 0, \ k < 0 \) and anticausal when \( h(k) = 0, \ k > 0 \). Show that if \( |h(k)| < \infty \) for all \( k \), the transfer function \( H(z) = \sum_{k=-\infty}^{\infty} h(k) z^{-k} \) of a causal LTI has no singularities outside the unit circle, i.e. \( |H(z)| < \infty, \ |z| > 1 \) while an anticausal LTI has no singularities inside the unit circle, i.e. \( |H(z)| < \infty, \ |z| < 1 \). (Hint: generalized triangle inequality: \( \sum |a_i| \leq \sum |a_i| \))

2.3 A discrete time LTI filter \( h(k) \) is said to be BIBO stable when \( \sum_{k=-\infty}^{\infty} |h(k)| < \infty \). Define the transfer function (Z-transform) \( H(z) = \sum_{k=-\infty}^{\infty} h(k) z^{-k} \), for \( z \) a complex variable.

(a) Show that \( H(z) \) has no singularities on the unit circle, i.e. \( |H(z)| < \infty, \ |z| = 1 \).

(b) Show that if a BIBO stable \( h(k) \) is causal then \( H(z) \) has all its singularities (poles) strictly inside the unit circle, i.e. \( |H(z)| < \infty, \ |z| \geq 1 \).

(c) Show that if a BIBO stable \( h(k) \) is anticausal, i.e. \( h(k) = 0, \ k > 0 \), then \( H(z) \) has all its singularities (poles) strictly outside the unit circle, i.e. \( |H(z)| < \infty, \ |z| \leq 1 \).

2.4 If you are only given the mathematical form of the transfer function \( H(z) \) of an LTI, and not told whether it corresponds to an LTI which is causal, anticausal, or stable, then it is not possible to uniquely specify the impulse response \( \{h_k\}_k \). This simple example illustrates this fact. The regions \( \{z: |z| > a\} \) and \( \{z: |z| \leq a\} \), specified in (a) and (b) are called the regions of convergence of the filter and specify whether the filter is stable, causal or anticausal.

Let \( H(z) \) be

\[
H(z) = \frac{1}{1 - az^{-1}}.
\]

(a) Show that if the LTI is causal, then for \( |z| > |a| \) you can write \( H(z) \) as the convergent series

\[
H(z) = \sum_{k=0}^{\infty} a^k z^{-k}, \quad |z| > |a|
\]

which corresponds to \( h_k = a^k, \ k = 0, 1, \ldots \) and \( h_k = 0, \ k < 0 \).

(b) Show that if the LTI is anticausal, then for \( |z| < |a| \) you can write \( H(z) \) as the convergent series

\[
H(z) = -\sum_{k=0}^{\infty} a^{-k} z^{k+1}, \quad |z| < |a|
\]

which corresponds to \( h_k = -a^{-k}, \ k = 1, 2 \ldots \) and \( h_k = 0, \ k \geq 0 \).

(c) Show that if \( |a| < 1 \) then the causal LTI is BIBO stable while the anti-causal LTI is BIBO unstable while if \( |a| > 1 \) then the reverse is true. What happens to stability when \( |a| = 1 \)?

2.5 An LTI has transfer function

\[
H(z) = \frac{3 - 4z^{-1}}{1 - 3.5z^{-1} + 1.5z^{-2}}
\]
(a) If you are told that the LTI is stable specify the region of convergence (ROC) in the $z$-plane, i.e. specify the range of values of $|z|$ for which $|H(z)| < \infty$, and specify the impulse response.

(b) If you are told that the LTI is causal specify the region of convergence (ROC) in the $z$-plane, and specify the impulse response.

(c) If you are told that the LTI is anticausal specify the region of convergence (ROC) in the $z$-plane, and specify the impulse response.

End of chapter
3  STATISTICAL MODELS

Keywords: sampling distributions, sufficient statistics, exponential families.

Estimation, detection and classification can be grouped under the broad heading of statistical inference which is the process of inferring properties about the distribution of a random variable $X$ given a realization $x$, which is also called a data sample, a measurement, or an observation. A key concept is that of the statistical model which is simply a hypothesized probability distribution or density function $f(x)$ for $X$. Broadly stated statistical inference explores the possibility of fitting a given model to the data $x$. To simplify this task it is common to restrict $f(x)$ to a class of parameteric models $\{f(x; \theta)\}_{\theta \in \Theta}$, where $f(x; \bullet)$ is a known function and $\theta$ is a vector of unknown parameters taking values in a parameter space $\Theta$. In this special case statistical inference boils down to inferring properties of the true value of $\theta$ parameterizing $f(x; \theta)$ that generated the data sample $x$.

In this chapter we discuss several models that are related to the ubiquitous Gaussian distribution, the more general class of exponential families of distributions, and the important concept of a sufficient statistic for inferring properties about $\theta$.

3.1  THE GAUSSIAN DISTRIBUTION AND ITS RELATIVES

The Gaussian distribution and its close relatives play a major role in parameteric statistical inference due to the relative simplicity of the Gaussian model and its broad applicability (recall the Central Limit Theorem!). Indeed, in engineering and science the Gaussian distribution is probably the most commonly invoked distribution for random measurements. The Gaussian distribution is also called the Normal distribution. The probability density function (pdf) of a Gaussian random variable (rv) $X$ is parameterized by two parameters, $\theta_1$ and $\theta_2$, which are the location parameter, denoted $\mu$ ($\mu \in \mathbb{R}$), and the (squared) scale parameter, denoted $\sigma^2$ ($\sigma^2 > 0$). The pdf of this Gaussian rv has the form

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} , \quad -\infty < x < \infty$$

When $\mu = 0$ and $\sigma^2 = 1$, $X$ is said to be a standard Gaussian (Normal) rv. A Gaussian random variable with location parameter $\mu$ and scale parameter $\sigma > 0$ can be represented by

$$X = \sigma Z + \mu,$$

where $Z$ is a standard Gaussian rv.

The cumulative density function (cdf) of a standard Gaussian random variable $Z$ is denoted $\mathcal{N}(z)$ and is defined in the conventional manner

$$\mathcal{N}(z) = P(Z \leq z).$$

Equivalently,

$$\mathcal{N}(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv.$$

Using (10) the cdf of a non-standard Gaussian rv $X$ with parameters $\mu$ and $\sigma^2$ can be expressed in terms of the cdf $\mathcal{N}(z)$ of a standard Gaussian rv $Z$:

$$P(X \leq x) = P\left(\frac{(X - \mu)}{\sigma} \leq \frac{(x - \mu)}{\sigma}\right) = \mathcal{N}\left(\frac{x - \mu}{\sigma}\right).$$
The standard Normal cdf \( N(x) \) can be related to the error function or error integral \([1]\): 
\[
\text{erf}(u) = \frac{2}{\sqrt{\pi}} \int_{0}^{u} e^{-t^2} dt, \quad x \geq 0,
\]
through the relation
\[
N(x) = \begin{cases} 
\frac{1}{2}[1 + \text{erf}(|x|/\sqrt{2})] & x \geq 0 \\
\frac{1}{2}[1 - \text{erf}(|x|/\sqrt{2})], & x < 0
\end{cases}.
\]

For positive integer order \( \nu \), the moments of a standard Gaussian random variable \( Z \) are \([34, 13.3]\)
\[
E[Z^\nu] = \begin{cases} 
(\nu - 1)(\nu - 3) \ldots 3 \cdot 1, & \nu \text{ even} \\
0, & \nu \text{ odd}
\end{cases}
\]
where \( E[g(Z)] = \int_{-\infty}^{\infty} g(z)f(z)dz \) denotes statistical expectation of the rv \( g(Z) \) under the pdf \( f(z) \) for rv \( Z \). These moment relations can easily be derived by looking at the coefficients of \( (ju)^k/k! \), \( k = 1, 2, \ldots \) in the power series expansion about \( ju = 0 \) of the characteristic function \( \Phi_Z(u) = E[e^{juZ}] = e^{-u^2/2} \).

In particular, using (10), this implies that the first and second moments of a non-standard Gaussian rv \( X \) are \( E[X] = \mu \) and \( E[X^2] = \mu^2 + \sigma^2 \), respectively. Thus for a Gaussian rv \( X \) we can identify the (ensemble) mean \( E[X] = \mu \) and variance \( \text{var}(X) = E[(X - E[X])^2] = E[X^2] - E^2[X] = \sigma^2 \) as the location and (squared) scale parameters, respectively, of the pdf \( f(x; \mu, \sigma^2) \) of \( X \). In the sequel we will need the following expression for the (non-central) mean deviation \( E[|X + a|] \) for Gaussian \( X \) \([35, 29.6]\):
\[
E[|X + a|] = \sqrt{2} e^{-a^2/2} + a(1 - 2N(-a)). \quad (11)
\]

In referring to rv’s and operations on rv’s in this book the following compact notations are sometimes used:

* “\( X \) is distributed as a Gaussian random variable with mean \( \mu \) and variance \( \sigma^2 \)”

\[
X \sim N(\mu, \sigma^2) \quad (12)
\]

* “\( X \) is equal to a scaled and shifted standard Gaussian random variable”

\[
X = a Z + b \Leftrightarrow X \sim N(b, a^2)
\]
or, in shorthand notation,
\[
X = a N(0, 1) + b \Leftrightarrow X \sim N(b, a^2). \quad (13)
\]

For example, in the following shorthand notation \( X_1, \ldots, X_n \) are independent identically distributed (iid) \( N(0, 1) \) rv’s
\[
\sum_{i=1}^{n} N(0, 1) = \sum_{i=1}^{n} X_i.
\]
Note that the above is an abuse of notation since \( N(0, 1) \) is being used to denote both a Gaussian probability distribution in (12) and a Gaussian random variable in (13). As in all abuses of this
type the ambiguity is resolved from the context: we will never write \( \mathcal{N}(0, 1) \) into an algebraic or other type of equation like the one in (13) when \( \mathcal{N}(0, 1) \) is meant to denote a Gaussian distribution function as opposed to a Gaussian random variable.

Other notational shortcuts are the following. When we write

\[ \mathcal{N}(v) = \alpha \]

we mean that “the cdf of a \( \mathcal{N}(0, 1) \) rv equals \( \alpha \) when evaluated at a point \( v \in \mathbb{R} \).” Likewise

\[ \mathcal{N}^{-1}(\alpha) = v \]

is to be read as “the inverse cdf of a \( \mathcal{N}(0, 1) \) rv equals \( v \) when evaluated at a point \( \alpha \in [0, 1] \).” Finally, by

\[ X \sim \mathcal{N}_n(\mu, R) \]

we mean “\( X \) is distributed as an \( n \)-dimensional Gaussian random vector with mean \( \mu \) and covariance matrix \( R \).”

### 3.1.1 MULTIVARIATE GAUSSIAN DISTRIBUTION

When one passes an i.i.d. Gaussian random sequence through a linear filter the output remains Gaussian but is no longer i.i.d; the filter smooths the input and introduces correlation. Remarkably, if the input to the filter is Gaussian then the output is also Gaussian, i.e., the joint distribution of any \( p \) samples of the output is multivariate Gaussian. To be specific, a random vector \( X = [X_1, \ldots, X_p]^T \) is multivariate Gaussian with mean parameter \( \mu \) and covariance matrix parameter \( \Lambda \) if it has a joint density of the form

\[
f(x) = \frac{1}{(2\pi)^{p/2} |\Lambda|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)\Lambda^{-1}(x - \mu) \right) \quad x \in \mathbb{R}^p.
\]

where \( |\Lambda| \) denotes the the determinant of \( \Lambda \). The \( p \)-variate Gaussian distribution depends on \( p(p + 3)/2 \) parameters, which we can concatenate into a parameter vector \( \theta \) consisting of the \( p \) elements of the mean vector

\[ \mu = [\mu_1, \ldots, \mu_p]^T = \mathbb{E}[X], \]

and the \( p(p + 1)/2 \) distinct parameters of the symmetric positive definite \( p \times p \) covariance matrix

\[ \Lambda = \text{cov}(Z) = \mathbb{E} \left[ (Z - \mu)(Z - \mu)^T \right]. \]

Some useful facts about the multivariate Gaussian random variables are (for derivations of these properties see Morrison [58]):

- **Unimodality and symmetry of the Gaussian density**: The multivariate Gaussian density (14) is unimodal (has a unique maximum) and is symmetric about its mean parameter.

- **Uncorrelated Gaussians are independent**: When the covariance matrix \( \Lambda \) is diagonal, i.e., \( \text{cov}(X_i, X_j) = 0, \ i \neq j \), then the multivariate Gaussian density reduces to a product of univariate densities

\[
f(X) = \prod_{i=1}^{n} f(X_i)
\]
where
\[ f(X_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}(X_i - \mu_i)^2} \]
is the univariate Gaussian density with \( \sigma_i^2 = \text{var}(X_i) \). Thus uncorrelated Gaussian random variables are in fact independent random variables.

- **Marginals of a Gaussian density are Gaussian:** If \( \overline{X} = [X_1, \ldots, X_m]^T \) is multivariate Gaussian then any subset of the elements of \( \overline{X} \) is also Gaussian. In particular \( X_1 \) is univariate Gaussian and \([X_1, X_2]\) is bivariate Gaussian.

- **Linear combination of Gaussian random variables are Gaussian:** Let \( \overline{X} = [X_1, \ldots, X_m]^T \) be a multivariate Gaussian random vector and let \( \mathbf{H} \) be a \( p \times m \) non-random matrix. Then \( \overline{Y} = \mathbf{H} \overline{X} \) is a vector of linear combinations of the \( X_i \)'s. The distribution of \( \overline{Y} \) is multivariate (\( p \)-variate) Gaussian with mean \( \mu_Y = E[\overline{Y}] = \mathbf{H}\mu \) and \( p \times p \) covariance matrix \( \mathbf{A}_Y = \text{cov}(\overline{Y}) = \mathbf{H}\text{cov}(\overline{X})\mathbf{H}^T \).

- **A vector of i.i.d. zero mean Gaussian random variables is invariant to rotation:** Let \( \overline{X} = [X_1, \ldots, X_m]^T \) be vector of zero mean Gaussian random variables with covariance \( \text{cov}(\overline{X}) = \sigma^2 \mathbf{I} \). If \( \mathbf{U} \) is an orthogonal \( m \times m \) matrix, i.e., \( \mathbf{U}^T \mathbf{U} = \mathbf{I} \), then \( \overline{Y} = \mathbf{U}^T \overline{X} \) has the same distribution as \( \overline{X} \).

- **The conditional distribution of a Gaussian given another Gaussian is Gaussian:** Let the vector \( \overline{Z}^T = [\overline{X}^T, \overline{Y}^T] = [X_1, \ldots, X_p, Y_1, \ldots, Y_q]^T \) be multivariate ((\( p + q \))-variate) Gaussian with mean parameters \( \mu_Z = [\mu_X, \mu_Y]^T \) and covariance parameters \( \mathbf{A}_Z \). Then the conditional density \( f_{Y|X}(y|x) \) of \( \overline{Y} \) given \( \overline{X} = \overline{x} \) is multivariate (\( q \)-variate) Gaussian of the form (14) with mean and covariance parameters \( \mu_Y \) and \( \mathbf{A}_Y \) respectively given by (15) and (16) below.

- **Conditional mean of a Gaussian given another Gaussian is linear and conditional covariance is constant:** For the aforementioned multivariate Gaussian vector \( \overline{Z}^T = [\overline{X}^T, \overline{Y}^T]^T \) partition its covariance matrix as follows

\[
\mathbf{A}_Z = \begin{bmatrix} \mathbf{A}_X & \mathbf{A}_{X,Y} \\ \mathbf{A}_{Y,X}^T & \mathbf{A}_Y \end{bmatrix},
\]

where \( \mathbf{A}_X = \text{cov}(\overline{X}) = E[(\overline{X} - \mu_X)(\overline{X} - \mu_X)^T] \) is \( p \times p \), \( \mathbf{A}_Y = \text{cov}(\overline{Y}) = E[(\overline{Y} - \mu_Y)(\overline{Y} - \mu_Y)^T] \) is \( q \times q \), and \( \mathbf{A}_{X,Y} = \text{cov}_G(\overline{X}, \overline{Y}) = E[(\overline{X} - \mu_X)(\overline{Y} - \mu_Y)^T] \) is \( p \times q \). The mean of the multivariate Gaussian conditional density \( f(y|x) \), the conditional mean, is linear in \( x \)

\[
\mu_{Y|X}(\overline{x}) = E[\overline{Y}|\overline{X} = \overline{x}] = \mu_Y + \mathbf{A}_{X,Y}^T \mathbf{A}_X^{-1} (\overline{x} - \mu_X) \tag{15}
\]

and the conditional covariance does not depend on \( \overline{x} \)

\[
\mathbf{A}_{Y|X} = \text{cov}(\overline{Y}|\overline{X} = \overline{x}) = \mathbf{A}_Y - \mathbf{A}_{X,Y}^T \mathbf{A}_X^{-1} \mathbf{A}_{X,Y}. \tag{16}
\]

### 3.1.2 CENTRAL LIMIT THEOREM

One of the most useful results in statistics is the central limit theorem, abbreviated to CLT. This theorem allows one to approximate the distribution of sums of i.i.d. finite variance random variables by a Gaussian distribution. Below we give a general version of the CLT that applies to vector valued r.v.s. For a simple proof of the scalar case see Mood, Graybill and Boes [56]. For proof in the multivariate case see Serfling [Ch. 1][72], which also covers the CLT for the non i.i.d. case.
**Lindeberg-Lévy Central Limit Theorem**: Let \( \{X_i\}_{i=1}^n \) be i.i.d. random vectors in \( \mathbb{R}^p \) with common mean \( E[X_i] = \mu \) and finite positive definite covariance matrix \( \text{cov}(X_i) = \Lambda \). Then as \( n \) goes to infinity the distribution of the random vector \( Z_n = n^{-1/2} \sum_{i=1}^n (X_i - \mu) \) converges to a \( p \)-variate Gaussian distribution with zero mean and covariance \( \Lambda \).

The CLT can also be expressed in terms of the sample mean \( \overline{X} = \overline{X}(n) = n^{-1} \sum_{i=1}^n X_i \): as \( n \to \infty \)

\[
\sqrt{n}(\overline{X}(n) - \mu) \to Z
\]

where \( Z \) is a zero mean Gaussian random vector with covariance matrix \( \Lambda / n \). Thus, for large but finite \( n \), \( \overline{X} \) is approximately Gaussian

\[
\overline{X} \approx (Z/\sqrt{n} + \mu),
\]

with mean \( \mu \) and covariance \( \Lambda / n \). For example, in the case of a scalar \( X_i \), the CLT gives the useful large \( n \) approximation

\[
P(n^{-1} \sum_{i=1}^n X_i \leq y) \approx \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2/n}\right) dy.
\]

The approximation error can be bounded by using the Berry-Esseen Theorems. See Serfling [72] for details.

### 3.1.3 CHI-SQUARE

The (central) **Chi-square** density with \( k \) degrees of freedom (df) is of the form:

\[
f_\theta(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \quad x > 0,
\]

where \( \theta = k \), a positive integer. Here \( \Gamma(u) \) denotes the Gamma function,

\[
\Gamma(u) = \int_0^\infty x^{u-1} e^{-x} dx,
\]

For \( n \) integer valued \( \Gamma(n+1) = n! = n(n-1) \ldots 1 \) and \( \Gamma(n+1/2) = \frac{(2n-1)!!}{(2n)^{n}} \sqrt{\pi} \).

If \( Z_i \sim \mathcal{N}(0, 1) \) are i.i.d., \( i = 1, \ldots, n \), then \( X = \sum_{i=1}^n Z_i^2 \) is distributed as Chi-square with \( n \) degrees of freedom (df). Our shorthand notation for this is

\[
\sum_{i=1}^n [N(0, 1)]^2 = \chi_n.
\]

This characterization of a Chi square r.v. is sometimes called a stochastic representation since it is defined via operations on other r.v.s. The fact that (17) is the density of a sum of squares of independent \( \mathcal{N}(0, 1) \)'s is easily derived. Start with the density function \( f(z) = e^{-z^2/2}/\sqrt{2\pi} \) of a standard Gaussian random variable \( Z \). Using the relation \( (\sqrt{2\pi\sigma})^{-1} \int_{-\infty}^{\infty} e^{-u^2/(2\sigma^2)} du = 1 \), the characteristic function of \( Z^2 \) is simply found as \( \Phi_{Z^2}(u) = E[e^{juZ^2}] = (1 + j2u)^{-1/2} \). Applying the summation-convolution theorem for independent r.v.s \( Y_i, \Phi_{\sum_{i=1}^n Y_i}(u) = \prod \Phi_{Y_i}(u) \), we obtain

\[
\Phi_{\sum_{i=1}^n Z_i^2}(u) = (1 + j2u)^{-n/2}.
\]

Finally, using a table of Fourier transform relations, identify (17) as the inverse fourier transform of \( \Phi_{\sum_{i=1}^n Z_i^2}(u) \).
Some useful properties of the Chi-square random variable are as follows:

* $E[\chi_n] = n$, $\text{var}(\chi_n) = 2n$
* Asymptotic relation for large $n$:
  \[ \chi_n = \sqrt{2n}N(0, 1) + n \]
* $\chi^2$ an exponential r.v. with mean 2, i.e. $X = \chi^2$ is a non-negative r.v. with probability density $f(x) = \frac{1}{2}e^{-x/2}$.
* $\sqrt{\chi^2}$ is a Rayleigh distributed random variable.

### 3.1.4 GAMMA

The Gamma density function is

\[ f_\theta(x) = \frac{x^{\theta-1}e^{-\lambda x}}{\Gamma(\theta)}, \quad x > 0, \]

where $\theta$ denotes the pair of parameters $(\lambda, \theta)$, $\lambda > 0$. Let $\{Y_i\}_{i=1}^n$ be i.i.d. exponentially distributed random variables with mean $1/\lambda$, specifically $Y_i$ has density

\[ f_i(y) = \lambda e^{-\lambda y}, \quad y > 0. \]

Then the sum $X = \sum_{i=1}^n Y_i$ has a Gamma density $f_{(\lambda, n)}$. Other useful properties of a Gamma distributed random variable $X$ with parameters $\theta = (\lambda, \theta)$ include:

* $E[X] = \frac{\theta}{\lambda}$
* $\text{var}(X) = \frac{\theta}{\lambda^2}$
* The Chi-square distribution with $k$ df is a special case of the Gamma distribution obtained by setting Gamma parameters as follows: $\lambda = 1/2$ and $\theta = k/2$.

### 3.1.5 NON-CENTRAL CHI SQUARE

The sum of squares of independent Gaussian r.v.s with unit variances but non-zero means is called a non-central Chi-square r.v. Specifically, if $Z_i \sim \mathcal{N}(\mu_i, 1)$ are independent, $i = 1, \ldots, n$, then $X = \sum_{i=1}^n Z_i^2$ is distributed as non-central Chi-square with $n$ df and non-centrality parameter $\delta = \sum_{i=1}^n \mu_i^2$. In our shorthand we write

\[ \sum_{i=1}^n [\mathcal{N}(0, 1) + \mu_i]^2 = \sum_{i=1}^n [\mathcal{N}(\mu_i, 1)]^2 = \chi_{n, \delta}. \quad (19) \]

The non-central Chi-square density has no simple expression of closed form. There are some useful asymptotic relations, however:

* $E[\chi_{n, \delta}] = n + \delta$, $\text{var}(\chi_{n, \delta}) = 2(n + 2\delta)$
* $\sqrt{\chi_2 + \mu^2 + \mu_2^2}$ is a Rician r.v.
3.1.6 CHI-SQUARE MIXTURE

The distribution of the sum of squares of independent Gaussian r.v.s with zero mean but different variances is not closed form either. However, many statisticians have studied and tabulated the distribution of a weighted sum of squares of i.i.d. standard Gaussian r.v.s  

\[ Z_1, \ldots, Z_n, Z_i \sim \mathcal{N}(0, 1). \]

Specifically, the following has a (central) Chi-square mixture (also known as the Chi-bar square [34]) with \( n \) degrees of freedom and mixture parameter \( \xi = [c_1, \ldots, c_n]^T, c_i \geq 0: \)

\[ \sum_{i=1}^{n} c_i Z_i^2 = \bar{\chi}_{n, \xi} \]

An asymptotic relation of interest to us will be:

* \( E[\bar{\chi}_{n, \xi}] = 1, \quad \text{var}(\bar{\chi}_{n, \xi}) = 2 \sum_{i=1}^{n} \frac{c_i}{\sum_{j} c_j} \)

Furthermore, there is an obvious a special case where the Chi square mixture reduces to a scaled (central) Chi square: \( \bar{\chi}_{n, 1} = \frac{1}{n} \chi_n \) for any \( c \neq 0. \)

3.1.7 STUDENT-T

For \( Z \sim \mathcal{N}(0, 1) \) and \( Y \sim \chi_n \) independent r.v.s the ratio \( X = Z/\sqrt{Y/n} \) is called a Student-t r.v. with \( n \) degrees of freedom, denoted \( T_n \). Or in our shorthand notation:

\[ \frac{\mathcal{N}(0,1)}{\sqrt{\chi_n/n}} = T_n. \]

The density of \( T_n \) is the Student-t density with \( n \) df and has the form

\[ f_\theta(x) = \frac{\Gamma([n+1]/2)}{\Gamma(n/2)} \frac{1}{\sqrt{n\pi}} \frac{1}{(1 + x^2/n)^{(n+1)/2}}, \quad x \in \mathbb{R}, \]

where \( \theta = n \) is a positive integer. Properties of interest to us are:

* \( E[T_n] = 0 \) \( (n > 1) \), \( \text{var}(T_n) = \frac{n}{n-2} \) \( (n > 2) \)

* Asymptotic relation for large \( n: \)

\[ T_n \approx \mathcal{N}(0,1). \]

For \( n = 1 \) the mean of \( T_n \) does not exist and for \( n \leq 2 \) its variance is infinite.

3.1.8 FISHER-F

For \( U \sim \chi_m \) and \( V \sim \chi_n \) independent r.v.s the ratio \( X = (U/m)/(V/n) \) is called a Fisher-F r.v. with \( m, n \) degrees of freedom, or in shorthand:

\[ \frac{\chi_m/m}{\chi_n/n} = F_{m,n}. \]

The Fisher-F density with \( m \) and \( n \) df is defined as

\[ f_\theta(x) = \frac{\Gamma([m+n]/2)}{\Gamma(m/2)\Gamma(n/2)} \left( \frac{m}{n} \right)^{m/2} \frac{x^{(m-2)/2}}{(1 + x/m)^{(m+n)/2}}, \quad x > 0 \]
where \( \theta = [m, n] \) is a pair of positive integers. It should be noted that moments \( E[X^k] \) of order greater than \( k = n/2 \) do not exist. A useful asymptotic relation for \( n \) large and \( n \gg m \) is

\[
\mathcal{F}_{m,n} \approx \chi_m.
\]

### 3.1.9 CAUCHY

The ratio of independent \( \mathcal{N}(0, 1) \) r.v.’s \( U \) and \( V \) is called a standard Cauchy r.v.

\[
X = U/V \sim \mathcal{C}(0, 1).
\]

It’s density has the form

\[
f(x) = \frac{1}{\pi} \frac{1}{1 + x^2} \quad x \in \mathbb{R}
\]

If \( \theta = [\mu, \sigma] \) are location and scale parameters \((\sigma > 0)\) \( f_\theta(x) = f((x - \mu)/\sigma) \) is a translated and scaled version of the standard Cauchy density denoted \( \mathcal{C}(\mu, \sigma^2) \). Some properties of note:

1. the Cauchy distribution has no moments of any (positive) integer order; and
2. the Cauchy distribution is the same as a Student-t distribution with 1 d.f.

### 3.1.10 BETA

For \( U \sim \chi_m \) and \( V \sim \chi_n \) independent Chi-square r.v.s with \( m \) and \( n \) df, respectively, the ratio \( X = U/(U + V) \) has a Beta distribution, or in shorthand

\[
\frac{\chi_m}{\chi_m + \chi_n} = \mathcal{B}(m/2, n/2)
\]

where \( \mathcal{B}(p,q) \) is a r.v. with Beta density having parameters \( \theta = [p, q] \). The Beta density has the form

\[
f_\theta(x) = \frac{1}{\beta_{r,t}} x^{r-1}(1-x)^{t-1}, \quad x \in [0, 1]
\]

where \( \theta = [r, t] \) and \( r, t > 0 \). Here \( \beta_{r,t} \) is the Beta function:

\[
\beta_{r,t} = \int_0^1 x^{r-1}(1-x)^{t-1} dx = \frac{\Gamma(r)\Gamma(t)}{\Gamma(r+t)}
\]

Some useful properties:

* The special case of \( m = n = 1 \) gives rise to \( X \) an arcsin distributed r.v.
* \( E_\theta[\mathcal{B}(p,q)] = p/(p+q) \)
* \( \text{var}_\theta(\mathcal{B}(p,q)) = pq/((p + q + 1)(p + q)^2) \)

### 3.2 REPRODUCING DISTRIBUTIONS

A random variable \( X \) is said to have a reproducing distribution if the sum of two independent realizations, say \( X_1 \) and \( X_2 \), of \( X \) have the same distribution, possibly with different parameter values, as \( X \). A Gaussian r.v. has a reproducing distribution:

\[
\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2),
\]
which follows from the fact that the convolution of two Gaussian density functions is a Gaussian density function \[56\]. Noting the stochastic representations (18) and (19) of the Chi square and non-central Chi square distributions, respectively, it is obvious that they are reproducing distributions:

* \(\chi_n + \chi_m = \chi_{m+n}\), if \(\chi_m, \chi_n\) are independent.
* \(\chi_{m,\delta_1} + \chi_{n,\delta_2} = \chi_{m+n,\delta_1+\delta_2}\), if \(\chi_{m,\delta_1}, \chi_{n,\delta_2}\) are independent.

The Chi square mixture, Fisher-F, and Student-t are not reproducing densities.

### 3.3 FISHER-COCHRAN THEOREM

This result gives a very useful tool for finding the distribution of quadratic forms of Gaussian random variables. A more general result that covers the joint distribution of quadratic forms is given in [66].

**Theorem 1** Let \(\mathbf{X} = [X_1, \ldots, X_n]^T\) be a vector of i.i.d. \(\mathcal{N}(0,1)\) r.v.'s and let \(\mathbf{A}\) be a symmetric idempotent matrix (\(\mathbf{AA} = \mathbf{A}\)) of rank \(p\). Then

\[\mathbf{X}^T \mathbf{A} \mathbf{X} = \chi_p\]

A simple proof is given below.

**Proof:** Let \(\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\) be the eigendecomposition of \(\mathbf{A}\). Then

* All eigenvalues \(\lambda_i\) of \(\mathbf{A}\) are either 0 or 1

\[
\mathbf{AA} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T
\]

and therefore

\[
\mathbf{X}^T \mathbf{A} \mathbf{X} = \mathbf{X}^T \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \mathbf{X} \\
= \sum_{i=1}^{n} \lambda_i Z_i^2 = \sum_{i=1}^{p} [\mathcal{N}(0,1)]^2
\]

\(\diamondsuit\)

### 3.4 SAMPLE MEAN AND SAMPLE VARIANCE

Let \(X_i\)'s be i.i.d. \(\mathcal{N}(\mu, \sigma^2)\) r.v.'s. The sample mean and sample variance respectively approximate the location \(\mu\) and spread \(\sigma\) of the population.

* Sample mean: \(\bar{X} = n^{-1} \sum_{i=1}^{n} X_i\)
* Sample variance: \(s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2\)
In the Gaussian case the joint distribution of the sample mean and variance can be specified.

(1). $\overline{X} = \mathcal{N}(\mu, \sigma^2/n)$

(2). $s^2 = \frac{\sigma^2}{n-1} \chi_{n-1}$

(3). $\overline{X}$ and $s^2$ are independent rv’s.

These results imply that a weighted ratio of sample mean and sample variance is distributed as Student t.

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} = T_{n-1}. \quad (21)$$

Proof of assertions (2) and (3): In view of the representation (13), it suffices consider the the case of a standard Gaussian sample: $\mu = 0$ and $\sigma = 1$.

First we show that the sample mean and the sample variance are independent random variables. Define the vector of random variables $Y = [Y_1, \ldots, Y_n]^T$ as follows. First define

$$Y_1 = \sqrt{n}X = h_1^T X,$$

where

$$h_1 = [1/\sqrt{n}, \ldots, 1/\sqrt{n}]^T.$$

Note that $h_1$ has unit norm. Next apply the Gramm-Schmidt orthonormalization procedure of Sec. 2.3.6 to complete the basis with respect to $h_1$. This generates $n-1$ vectors $h_2, \ldots, h_n$ that are orthonormal, mutually orthogonal, and orthogonal to $h_1$. The random vector $Y$ is now defined as

$$Y = H^T X$$

where $H = [h_1, \ldots, h_n]$ is an $n \times n$ orthogonal matrix.

Since, $X = HY$, the orthogonality of $H$ implies the following properties

1. The $Y_i$’s are zero mean unit variance independent Gaussian random variables: $Y \sim \mathcal{N}_n(0, I)$

2. $Y^T Y = X^T X$

As $Y_1 = \sqrt{n}X$ Property 1 implies that $\overline{X}$ is independent of $Y_2, \ldots, Y_n$. Furthermore, using the equivalence:

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2,$$

Property 2 and the definition of $Y_1$ imply that

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = Y_2^2 + \cdots + Y_n^2, \quad (20)$$

that is, the sample variance is only a function of $Y_2, \ldots, Y_n$ and is therefore independent of $Y_1 = \overline{X}$ the sample mean.

Furthermore, as $Y_2, \ldots, Y_n$ are independent $\mathcal{N}(0, 1)$ random variables, the representation (20) implies that the (normalized) sample variance has a Chi-square distribution with $n - 1$ degrees of freedom.

This completes the proof of assertions (2) and (3).
The Chi-square property in assertion (3) can also be shown directly using the Fisher-Cochran theorem (Thm. 1). Note that the normalized sample variance on the extreme left of the equalities (20) can be expressed as a quadratic form

\[ [X - 1\overline{X}]^T[X - 1\overline{X}] = X^T [I - 11^T \frac{1}{n}] [I - 11^T \frac{1}{n}] X^T \]

\[ = X^T [I - 11^T \frac{1}{n}] X \]

where \( 1 = [1, \ldots, 1]^T \). Observe: since rank\([I - 11^T \frac{1}{n}] = n - 1\), we have that \( [X - 1\overline{X}]^T[X - 1\overline{X}] = (n - 1) s^2 \) is \( \chi_{n-1} \).

### 3.5 SUFFICIENT STATISTICS

Many detection/estimation/classification problems have the following common structure. A continuous time waveform \( \{x(t) : t \in \mathbb{R}\} \) is measured at \( n \) time instants \( t_1, \ldots, t_n \) producing the vector

\[ x = [x_1, \ldots, x_n]^T, \]

where \( x_i = x(t_i) \). The vector \( x \) is modelled as a realization of a random vector \( X \) with a joint distribution which is of known form but depends on a handful \( (p) \) of unknown parameters \( \theta = [\theta_1, \ldots, \theta_p]^T \).

More concisely:

* \( X = [X_1, \ldots, X_n]^T, \) \( X_i = X(t_i) \), is a vector of random measurements or observations taken over the course of the experiment
* \( \mathcal{X} \) is sample or measurement space of realizations \( x \) of \( X \)
* \( \mathcal{B} \) is the event space induced by \( X \), e.g., the Borel subsets of \( \mathbb{R}^n \)
* \( \theta \in \Theta \) is an unknown parameter vector of interest
* \( \Theta \) is parameter space for the experiment
* \( P_{\theta} \) is a probability measure on \( \mathcal{B} \) for given \( \theta \). \( \{P_{\theta}\}_{\theta \in \Theta} \) is called the statistical model for the experiment.

The probability model induces the joint cumulative distribution function (j.c.d.f.) associated with \( X \)

\[ F_X(x; \theta) = P_{\theta}(X_1 \leq x_1, \ldots, X_n \leq x_n), \]

which is assumed to be known for any \( \theta \in \Theta \). When \( X \) is a continuous random variable the j.c.d.f. is specified by the joint probability density function (j.p.d.f.) that we will write in several different ways, depending on the context: \( f_{\theta}(x) \) or \( f(x; \theta) \), or, when we need to explicitly call out the r.v. \( X \), \( f_X(x; \theta) \). We will denote by \( E_{\theta}[Z] \) the statistical expectation of a random variable \( Z \) with respect to the j.p.d.f. \( f_Z(z; \theta) \)

\[ E_{\theta}[Z] = \int z f_{\theta}(z; \theta) \, dz. \]

The family of functions \( \{f(x; \theta)\}_{x \in \mathcal{X}, \theta \in \Theta} \) then defines the statistical model for the experiment.
The general objective of statistical inference can now be stated. Given a realization \( x \) of \( X \) infer properties of \( \theta \) knowing only the parametric form of the statistical model. Thus we will want to come up with a function, called an inference function, which maps \( X \) to subsets of the parameter space, e.g., an estimator, classifier, or detector for \( \theta \). As we will see later there are many ways to design inference functions but a more fundamental question is: are there any general properties that good inference functions should have? One such property is that the inference function only need depend on the \( n \)-dimensional data vector \( X \) through a lower dimensional version of the data called a sufficient statistic.

### 3.5.1 SUFFICIENT STATISTICS AND THE REDUCTION RATIO

First we define a statistic as any function \( T = T(X) \) of the data (actually, for \( T \) to be a valid random variable derived from \( X \) it must be a measurable function, but this theoretical technicality is beyond our scope here).

There is a nice interpretation of a statistic in terms of its memory storage requirements. Assume that you have a special computer that can store any one of the time samples in \( X = [X_1, \ldots, X_n] \), \( X_k = X(t_k) \) say, in a "byte" of storage space and the time stamp \( t_k \) in another "byte" of storage space. Any non-invertible function \( T \), e.g., which maps \( \mathbb{R}^n \) to a lower dimensional space \( \mathbb{R}^m \), can be viewed as a dimensionality reduction on the data sample. We can quantify the amount of reduction achieved by \( T \) by defining the reduction ratio (RR):

\[
RR = \frac{\text{# bytes of storage required for } T(X)}{\text{# bytes of storage required for } X}
\]

This ratio is a measure of the amount of data compression induced by a specific transformation \( T \). The number of bytes required to store \( X \) with its time stamps is:

\[
\text{# bytes}\{X\} = \text{# bytes}\{X_1, \ldots, X_n\}^T = \text{# bytes}\{\text{timestamps}\} + \text{# bytes}\{\text{values}\} = 2n
\]

Consider the following examples:

<table>
<thead>
<tr>
<th>Statistic used</th>
<th>Meaning in plain English</th>
<th>Reduction ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T(X) = [X_1, \ldots, X_n]^T )</td>
<td>entire data sample</td>
<td>RR = 1</td>
</tr>
<tr>
<td>( T(X) = [X_{(1)}, \ldots, X_{(n)}]^T )</td>
<td>rank ordered sample</td>
<td>RR = 1/2</td>
</tr>
<tr>
<td>( T(X) = \bar{X} )</td>
<td>sample mean</td>
<td>RR = 1/(2n)</td>
</tr>
<tr>
<td>( T(X) = [\bar{X}, s^2]^T )</td>
<td>sample mean and variance</td>
<td>RR = 1/n</td>
</tr>
</tbody>
</table>

A natural question is: what is the maximal reduction ratio one can get away with without loss of information about \( \theta \)? The answer is: the ratio obtained by compression to a quantity called a minimal sufficient statistic. But we are getting ahead of ourselves. We first need to define a plain old sufficient statistic.
3.5.2 DEFINITION OF SUFFICIENCY

Here is a warm up before making a precise definition of sufficiency. \( T = T(X) \) is a sufficient statistic (SS) for a parameter \( \theta \) if it captures all the information in the data sample useful for inferring the value of \( \theta \). To put it another way: once you have computed a sufficient statistic you can store it and throw away the original sample since keeping it around would not add any useful information.

More concretely, let \( X \) have a cumulative distribution function (CDF) \( F_X(x; \theta) \) depending on \( \theta \). A statistic \( T = T(X) \) is said to be sufficient for \( \theta \) if the conditional CDF of \( X \) given \( T = t \) is not a function of \( \theta \), i.e.,

\[
F_X|T(x|T = t; \theta) = G(x, t),
\]

where \( G \) is a function that does not depend on \( \theta \).

Specializing to a discrete valued \( X \) with probability mass function \( p_X(x) = P(X = x) \), a statistic \( T = T(X) \) is sufficient for \( \theta \) if

\[
P_X(X = x|T = t) = G(x, t).
\]

For a continuous r.v. \( X \) with pdf \( f_X(x; \theta) \), the condition (21) for \( T \) to be a sufficient statistic (SS) becomes:

\[
f_X|T(x|t; \theta) = G(x, t).
\]

Sometimes the only sufficient statistics are vector statistics, e.g. \( T(X) = [T_1(X), \ldots, T_K(X)]^T \).

The definition (21) is often difficult to use since it involves derivation of the conditional distribution of \( X \) given \( T \). When the random variable \( X \) is discrete or continuous a simpler way to verify sufficiency is through the Fisher factorization (FF) property [66]

**Fisher factorization (FF):** \( T = T(X) \) is a sufficient statistic for \( \theta \) if the probability density \( f_X(x; \theta) \) of \( X \) has the representation

\[
f_X(x; \theta) = g(T, \theta) h(x),
\]

for some non-negative functions \( g \) and \( h \). The FF can be taken as the operational definition of a sufficient statistic \( T \). An important implication of the Fisher Factorization is that when the density function of a sample \( X \) satisfies (24) then the density \( f_T(t; \theta) \) of the sufficient statistic \( T \) is equal to \( g(t, \theta) \) up to a \( \theta \)-independent constant \( q(t) \) (see exercises at end of this chapter):

\[
f_T(t; \theta) = g(t, \theta)q(t).
\]

Examples of sufficient statistics:

**Example 1** *Entire sample*

\( X = [X_1, \ldots, X_n]^T \) is sufficient but not very interesting

**Example 2** *Rank ordered sample*
$X_{(1)}, \ldots, X_{(n)}$ is sufficient when $X_i$’s i.i.d.

**Proof:** Since $X_i$’s are i.i.d., the joint pdf is

$$f_{\theta}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{\theta}(x_i) = \prod_{i=1}^{n} f_{\theta}(x_{(i)}).$$

Hence sufficiency of the rank ordered sample $X_{(1)}, \ldots, X_{(n)}$ follows from Fisher factorization.

**Example 3** Binary likelihood ratios

Let $\theta$ take on only two possible values $\theta_0$ and $\theta_1$, e.g., a bit taking on the values “0” or “1” in a communication link. Then, as $f(x; \theta)$ can only be $f(x; \theta_0)$ or $f(x; \theta_1)$, we can reindex the pdf as $f(x; \theta)$ with the scalar parameter $\theta \in \Theta = \{0, 1\}$. This gives the binary decision problem: “decide between $\theta = 0$ versus $\theta = 1$.” If it exists, i.e. it is finite for all values of $X$, the “likelihood ratio” $\Lambda(X) = f_1(X)/f_0(X)$ is sufficient for $\theta$, where $f_1(x) \overset{\text{def}}{=} f(x; 1)$ and $f_0(x) \overset{\text{def}}{=} f(x; 0)$.

**Proof:** Express $f_{\theta}(X)$ as function of $\theta, f_0, f_1$, factor out $f_0$, identify $\theta$, and invoke FF

$$f_{\theta}(X) = \theta f_1(X) + (1 - \theta) f_0(X)$$

$$= \left( \frac{\theta \Lambda(X) + (1 - \theta)}{g(T, \theta)} \right) \frac{f_0(X)}{h(X)}.$$

Therefore to discriminate between two values $\theta_1$ and $\theta_2$ of a parameter vector $\theta$ we can throw away all data except for the scalar sufficient statistic $T = \Lambda(X)$

**Example 4** Discrete likelihood ratios

Let $\Theta = \{\theta_1, \ldots, \theta_p\}$ and assume that the vector of $p - 1$ likelihood ratios

$$T(X) = \begin{bmatrix} f_{\theta_1}(X) \\ f_{\theta_p}(X) \\ \vdots \\ f_{\theta_{p-1}}(X) \\ f_{\theta_p}(X) \end{bmatrix}^T = [\Lambda_1(X), \ldots, \Lambda_{p-1}(X)]^T$$

is finite for all $X$. Then this vector is sufficient for $\theta$. An equivalent way to express this vector is as the sequence $\{\Lambda_{\theta}(X)\}_{\theta \in \Theta} = \Lambda_1(X), \ldots, \Lambda_{p-1}(X)$, and this is called the likelihood trajectory over $\theta$.

**Proof**

Define the $p - 1$ element selector vector $u_\theta = e_k$ when $\theta = \theta_k, \ k = 1, \ldots, p - 1$ (recall that $e_k = [0, \ldots, 0, 1, 0, \ldots 0]^T$ is the $k$-th column of the $(p - 1) \times (p - 1)$ identity matrix). Now for any $\theta \in \Theta$ we can represent the j.p.d.f. as

$$f_{\theta}(x) = \frac{u_\theta^T \overbrace{T(X)}^{g(T, \theta)} \overbrace{f_{\theta_p}(X)}^{h(x)}},$$

which establishes sufficiency by the FF.  \(\diamond\)
Example 5 Likelihood ratio trajectory

When $\Theta$ is a set of scalar parameters $\theta$ the likelihood ratio trajectory over $\Theta$

$$\Lambda(X) = \left\{ \frac{f_\theta(X)}{f_{\theta_0}(X)} \right\}_{\theta \in \Theta},$$

is sufficient for $\theta$. Here $\theta_0$ is an arbitrary reference point in $\Theta$ for which the trajectory is finite for all $X$. When $\theta$ is not a scalar (25) becomes a likelihood ratio surface, which is also a sufficient statistic.

3.5.3 MINIMAL SUFFICIENCY

What is the maximum possible amount of reduction one can apply to the data sample without losing information concerning how the model depends on $\theta$? The answer to this question lies in the notion of a minimal sufficient statistic. Such statistics cannot be reduced any further without loss in information. In other words, any other sufficient statistic can be reduced down to a minimal sufficient statistic without information loss. Since reduction of a statistic is accomplished by applying a functional transformation we have the formal definition.

**Definition:** $T_{\text{min}}$ is a minimal sufficient statistic if it can be obtained from any other sufficient statistic $T$ by applying a functional transformation to $T$. Equivalently, if $T$ is any sufficient statistic there exists a function $q$ such that $T_{\text{min}} = q(T)$.

Minimal sufficient statistics are not unique: if $T_{\text{min}}$ is minimal sufficient $h(T_{\text{min}})$ is also minimal sufficient if $h$ is any invertible function. Minimal sufficient statistics can be found in a variety of ways [56, 9, 48]. One way is to find a complete sufficient statistic; under broad conditions this statistic will also be minimal [48]. A sufficient statistic $T$ is complete if

$$E[ \varphi(T) ] = 0, \text{ for all } \theta \in \Theta$$

implies that the function $\varphi$ is identically zero, i.e., $\varphi(t) = 0$ for all values of $t$.

To see that a completeness implies minimality we can adapt the proof of Scharf in [69]. Let $M$ be a minimal sufficient statistic and let $C$ be complete sufficient statistic. As $M$ is minimal it is a function of $C$. Therefore $\varphi(C) \overset{\text{def}}{=} C - E_\theta[C|M]$ is a function of $C$ since the conditional expectation $E_\theta[X|M]$ is a function of $M$. Since, obviously, $E_\theta[\varphi(C)] = 0$ for all $\theta$ and $C$ is complete, $C = E_\theta[C|M]$ for all $\theta$. Thus $C$ is minimal since it is a function of $M$ which is a function of any other sufficient statistic. In other words, $C$ inherits minimality from $M$.

Another way to find a minimal sufficient statistic is through reduction of the data to the likelihood ratio surface.

As in Example 5, assume that there exists a reference point $\theta_0 \in \Theta$ such that the following likelihood-ratio function is finite for all $x \in X$ and all $\theta \in \Theta$

$$\Lambda_\theta(x) = \frac{f_\theta(x)}{f_{\theta_0}(x)}$$

For given $x$ let $\Lambda(x)$ denote the set of likelihood ratios (a likelihood ratio trajectory or surface)

$$\Lambda(x) = \{ \Lambda_\theta(x) \}_{\theta \in \Theta}.$$
Definition 1: We say that a (θ-independent) function of $x$, denoted $\tau = \tau(x)$, indexes the likelihood ratios $\Lambda$ when both

1. $\Lambda(x) = \Lambda(\tau)$, i.e., $\Lambda$ only depends on $x$ through $\tau = \tau(x)$.
2. $\Lambda(\tau) = \Lambda(\tau')$ implies $\tau = \tau'$, i.e., the mapping $\tau \to \Lambda(\tau)$ is invertible.

Condition 1 is an equivalent way of stating that $\tau(X)$ is a sufficient statistic for $\theta$.

Theorem: If $\tau = \tau(x)$ indexes the likelihood ratios $\Lambda(x)$ then $T_{\text{min}} = \tau(X)$ is minimally sufficient for $\theta$.

Proof:
We prove this only for the case that $X$ is a continuous r.v. First, condition 1 in Definition 1 implies that $\tau = \tau(X)$ is a sufficient statistic. To see this use FF and the definition of the likelihood ratios to see that $\Lambda(x) = \Lambda(x')$ implies: $f_\theta(x) = \Lambda_\theta(\tau)f_\theta(x) = g(\tau; \theta)h(x)$. Second, let $T$ be any sufficient statistic. Then, again by FF, $f_\theta(x) = g(T, \theta)h(x)$ and thus

$$\Lambda(\tau) = \left\{ \frac{f_\theta(X)}{f_\theta(x)} \right\}_{\theta \in \Theta} = \left\{ \frac{g(T, \theta)}{g(T, \theta, \theta)} \right\}_{\theta \in \Theta}.$$

so we conclude that $\Lambda(\tau)$ is a function of $T$. But by condition 2 in Definition 1 the mapping $\tau \to \Lambda(\tau)$ is invertible and thus $\tau$ is itself a function of $T$.

Another important concept in practical applications is that of finite dimensionality of a sufficient statistic.

Definition: a sufficient statistic $T(X)$ is said to be finite dimensional if its dimension is not a function of the number of data samples $n$.

Frequently, but not always (see Cauchy example below), minimal sufficient statistics are finite dimensional.

Example 6: Minimal sufficient statistic for mean of Gaussian density.

Assume $X \sim \mathcal{N}(\mu, \sigma^2)$ where $\sigma^2$ is known. Find a minimal sufficient statistic for $\theta = \mu$ given the iid sample $X = [X_1, \ldots, X_n]^T$.

Solution: the j.p.d.f. is

$$f_\theta(x) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

$$= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} (\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2)}$$

$$= e^{-\frac{n\mu^2}{2\sigma^2}} e^{\mu/\sigma^2 \sum_{i=1}^n x_i} g(T(x)) h(x)$$

$$= e^{-\frac{n\mu^2}{2\sigma^2}} e^{\mu/\sigma^2 \sum_{i=1}^n x_i} \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-1/(2\sigma^2) \sum_{i=1}^n x_i^2}$$

$$= e^{-\frac{n\mu^2}{2\sigma^2}} e^{\mu/\sigma^2 \sum_{i=1}^n x_i}$$

where $T(x)$ is the sufficient statistic for $\mu$.
Thus by FF

\[ T = \sum_{i=1}^{n} X_i \]

is a sufficient statistic for \( \mu \). Furthermore, as \( q(T) = n^{-1}T \) is a 1-1 function of \( T \)

\[ S = \bar{X} \]

is an equivalent sufficient statistic.

Next we show that the sample mean is in fact minimal sufficient by showing that it indexes the likelihood ratio trajectory \( \Lambda(x) = \{ \Lambda_\theta(x) \}_{\theta \in \Theta} \), with \( \theta = \mu, \Theta = \mathbb{R} \). Select the reference point \( \theta_o = \mu_o = 0 \) to obtain:

\[ \Lambda_\mu(x) = \frac{f_\mu(x)}{f_0(x)} = \exp \left( \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{1}{2} n \mu^2 / \sigma^2 \right). \]

Identifying \( \tau = \sum_{i=1}^{n} x_i \), condition 1 in Definition 1 is obviously satisfied since \( \Lambda_\mu(x) = \Lambda_\mu(\sum x_i) \) (we already knew this since we showed that \( \sum_{i=1}^{n} X_i \) was a sufficient statistic). Condition 2 in Definition 1 follows since \( \Lambda_\mu(\sum x_i) \) is an invertible function of \( \sum x_i \) for any non-zero value of \( \mu \) (summation limits omitted for clarity). Therefore the sample mean indexes the trajectories, and is minimal sufficient.

Example 7 Minimal sufficient statistics for mean and variance of Gaussian density.

Assume \( X \sim \mathcal{N}(\mu, \sigma^2) \) where both \( \mu \) and \( \sigma^2 \) are unknown. Find a minimal sufficient statistic for \( \theta = [\mu, \sigma^2]^T \) given the iid sample \( X = [X_1, \ldots, X_n]^T \).

Solution:

\[
\begin{align*}
 f_\theta(x) &= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2} \\
 &= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 \right)} \\
 &= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{n\mu^2}{2\sigma^2}} e^{\sum_{i=1}^{n} x_i} \left[ \sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2 \right]^T g(T, \theta) \left[ \frac{1}{k(x)} \right]
\end{align*}
\]

Thus

\[ T = \begin{bmatrix}
\sum_{i=1}^{n} X_i, \\
\sum_{i=1}^{n} X_i^2
\end{bmatrix}
\]
is a (jointly) sufficient statistic for \( \mu, \sigma^2 \). Furthermore, as \( q(T) = [n^{-1}T_1, (n-1)^{-1}(T_2 - T_1^2)] \) is a 1-1 function of \( T \) \( (T = [T_1, T_2]^T) \)

\[
S = [\bar{X}, s^2]
\]

is an equivalent sufficient statistic.

Similarly to Example 6, we can show minimal sufficiency of this statistic by showing that it indexes the likelihood ratio surface \( \{ \Lambda_\theta(X) \}_{\theta \in \Theta} \), with \( \theta = [\mu, \sigma^2] \), \( \Theta = \mathbb{R} \times \mathbb{R}^+ \). Arbitrarily select the reference point \( \theta_0 = [\mu_0, \sigma_0^2] = [0, 1] \) to obtain:

\[
\Lambda_\theta(x) = \frac{f_\theta(x)}{f_{\theta_0}(x)} = \left( \frac{\sigma_0}{\sigma} \right)^n e^{-n\mu^2/(2\sigma^2)} \ e^{[\mu/\sigma^2, -\delta/2][\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2]^T},
\]

where \( \delta = \frac{\sigma_0^2 - \sigma^2}{\sigma_0^2} \). Identifying \( \tau = [\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2] \), again condition 1 in Definition 1 is obviously satisfied. Condition 2 in Definition 1 requires a bit more work. While \( \Lambda_\theta(\tau) \) is no longer an invertible function of \( \tau \) for for any single value of \( \theta = [\mu, \sigma^2] \), we can find two values \( \theta \in \{\theta_1, \theta_2\} \) in \( \Theta \) for which the vector function \( [\Lambda_{\theta_1}(\tau), \Lambda_{\theta_2}(\tau)] \) of \( \tau \) is invertible in \( \tau \). Since this vector is specified by \( \Lambda(x) \), this will imply that \( \tau \) indexes the likelihood ratios.

To construct this invertible relation denote by \( \Lambda = [\lambda_1, \lambda_2]^T \) an observed pair of samples \( [\Lambda_{\theta_1}(\tau), \Lambda_{\theta_2}(\tau)]^T \) of the surface \( \Lambda(x) \). Now consider the problem of determining \( \tau \) from the equation \( \Lambda = [\Lambda_{\theta_1}(\tau), \Lambda_{\theta_2}(\tau)]^T \).

Taking the log of both sides and rearranging some terms, we see that this is equivalent to a 2 \times 2 linear system of equations of the form \( \Lambda' = A \tau \), where \( A \) is a matrix involving \( \theta_0, \theta_1, \theta_2 \) and \( \Lambda' \) is a linear function of \( \ln \Lambda \). You can verify that with the selection of \( \theta_0 = [0, 1], \theta_1 = [1, 1], \theta_2 = [0, 1/2] \) we obtain \( \delta = 0 \) or 1 for \( \theta = \theta_1 \) or \( \theta_2 \), respectively, and \( A = \text{diag}(1, -1/2) \), an invertible matrix.

We therefore conclude that the vector [sample mean, sample variance] indexes the trajectories, and this vector is therefore minimal sufficient.

**Example 8** Minimal sufficient statistic for the location of a Cauchy distribution

Assume that \( X_i \sim f(x; \theta) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2} \) and, as usual, \( X = [X_1, \ldots, X_n]^T \) is an i.i.d. sample. Then

\[
f(x; \theta) = \prod_{i=1}^n \frac{1}{\pi} \frac{1}{1+(x_i - \theta)^2} = \frac{1}{\pi^n} \prod_{i=1}^n \frac{1}{1+(x_i - \theta)^2}.
\]

Here we encounter a difficulty: the denominator is a 2n-degree polynomial in \( \theta \) whose coefficients cannot be determined without specifying the entire set of all possible cross products \( x_{i_1} \cdots x_{i_p}, \)

\( p = 1, 2, \ldots, n, \) of the \( x_i \)'s. Since this requires specifying the entire set of sample values there is no finite dimensional sufficient statistic. However, each of these cross products is independent of the ordering of its factors so the ordered statistic \( [X_{(1)}, \ldots, X_{(n)}]^T \) is minimally sufficient.

### 3.6 Establishing That a Statistic Is Not Sufficient

One can show that a statistic \( U \) is not sufficient for a parameter \( \theta \) by establishing that the conditional distribution of the sample \( X = [X_1, \ldots, X_n] \) given \( U \) is a function of \( \theta \). For example, for an i.i.d. sample from a Gaussian distribution with unknown mean \( \mu \) and known variance \( \sigma^2 \), we have seen in Example 6 that the sample mean of \( X \) is sufficient, and is in fact minimally sufficient, for estimation of \( \mu \). However, other functions of the samples are not generally sufficient.
Example 9 The $l_p$ norm of the samples is not a sufficient statistic for the mean of a Gaussian density.

The $l_p$ norm of $X$, defined as $U = \|X\|_p = (\sum_{i=1}^n |X_i|^p)^{1/p}$, is not a sufficient statistic for the mean $\mu$ when $X$ is a Gaussian sample. To show this we specialize to the case $n = 1$, for which $U = |X_1|$, and establish that the conditional CDF $F_{X_1|U}(x|u; \theta)$ is a function of $\theta = \mu$. The distribution of $X_1$ given $U = u$ concentrates its mass at two points $u$ and $-u$. This distribution can be represented as a density with dirac delta functions at these points:

$$f_{X_1|U}(x|u; \theta) = \frac{f_{X_1}(x; \theta)}{f_{X_1}(u; \theta) + f_{X_1}(-u; \theta)} \delta(|x| - u) = \frac{e^{-(x-\theta)^2/(2\sigma^2)}}{e^{-(x-\theta)^2/(2\sigma^2)} + e^{-(x+\theta)^2/(2\sigma^2)}} \delta(|x| - u),$$

which is a function of $\theta$. Thus the CDF is also a function of $\theta$ and we conclude that the absolute value of the sample mean is not a sufficient statistic for the mean of an i.i.d. Gaussian sample.

### 3.6.1 EXPONENTIAL FAMILY OF DISTRIBUTIONS

Let $\theta = [\theta_1, \ldots, \theta_p]^T$ take values in some parameter space $\Theta$. The distribution $f_\theta$ of a random variable $X$ is a member of the $p$-parameter exponential family if for all $\theta \in \Theta$

$$f_\theta(x) = a(\theta) b(x) e^{\theta^T \zeta(x)}, \quad -\infty < x < \infty$$

(26)

for some scalar functions $a, b$ and some $p$-element vector functions $\zeta, \xi$. A similar definition of exponential family holds for vector valued random variables $X$, see Bickel and Doksum [9, Ch. 2]. Note that for any $f_\theta$ in the exponential family its support set $\{x : f_\theta(x) > 0\}$ does not depend on $\theta$. Note that, according to our definition, for $f_\theta$ to be a member of the $p$-parameter exponential family the dimension of the vectors $\zeta(\theta)$ and $\xi(x)$ must be exactly $p$. This is to guarantee that the sufficient statistic has the same dimension as the parameter vector $\theta$. While our definition is the most standard [47, 56, 9], some other books, e.g., [64], allow the dimension of the sufficient statistic to be different from $p$. However, by allowing this we lose some important properties of exponential families [9].

The parameterization of an exponential family of distributions is not unique. In other words, the exponential family is invariant to changes in parameterization. For example, if $f_\theta$, $\theta > 0$, is a member of an exponential family then if one defines $\alpha = 1/\theta$ and $g_\alpha = f_{1/\alpha}$ then $g_\alpha$, $\alpha > 0$, is also in the exponential family, but possibly with different functions $a(\cdot), b(\cdot), c(\cdot)$ and $t(\cdot)$. More generally, if $f_\theta(x)$ is a member of the $p$-dimensional exponential family then transformation of the parameters by any invertible function of $\theta$ preserves membership in the exponential family.

To illustrate, let’s say that the user redefined the parameters by the mapping $\zeta : \theta \rightarrow \eta$ defined by the invertible transformation $\zeta(\theta) = \eta$. Then, using (26), $f_\theta$ would be replaced by

$$f_\eta(x) = a(\eta) b(x) e^{\eta^T \zeta(x)}, \quad -\infty < x < \infty,$$

(27)

where $\eta(a) = a(\zeta^{-1}(\eta))$. Thus $f_\eta$ remains in the exponential family. When expressed in the form (27), the exponential family density $f_\theta$ is said to be in canonical form with natural parameterization $\eta$. Under the natural parameterization the mean and covariance matrix of the sufficient statistic $T = \xi(X)$ are given by (assuming differentiable $a$)

$$E_\eta[T] = \nabla \ln a(\eta),$$
and
\[ \text{cov}_a[T] = \nabla^2 \ln a(\eta). \]

For a proof of these relations see Bickel and Doksum [9].

Another parameterization of an exponential family of densities is the mean value parameterization. In this parameterization, the functions \(t(\cdot), a(\cdot), b(\cdot)\) and \(c(\cdot)\) in (26) are manipulated so that
\[ E_a[T] = \theta. \] (28)

As we will see in the next chapter, when an exponential family is expressed in its mean value parameterization the sufficient statistic \(T\) is an unbiased minimum variance estimator of \(\theta\). Thus mean value parameterizations are very special and advantageous.

Examples of distributions in the exponential family include: Gaussian with unknown mean or variance, Poisson with unknown mean, exponential with unknown mean, gamma, Bernoulli with unknown success probability, binomial with unknown success probability, multinomial with unknown cell probabilities. Distributions which are not from the exponential family include: Cauchy with unknown median, uniform with unknown support, Fisher-F with unknown degrees of freedom.

When the statistical model is in the exponential family, sufficient statistics for the model parameters have a particularly simple form:

\[
\begin{align*}
    f_a(x) &= \prod_{i=1}^{n} a(\theta) b(x_i) e^{c(T(x_i))} \\
    &= a^n(\theta) e^{c^T(\theta) \sum_{i=1}^{n} t(x_i)} \prod_{i=1}^{n} b(x_i)
\end{align*}
\]

Therefore, the following is a \(p\)-dimensional sufficient statistic for \(\theta\)
\[
\sum_{i=1}^{n} t(X_i) = \left[ \sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_p(X_i) \right]^T
\]

In fact this is a finite dimensional suff. statistic which is complete and minimal [9].

### 3.6.2 CHECKING IF A DENSITY IS IN THE EXPONENTIAL FAMILY

Due to the many attractive properties of exponential families, in many situations the first question to be answered is: is the density of my data \(X\) a member of this exclusive club? This question might arise, for example, if the input to a known filter or other system has a known density and one can compute a mathematical representation of the density of the output of the filter. To check if the output density is exponential one has to try and manipulate the density into exponential form, as illustrated in the exercises. If this is difficult the next step is to try and show that the density is not in the exponential family. Some properties can be checked immediately, e.g. that
the parameters space $\Theta$ does not depend on the range of $X$, e.g. as in a uniform density with unknown region of support boundaries. Another simple test is to compute $\frac{\partial^2}{\partial \theta \partial x} \ln f_\theta(x)$ and verify that it is not of separable form $c' t'(x)$ for some functions $c$ and $t$. This type of question is explored in the exercises.

3.7 BACKGROUND REFERENCES

Mood, Graybill and Boes [56] offers an undergraduate introduction to mathematical statistics with lots of fun exercises and examples. Two of the classic graduate level text books on linear multivariate statistics are Rao [66] and Morrison [58]. Manoukian [51] is a reference book giving a concise compilation of principal results from sampling distribution theory. The book by Johnson et al [34], is the first of a set of several volumes of a very comprehensive encyclopedia of probability distributions, random variables, and their properties.

3.8 EXERCISES

3.1 Show that the matrix $\Pi = I_n - \frac{1}{n} 1 1^T$ is symmetric and idempotent, where $I_n$ is the $n \times n$ identity matrix and $1 = [1, \ldots, 1]^T$ is an $n$-element column vector of $1$’s. Show that for $\bar{x} \in \mathbb{R}^n$, $\Pi \bar{x}$ is the vector of residuals $[x_1 - \bar{x}_1, \ldots, x_n - \bar{x}_n]^T$ where $\bar{x}_i$ is the sample mean of elements of $\bar{x}$. Finally show that if $\bar{x}$ has the decomposition $y + c 1$ where $y$ has zero (sample) mean and $c$ is an arbitrary scalar, then $\Pi \bar{x} = y$, i.e. the matrix $\Pi$ extracts the zero (sample) mean component of $\bar{x}$. It is in this sense that $\Pi$ is an orthogonal projection matrix onto the space of zero (sample) mean vectors in $\mathbb{R}^n$.

3.2 Assume that a random vector $\bar{X} = [X_1, \ldots, X_n]^T$ has a density $p_\theta(\bar{x})$ which depends on an unknown parameter vector $\theta$. In this exercise you will show that if a statistic $S = S(\bar{X}) = [S_1(\bar{X}), \ldots, S_k(\bar{X})]^T$ satisfies the Fisher Factorization theorem then the conditional density $p_\theta(\bar{X}|S)$ is not a function of $\theta$ and thus $S$ is a sufficient statistic for $\theta$. In the following you should assume that $\bar{X}$ is a discrete random vector and that its joint density $p_\theta(\bar{x}) = P_\theta(\bar{X} = \bar{x})$ is a probability mass function (i.e. $p_\theta(\bar{x}) = 0$ except for a countable number of points $\bar{x} \in \{\bar{x}_1, \bar{x}_2, \ldots\}$ where $p_\theta(\bar{x}_i) > 0$, and $\sum_{\bar{x}} p_\theta(\bar{x}_i) = 1$).

(a) Use Bayes rule to establish that

$$p_\theta(\bar{x}|S) \overset{\text{def}}{=} P_\theta(\bar{X} = \bar{x}|S = s) = \frac{P_\theta(S = s|\bar{X} = \bar{x}) p_\theta(\bar{x})}{\sum_{\bar{x}'}: S(\bar{x}') = s p_\theta(\bar{x}')},$$

where the summation of $p_\theta(\bar{x})$ is over all possible realizations \{\bar{s}\} of the vector $\bar{X}$ such that $S(\bar{x}_i) = s$.

(b) Show that $P_\theta(S = s|\bar{X} = \bar{x})$ is equal to one or zero depending on whether $S(\bar{x}) = s$ or $S(\bar{x}) \neq s$, respectively. (Hint: express the conditional probability as a ratio and use the definition $S = S(\bar{X})$ to evaluate the intersection of the events $S = s$ and $\bar{X} = \bar{x}$).

(c) Using the Fisher Factorization $p_\theta(\bar{x}) = g_\theta(s) \cdot h(\bar{x})$ show that

$$p_\theta(\bar{x}|S) = \begin{cases} \frac{h(s)}{\sum_{\bar{x}}: S(\bar{x}) = s h(\bar{x})}, & S(\bar{x}) = s, \\ 0, & \text{o.w.} \end{cases}$$

which, as claimed, does not depend on $\theta$. 

3.8 EXERCISES

3.1 Show that the matrix $\Pi = I_n - \frac{1}{n} 1 1^T / n$ is symmetric and idempotent, where $I_n$ is the $n \times n$ identity matrix and $1 = [1, \ldots, 1]^T$ is an $n$-element column vector of $1$’s. Show that for $\bar{x} \in \mathbb{R}^n$, $\Pi \bar{x}$ is the vector of residuals $[x_1 - \bar{x}_1, \ldots, x_n - \bar{x}_n]^T$ where $\bar{x}_i$ is the sample mean of elements of $\bar{x}$. Finally show that if $\bar{x}$ has the decomposition $y + c 1$ where $y$ has zero (sample) mean and $c$ is an arbitrary scalar, then $\Pi \bar{x} = y$, i.e. the matrix $\Pi$ extracts the zero (sample) mean component of $\bar{x}$. It is in this sense that $\Pi$ is an orthogonal projection matrix onto the space of zero (sample) mean vectors in $\mathbb{R}^n$.

3.2 Assume that a random vector $\bar{X} = [X_1, \ldots, X_n]^T$ has a density $p_\theta(\bar{x})$ which depends on an unknown parameter vector $\theta$. In this exercise you will show that if a statistic $S = S(\bar{X}) = [S_1(\bar{X}), \ldots, S_k(\bar{X})]^T$ satisfies the Fisher Factorization theorem then the conditional density $p_\theta(\bar{X}|S)$ is not a function of $\theta$ and thus $S$ is a sufficient statistic for $\theta$. In the following you should assume that $\bar{X}$ is a discrete random vector and that its joint density $p_\theta(\bar{x}) = P_\theta(\bar{X} = \bar{x})$ is a probability mass function (i.e. $p_\theta(\bar{x}) = 0$ except for a countable number of points $\bar{x} \in \{\bar{x}_1, \bar{x}_2, \ldots\}$ where $p_\theta(\bar{x}_i) > 0$, and $\sum_{\bar{x}} p_\theta(\bar{x}_i) = 1$).

(a) Use Bayes rule to establish that

$$p_\theta(\bar{x}|S) \overset{\text{def}}{=} P_\theta(\bar{X} = \bar{x}|S = s) = \frac{P_\theta(S = s|\bar{X} = \bar{x}) p_\theta(\bar{x})}{\sum_{\bar{x}'}: S(\bar{x}') = s p_\theta(\bar{x}')},$$

where the summation of $p_\theta(\bar{x})$ is over all possible realizations \{\bar{s}\} of the vector $\bar{X}$ such that $S(\bar{x}_i) = s$.

(b) Show that $P_\theta(S = s|\bar{X} = \bar{x})$ is equal to one or zero depending on whether $S(\bar{x}) = s$ or $S(\bar{x}) \neq s$, respectively. (Hint: express the conditional probability as a ratio and use the definition $S = S(\bar{X})$ to evaluate the intersection of the events $S = s$ and $\bar{X} = \bar{x}$).

(c) Using the Fisher Factorization $p_\theta(\bar{x}) = g_\theta(s) \cdot h(\bar{x})$ show that

$$p_\theta(\bar{x}|S) = \begin{cases} \frac{h(s)}{\sum_{\bar{x}}: S(\bar{x}) = s h(\bar{x})}, & S(\bar{x}) = s, \\ 0, & \text{o.w.} \end{cases}$$

which, as claimed, does not depend on $\theta$. 

3.7 BACKGROUND REFERENCES

Mood, Graybill and Boes [56] offers an undergraduate introduction to mathematical statistics with lots of fun exercises and examples. Two of the classic graduate level text books on linear multivariate statistics are Rao [66] and Morrison [58]. Manoukian [51] is a reference book giving a concise compilation of principal results from sampling distribution theory. The book by Johnson et al [34], is the first of a set of several volumes of a very comprehensive encyclopedia of probability distributions, random variables, and their properties.
3.3 Show that the Poisson distribution \( p_\lambda(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \) \( x = 0, 1, 2, \ldots \) is a member of the one-parameter exponential family. For an i.i.d. sample \( \mathbf{X} = \{X_1, \ldots, X_n\}^T \) of these Poisson r.v.s find a one dimensional sufficient statistic for \( \lambda \). Define \( \alpha = 1/\lambda \) and show that the reparameterized Poisson distribution \( p_\alpha(x) \) is also in the exponential family. Which of these two parameterizations (\( \alpha \) or \( \lambda \)) is a mean value parameterization?

3.4 Let \( \mathbf{X} = \{X_1, \ldots, X_n\}^T \) be a vector of i.i.d. r.v.s \( X_i \) which are uniformly distributed over the interval \( (\theta_1, \theta_2) \), \( \theta_1 < \theta_2 \). Show that \( S(X) = [\min_i \{X_i\}, \max_i \{X_i\}]^T \) is a sufficient statistic for \( \theta = [\theta_1, \theta_2]^T \).

3.5 Let \( Z_i, i = 1, \ldots, n \), be a set of i.i.d. random variables each with the *alpha density*

\[
p_\theta(z) = \frac{\beta}{\sqrt{2\pi}\Phi(\alpha)z^2} \exp\left(-\frac{1}{2} [\alpha - \beta/z]^2\right),
\]

where \( \beta > 0 \) is unknown, \( \alpha \) is known and \( \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \) is the standard normal CDF. The alpha distribution is often used to model tool wear for rotating machinery.

(a) Is the joint density \( p_\theta(\mathbf{z}) \) a member of the exponential family of densities?

(b) using the Fisher Factorization find a two dimensional sufficient statistic for estimating the parameter \( \beta \) based on the observation \( \mathbf{Z} = [Z_1, \ldots, Z_n]^T \). Show that this reduces to a one dimensional (scalar) statistic when \( \alpha = 0 \).

3.6 Let \( \mathbf{X} = \{X_1, \ldots, X_n\}^T \) be a vector of i.i.d. Gaussian r.v.s with mean \( \mu \) and variance \( \sigma^2 = \mu^2 \) \( (X_i \sim \mathcal{N}(\mu, \sigma^2)) \).

(a) Show that the sample mean \( \mathbf{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \) is not a sufficient statistic for \( \mu \) by demonstrating that the conditional jpdf of \( \mathbf{X} \) given \( \mathbf{X} \) is a function of \( \mu \).

(b) Find a two dimensional sufficient statistic.

3.7 Let \( T = T(\mathbf{x}) \) be a sufficient statistic for \( \theta \), where \( \mathbf{x} \sim f(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta) h(\mathbf{x}) \) is a discrete random variable. Show that \( T \) has probability mass function

\[
f(t; \theta) = g(t, \theta)q(t),
\]

where

\[
q(t) = \sum_{\{\mathbf{x} : T(\mathbf{x}) = t\}} h(\mathbf{x}).
\]

3.8 Consider the case that \( \mathbf{X} = \{X_1, \ldots, X_n\}^T \) are drawn from a Bernoulli distribution, \( X_i \in \{0, 1\} \), \( P(X_i = 1) = 1 - P(X_i = 0) = p \), \( p \in [0, 1] \), and \( X_i \)’s are i.i.d. Show that the Binomial r.v. \( T = \sum_{i=1}^{n} X_i \) is a sufficient statistic for \( p \). Show that \( T \) is minimal. Also show that \( T \) is a complete sufficient statistic (Hint: for any function \( g \) express \( E_\theta[g(T)] \) as a polynomial in \( \theta = p \) and compute \( n \)-th order derivative wrt \( p \)).

3.9 Let \( X_1, \ldots, X_n \) be i.i.d. uniform r.v.s having common density \( f_{X_i}(x; \theta) = \frac{1}{\theta} I_{(0, \theta]}(x) \) \( (\theta > 0) \), where \( I_A(x) \) denotes the indicator function of the set \( A \). Show that \( T = \max(X_1, \ldots, X_n) \) is a complete sufficient statistic for \( \theta \) by the following steps:

(a) Show the sufficiency of \( T \).

(b) Derive the density function of \( T \).

(c) Show that \( E_\theta[g(T)] = 0 \), for all \( \theta > 0 \) implies \( g \) is identically zero.

End of chapter
4 FUNDAMENTALS OF PARAMETRIC ESTIMATION

In the last chapter we explored the foundation of statistical inference: the formulation of a statistical model and sufficient statistics for model parameters. In this chapter we go on to develop explicit methods to estimate the parameters from random samples from the model, paying close attention to how well the accuracy of these estimates hold up over different sample realizations.

We will start off with the basic mathematical formulation of estimation and then, specializing to the case of scalar one-dimensional parameters, consider two different models: random parameters and non-random parameters. It turns out, perhaps surprisingly, that estimation of random parameters has a cleaner theory. This is because for random parameters one can more straightforwardly assess the estimator’s mean accuracy and specify procedures for finding optimal estimators, called Bayes estimators, having highest possible accuracy. In particular we define three different optimality criteria mean squared error (MSE), mean absolute error (MAE), and mean uniform error, also called probability of large error ($P_e$). We then turn to deterministic scalar parameters for which we focus on bias and variance as measures of estimator accuracy. This leads to the concept of Fisher information and the Cramér-Rao lower bound on variance of unbiased estimators. Finally we generalize the treatment to multiple (vector) parameters.

4.1 ESTIMATION: MAIN INGREDIENTS

We follow the same notation as in the last chapter, summarized below.

- $X \in \mathcal{X}$ is a random measurement or observation
- $\mathcal{X}$ is the sample space of measurement realizations $x$
- $\theta \in \Theta$ is an unknown parameter vector of interest
- $\Theta \subset \mathbb{R}^p$ is the parameter space
- $f(x; \theta)$ is the pdf of $X$ for given $\theta$ (a known function)

With these definitions, the objective of parameter estimation is to design an estimator function

$$\hat{\theta} = \hat{\theta}(x)$$

which maps $\mathcal{X}$ to $\mathbb{R}^p \supset \Theta$. The concept is illustrated in Fig. 1.

It is important to distinguish between an estimator, which is a function of the sample $X$, and an estimate, which is an evaluation of the function at a particular realization $x$ of $X$, i.e.:

- the function $\hat{\theta}$ is an estimator.
- the point $\hat{\theta}(x)$ is an estimate.

A natural question arises. What is an appropriate design criterion for constructing an estimator? There are many possible approaches to this. In this chapter we will describe two of the principal approaches. The first assumes that $\theta$ is random and the second assumes it is deterministic. Common to both approaches is the specification of a loss function, also called a risk function, associated with an estimator that measures the estimation error as a function of both the sample and the parameter values.

Define $c(\hat{\theta}(x), \theta)$ a loss function associated with $\hat{\theta}$ for given $\theta$ and $X = x$. The optimum estimator, should it exist, might be found by minimizing average loss $\mathbb{E}[C]$, where as usual, the capitalization $C$ denotes the random variable $c(\hat{\theta}(X), \theta)$. 
4.2 ESTIMATION OF RANDOM SCALAR PARAMETERS

For the case that $\theta$ is a random scalar parameter $\theta$ we have access to the following information:

- $f(\theta)$: a prior p.d.f. for $\theta$.
- $f(x|\theta)$: a conditional p.d.f. (the response model) for $X$.
- $f(\theta|x)$: the posterior p.d.f. for $\theta$ that is determined by Bayes rule:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}.$$  

- $f(x)$: the marginal p.d.f. determined by marginalization over $\theta$

$$f(x) = \int_{\Theta} f(x|\theta)f(\theta) d\theta.$$  

With the above we can compute the average loss, also called Bayes risk, as

$$E[C] = \int_{\Theta} \int_X c(\hat{\theta}(x), \theta)f(x|\theta)f(\theta) dx d\theta.$$  

We now can naturally define an optimal estimator. A scalar estimator $\hat{\theta}$ which minimizes the average loss is called a Bayes estimator. Some reasonable loss functions for this estimation problem are:

- $c(\hat{\theta}; \theta) = (\hat{\theta} - \theta)^2$: squared error
- $c(\hat{\theta}; \theta) = |\hat{\theta} - \theta|$: absolute error
- $c(\hat{\theta}; \theta) = I(|\hat{\theta} - \theta| > \epsilon)$: uniform error

Figure 1: An estimator of a $p$-dimensional parameter $\Theta$ given an $n$-dimensional random sample $X$ is a mapping of $X$ to $\mathbb{R}^p$. 
Figure 2: Three loss functions for scalar parameter estimation: (a) squared error, (b) absolute error, (c) uniform error.

Figure 2 illustrates these three loss functions as a function of the estimator error difference \( \hat{\theta} - \theta \).

For each of the three loss functions we can compute the mean loss and obtain the Bayes risk functions (functions of \( f(\theta), f(x|\theta) \) and \( \hat{\theta} \)):

**Estimator MSE:**

\[
\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]
\]

**Estimator MAE:**

\[
\text{MAE}(\hat{\theta}) = E[|\hat{\theta} - \theta|]
\]

**Error Probability:**

\[
P_e(\hat{\theta}) = P(|\hat{\theta} - \theta| > \epsilon)
\]

It remains to find the estimators \( \hat{\theta} \), called **optimal estimators**, which minimize each of these criteria.

### 4.2.1 MINIMUM MEAN SQUARED ERROR ESTIMATION

The MSE is the most widespread estimation criterion and arguably the one with the longest history. The optimal minimum mean squared error estimator (MMSEE) is the **conditional mean estimator** (CME) defined as

\[
\hat{\theta}(X) = E[\theta|X] = \text{mean}_{\theta \in \Theta}\{f(\theta|X)\},
\]

where

\[
\text{mean}_{\theta \in \Theta}\{f(\theta|X)\} = \int_{-\infty}^{\infty} \theta f(\theta|X)d\theta.
\]

The CME has an intuitive mechanical interpretation as the center of mass (1st moment of inertia) of the mass density \( f(\theta|x) \) (Fig. 3). The CME corresponds to the posterior average value of the parameter after you have observed the data sample.
The CME satisfies an orthogonality condition: the Bayes estimator error is orthogonal to any (linear or non-linear) function of the data. This condition is mathematically expressed below for the general case of complex rv's,

$$E[(\theta - \hat{\theta}(X))g(X)^*] = 0,$$

for any function $g$ of $x$. Here $u^*$ denotes complex conjugate of $u$.

Proof: Write the MSE as

$$E[|\hat{\theta} - \theta|^2] = E[|\hat{\theta} - E[\theta|X]|^2] - (\theta - E[\theta|X])^2]$$

$$= E[|\hat{\theta} - E[\theta|X]|^2] + E[|\theta - E[\theta|X]|^2]$$

$$-E[g(X)^*(\theta - E[\theta|X])] - E[g(X)(\theta - E[\theta|X])^*]$$

where $g(X) = \hat{\theta} - E[\theta|X]$ is a function of $X$ only.

Step 1: show orthogonality condition

$$E[g(X)(\theta - E[\theta|X])] = E[E[g(X)(\theta - E[\theta|X])^*|X]]$$

$$= E \left[ g(X) E_{\theta = 0} [\theta - E[\theta|X] | X] \right] = 0$$

Step 2: Next show $E[\theta|X]$ minimizes MSE
\[ E[|\hat{\theta} - \theta|^2] = E[|\hat{\theta} - E[\theta | X]|^2] + E[|\theta - E[\theta | X]|^2] \]
\[ \geq E[|\theta - E[\theta | X]|^2] \]
where “=” occurs iff \( \hat{\theta} = E[\theta | X] \)

4.2.2 MINIMUM MEAN ABSOLUTE ERROR ESTIMATOR

For convenience we assume \( \theta \) is a real valued scalar and \( F(\theta | x) = \int_{-\infty}^{\theta} f(\theta' | x) d\theta' \) is a continuous function of \( \theta \). The minimal mean absolute error estimator (MMAEE) is the conditional median estimator (CmE)

\[ \hat{\theta}(X) = \text{median}_{\theta \in \Theta} \{ f(\theta | X) \}, \]

where

\[
\text{median}_{\theta \in \Theta} \{ f(\theta | X) \} = \min \left\{ u : \int_{-\infty}^{u} f(\theta | X) d\theta = 1/2 \right\} \tag{29}
\]
\[
= \min \left\{ u : \int_{-\infty}^{u} f(X | \theta) f(\theta) d\theta = \int_{u}^{\infty} f(X | \theta) f(\theta) d\theta \right\}. \tag{30}
\]

The median of a density separates the density into two halves of equal mass (Fig. 4). When \( F(\theta | x) \) is strictly increasing over \( \Theta \) the ”min” in the definition of the median is not necessary - but it may be required when there are regions of \( \Theta \) where the density \( f(\theta | X) \) is equal to zero. If \( f(\theta | X) \) is continuous in \( \theta \) the CmE also satisfies an orthogonality condition:

\[ E[\text{sgn}(\theta - \hat{\theta}(X)) g(X)] = 0, \]

and thus for minimum MAE estimation it is the sign of the optimum estimation error that is orthogonal to any function of the data sample.

Proof: Let \( \hat{\theta}_m = \text{median of} \ f(\theta | X) \).

Then by definition of median for continuous densities

\[
E[\text{sgn}(\theta - \hat{\theta}_m) \mid X] = \int_{\Theta} \text{sgn}(\theta - \hat{\theta}_m(X)) f(\theta | X) d\theta
\]
\[
= \int_{\theta > \hat{\theta}_m(X)} f(\theta | X) d\theta - \int_{\theta \leq \hat{\theta}_m(X)} f(\theta | X) d\theta
\]
\[
= 0
\]

Step 1: show orthogonality condition:

\[ E[\text{sgn}(\theta - \hat{\theta}_m) g(X)] = E[ E[\text{sgn}(\theta - \hat{\theta}_m) | X] g(X) ] \]

=0
Step 2: for $\hat{\theta}$ arbitrary we have (apply “useful formula” below)

$$\text{MAE}(\hat{\theta}) = E[|\theta - \hat{\theta}_m + \hat{\theta}_m - \hat{\theta}|]$$

$$= E[|\theta - \hat{\theta}_m|] + E[\text{sgn}(\theta - \hat{\theta})\Delta]_{\text{sgn}(\theta - \hat{\theta}) = 0}$$

$$+ E[\text{sgn}(a + \Delta) - \text{sgn}(a)](a + \Delta)_{\text{sgn}(a + \Delta) \geq |\text{sgn}(a + \Delta) - \text{sgn}(a)|}$$

$$\geq E[|\theta - \hat{\theta}_m|]$$

Useful formula: $|a + \Delta| = |a| + \text{sgn}(a)\Delta + [\text{sgn}(a + \Delta) - \text{sgn}(a)](a + \Delta)$

### 4.2.3 Minimum Mean Uniform Error Estimation

Unlike the MSE or MAE, the MUE penalizes only those errors that exceed a tolerance level $\epsilon > 0$ and this penalty is uniform. For small $\epsilon$ the optimal estimator is the maximum a posteriori (MAP) estimator, which is also called the posterior mode estimator (Fig. 5)

$$\hat{\theta}(X) = \arg\max_{\theta \in \Theta} \{ f(\theta|X) \}$$

$$= \arg\max_{\theta \in \Theta} \left\{ \frac{f(X|\theta)f(\theta)}{f(X)} \right\}$$

$$= \arg\max_{\theta \in \Theta} \{ f(X|\theta)f(\theta) \}. \tag{33}$$
Notice that the third line of (33) is best suited to computation of the MAP estimator since it does not require the marginal $f(x)$, which can be difficult to compute.

**Proof:**

Assume that $\epsilon$ is a small and positive number. The probability that the magnitude estimator error exceeds $\epsilon$ is simply expressed

$$P_e(\hat{\theta}) = 1 - P(|\theta - \hat{\theta}| \leq \epsilon)$$

$$= 1 - \int_X dx f(x) \int_{\{\theta : |\theta - \hat{\theta}(x)| \leq \epsilon\}} f(\theta|x)d\theta.$$

Consider the inner integral (over $\theta$) in the above expression. This is an integral over $\theta$ within a window, which we call the **length $2\epsilon$ window**, centered at $\hat{\theta}$. Referring to Fig. 6, it should be evident to the reader that, if $\epsilon$ is sufficiently small, this integral will be maximized by centering the length $2\epsilon$ window at the value of $\theta$ that maximizes the integrand $f(\theta|x)$. This value is of course the definition of the MAP estimate $\hat{\theta}$.

Now that we have seen three different estimator criteria, and their associated optimal estimators, we make several general remarks.

1. The CmE may not exist for discrete $\Theta$ since the median may not be well defined.
2. Only the CME requires (often difficult) computation of the normalization factor $f(x)$ in the posterior $f(\theta|x) = f(x|\theta)f(\theta)/f(x)$.
3. Each of these estimators depends on $x$ only through posterior $f(\theta|x)$.
4. When the posterior is continuous, unimodal, and symmetric then each of the above estimators are identical (VanTrees [84])! See Fig. 7 for illustration.
5. If $T = T(X)$ is a sufficient statistic the posterior depends on $X$ only through $T$. Indeed, if $f(X|\theta) = g(T;\theta)h(X)$, then by Bayes rule

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{\int_\Theta f(X|\theta)f(\theta)d\theta} = \frac{g(T;\theta)f(\theta)}{\int_\Theta g(T;\theta)f(\theta)d\theta}$$
which is only a function of $X$ through $T$. Thus, in terms of optimal estimation performance, one loses nothing by compressing $X$ to a sufficient statistic.

6. The CME has the following linearity property. For any random parameter variables $\theta_1$ and $\theta_2$: $E[\theta_1 + \theta_2 | X] = E[\theta_1 | X] + E[\theta_2 | X]$. This property is not shared by the CmE or the MAP estimator.

4.2.4 BAYES ESTIMATOR EXAMPLES

Here we give four examples of statistical models, priors, and derive their optimal estimators under various criteria.

These are the examples we will cover (hotlinks on the web version)

* Estimation of width of uniform density
* Estimation of a Gaussian signal
* Estimation of magnitude of Gaussian signal
* Estimation of a binary signal in Gaussian noise

Example 10 ESTIMATION OF WIDTH OF UNIFORM PDF

Consider the following motivating problem. A networked computer terminal takes a random amount of time to connect to another terminal after sending a connection request at time $t = 0$. You, the user, wish to schedule a transaction with a potential client as soon as possible after sending the request. However, if your machine does not connect within the scheduled time then your client will go elsewhere. If one assumes that the connection delay is a random variable $X$ that is uniformly distributed over the time interval $[0, \theta]$ you can ensure your client that the delay will not exceed $\theta$. The problem is that you do not know $\theta$ so it must be estimated from past
experience, e.g., the sequence of previously observed connection delays $X_1, \ldots, X_n$. By assuming a prior distribution on $\theta$ an optimal estimate can be obtained using the theory developed above. So now let’s formulate this in our language of estimation theory.

We assume that $X_1, \ldots, X_n$ are conditionally i.i.d. uniform samples each with conditional density

$$f(x_1|\theta) = \frac{1}{\theta} I_{[0,\theta]}(x_1).$$

Let’s say that based on your experience with lots of different clients you determine that a reasonable prior on $\theta$ is

$$f(\theta) = \theta e^{-\theta}, \quad \theta > 0.$$

Figure 8 illustrates these two densities.

We will derive the CME, CmE, and MAP estimators of $\theta$. There are two steps.

Step 1: Find the posterior $f(\theta|x) = f(x|\theta)f(\theta)/f(x)$

$$f(x|\theta)f(\theta) = \left( \prod_{i=1}^{n} \frac{1}{\theta} I_{[x_i,\infty)}(\theta) \right) \left( \theta e^{-\theta} \right)$$

$$= \frac{e^{-\theta}}{\theta^{n-1}} \prod_{i=1}^{n} I_{[x_i,\infty)}(\theta)$$

$$= \frac{e^{-\theta}}{\theta^{n-1}} I_{[x(1),\infty)}(\theta),$$

where $x(1) = \max\{x_i\}$. Observe that the function $\frac{e^{-\theta}}{\theta^{n-1}}$ is monotone decreasing over $\theta > 0$ (verify that the derivative of its logarithm is negative).
Furthermore,

\[ f(x) = \int_0^\infty f(x|\theta)f(\theta)d\theta = q_{-n+1}(x_{(1)}) \]

where \( q_n \) is the monotone decreasing function

\[ q_n(x) \triangleq \int_x^\infty \theta^n e^{-\theta}d\theta \]

Recursive formula: \( q_{n-1}(x) = \frac{1}{n} \left( \frac{1}{\sqrt{\pi}} e^{-x} - q_n(x) \right), \) \( n = 0, -1, -2, \ldots \)

Step 2: find optimal estimator functions:

\[ \hat{\theta}_{\text{MAP}} = X_{(1)} \]

\[ \hat{\theta}_{\text{CME}} = q_{n+2}(X_{(1)})/q_{n+1}(X_{(1)}) \]

\[ \hat{\theta}_{\text{CmE}} = q_{n+1}^{-1} \left( \frac{1}{2} q_{n+1}(X_{(1)}) \right). \]

Note that only the MAP estimator is a simple function of \( X \) while the two others require more difficult computation of integrals \( q_n \) and/or an inverse function \( q_n^{-1} \). These estimators are illustrated in Fig. 9 along with the posterior density \( f(\theta|x) \).

Example 11  ESTIMATION OF GAUSSIAN AMPLITUDE
A very common assumption arising in many signal extraction problems is the assumption of a Gaussian distributed signal observed in additive Gaussian noise. For example, a radar target acquisition system might transmit a pulse to probe for possible targets in a cell located at a particular point in space. If a strong reflecting target is present at that point then it reflects some of the energy in the radar pulse back to the radar, resulting in a high energy signal, called a radar return, at the radar receiver. The amplitude of this signal might contain useful information about the identity of the target. Estimation of the radar return is complicated by the presence of ambient noise generated in the radar receiver (thermal noise) or by interference from other sources (clutter) in the cell. Based on field trials of the radar system prior mean and variances of the received signal and the noise might be available.

To set this up more formally as an estimation problem we define two jointly Gaussian r.v.s: $S, X$ with known means, variances, and covariance

\[
E[S] = \mu_S, \quad E[X] = \mu_X, \\
\text{var}(S) = \sigma_S^2, \quad \text{var}(X) = \sigma_X^2 \\
\text{cov}(S, X) = \rho \sigma_S \sigma_X.
\]

$S$ will play the role of the signal and $X$ will be the measurement. Of course the specific form of the covariance function will depend on the receiver structure, e.g., it reduces to a simple function of $\sigma_S$ and $\sigma_X$ for an additive noise model.

The objective is to find an optimal estimator of $S$ given measured $X$. As in the previous example the derivation of CME, CmE and MAP estimators is divided into two parts.

Step 1: find the posterior density.

A fundamental fact about jointly Gaussian random variables is that if you condition on one of the variables then the other variable is also Gaussian, but with different mean and variance equal to its conditional mean and variance (see Fig. 10 and Exercise 4.25 at the end of chapter). In
particular, the conditional density of $S$ given $X = x$ is Gaussian with mean parameter

$$
\mu_{S|X}(x) = E[S|X = x] = \mu_S + \rho \frac{\sigma_S}{\sigma_X} (x - \mu_X),
$$

and variance parameter

$$
\sigma^2_{S|X} = E[(S - E[S|X])^2|X = x] = (1 - \rho^2)\sigma^2_S,
$$

so that the conditional density takes the form

$$
f_{S|X}(s|x) = \frac{f_X|S(x|s)f_S(s)}{f_X(x)} = \frac{1}{\sqrt{2\pi\sigma^2_{S|X}}} \exp\left\{ -\frac{(s - \mu_{S|X}(x))^2}{2\sigma^2_{S|X}} \right\}.
$$

Figure 10: The posterior $f(s|x)$ when $s, x$ are jointly Gaussian is a Gaussian density.

Step 2: find the form of the optimal estimators

We immediately note that, as the posterior is continuous, symmetric and unimodal, the MAP, CME, and CmE estimators are of identical form. Bringing out the explicit dependency of the estimator $\hat{S}$ on the observed realization $x$ we have:

$$
\hat{S}(x) = \mu_{S|X}(x) = \text{linear in } x.
$$

An interesting special case, relevant to the radar example discussed above, is the independent additive noise model where $X = S + V$. For this case $\sigma^2_X = \sigma^2_S + \sigma^2_V$, $\rho^2 = \sigma^2_S/(\sigma^2_S + \sigma^2_V)$ and therefore

$$
\hat{S}(x) = \mu_S + \frac{\sigma^2_S}{\sigma^2_S + \sigma^2_V} (x - \mu_X).
$$
We can easily derive the performance of the optimal estimator under the MSE criterion

Minimum MSE: $E[(S - \hat{S})^2] = (1 - \rho^2)\sigma_S^2$.

A little more work produces expressions for the performances of this optimal estimator under the MAE and Pe (MUE) criteria:

Minimum MAE: $E[|S - \hat{S}|] = \sqrt{(1 - \rho^2)\sigma_S^2} \sqrt{\frac{2}{\pi}}$

Minimum Pe: $P(|S - \hat{S}| > \epsilon) = 1 - \text{erf} \left( \frac{\epsilon}{\sqrt{2(1 - \rho^2)\sigma_S^2}} \right)$

**Example 12 Estimation of magnitude of Gaussian signal**

Now we change Example 11 a little bit. What if the radar operator was only interested in the energy of the received signal and not its sign (phase)? Then the proper objective would be to estimate the magnitude $|S|$ instead of the magnitude and phase $S$. Of course, an ad hoc estimation procedure would be to simply take the previously derived estimator $\hat{S}$ and use its magnitude $|\hat{S}|$ to estimate $|S|$ but is this the best we can do?

Let’s see what the form of the best estimators of $|S|$ are.

Again we define two jointly Gaussian r.v.s: $S, X$ with means, variances, and covariance

\[
E[S] = \mu_S, \quad E[X] = \mu_X, \\
\text{var}(S) = \sigma_S^2, \quad \text{var}(X) = \sigma_X^2, \\
\text{cov}(S, X) = \rho \sigma_S \sigma_X.
\]

Now the objective is to estimate the random variable $Y = |S|$ based on $X$. Note: the pair $Y, X$ no longer obeys a jointly Gaussian model. But, using first principles, we can easily derive the optimal estimators. The first step is to compute the posterior density $f_{Y|X}$.

![Figure 11: Illustration of the method of differentials for finding conditional density of $Y = |S|$ given $X$ from the probability $P(y < Y \leq y + \Delta|X = x) \approx f_{Y|X}(y|x)\Delta, 0 < \Delta \ll 1$.](image-url)
Since we know \( f_{S|X} \) from the previous example this is a simple transformation of variables problem of elementary probability. We use the method of differentials (see Fig. 11) to obtain the following relation, valid for small \( \Delta \)

\[
f_{Y|X}(y|x) = f_{S|X}(y|x) \Delta + f_{S|X}(-y|x) \Delta, \quad y \geq 0,
\]
or more explicitly

\[
f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi \sigma_{S|X}^2}} \left( \exp\left\{ -\frac{(y - \mu_{S|X}(x))^2}{2\sigma_{S|X}^2} \right\} + \exp\left\{ -\frac{(y + \mu_{S|X}(x))^2}{2\sigma_{S|X}^2} \right\} \right) I_{[0,\infty)}(y). \tag{34}
\]

![Figure 12: Posterior density of \( Y = |S| \) given \( X \)](image)

Unlike Example 11 this posterior density, shown in Fig. 12 is no longer symmetric in \( y \). Hence we expect the CME, CmE, and MAP estimators to be different.

The CME can be derived in explicit closed form by integration over \( y \in [0,\infty) \) of the function \( yf_{Y|X}(y|x) \) specified in (34)

\[
\hat{Y}_{\text{CME}}(x) = E[Y|X = x] = |\mu_{S|X}(x)| \text{ erf} \left( \frac{|\mu_{S|X}(x)|}{\sigma_{S|X} \sqrt{2}} \right) + \sqrt{\frac{\sigma_{S|X}}{\pi}} e^{-\mu_{S|X}^2/2\sigma_{S|X}^2}.
\]

On the other hand, by investigating the MMAE equation \( \int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = \int_{0}^{\hat{Y}} f_{Y|X}(y|x)dy \) it is easily seen that the CmE can only be implicitly given as the solution \( \hat{Y} = \hat{Y}_{\text{CmE}} \) of the following

\[
\text{erf} \left( \frac{\hat{Y} - \mu_{S|X}(x)}{\sigma_{S|X} \sqrt{2}} \right) + \text{erf} \left( \frac{\hat{Y} + \mu_{S|X}(x)}{\sigma_{S|X} \sqrt{2}} \right) = \frac{1}{2}.
\]
Finally, as $f_{Y|X}(y|x)$ is concave and smooth in $y$, the MAP estimator $\hat{Y} = \hat{Y}_{\text{MAP}}$ occurs at a stationary point in $y$ of the so called “MAP equation”

$$0 = \frac{\partial f(y|x)}{\partial y}.$$ 

Using (34) this yields

$$\hat{Y}(x) = \mu_{S|X}(x) \frac{\exp \left\{ -\frac{(\hat{Y} - \mu_{S|X}(x))^2}{2\sigma_{S|X}^2} \right\} - \exp \left\{ -\frac{(\hat{Y} + \mu_{S|X}(x))^2}{2\sigma_{S|X}^2} \right\}}{\exp \left\{ -\frac{(\hat{Y} - \mu_{S|X}(x))^2}{2\sigma_{S|X}^2} \right\} + \exp \left\{ -\frac{(\hat{Y} + \mu_{S|X}(x))^2}{2\sigma_{S|X}^2} \right\}}.$$ 

Figure 13: *Three optimal estimators of $Y = |S|$ when $S, X$ are jointly Gaussian.*

The above optimal estimators are illustrated in Fig. 13. It can be verified that as $\mu_{S|X}/\sigma_{S|X} \to \infty$ all three estimators converge to an identical limit:

$$\hat{Y}(x) \to |\mu_{S|X}(x)|.$$ 

This limiting case occurs since the posterior density becomes a dirac delta function concentrated at $y = \mu_{S|Y}(x)$ as $\mu_{S|X}/\sigma_{S|X} \to \infty$. Observe that none of these estimators of $|S|$ are given by $|\hat{S}|$ where $\hat{S}$ is the corresponding MAP/CME/CmE estimate of $S$ derived in Example 11. This illustrates an important fact: estimation of random parameters is not invariant to functional transformation,

**Example 13 Estimation of sign of Gaussian signal**

Above we derived optimal estimators for magnitude of a Gaussian random variable based on Gaussian observations. Well, how about when only the phase is of interest, e.g., when the radar
operator wants to estimate the sign as opposed to the magnitude of the signal? We treat a simplified version of this problem in this example.

Assume that the model for the observation is

$$X = \theta + W$$

where $W$ is a zero mean Gaussian noise with variance $\sigma^2$ and $\theta$ is an equally likely binary random variable: $P(\theta = 1) = P(\theta = -1) = \frac{1}{2}$, $\Theta = \{-1, 1\}$. This corresponds to our radar problem when the prior mean $\mu_S$ is zero (why?) and an additive noise model is assumed.

Here the posterior density is a probability mass function since the signal $\theta$ is discrete valued:

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)},$$

where $p(\theta) = 1/2$. For convenience we have eliminated subscripts on densities. Furthermore, as illustrated in Fig. 14,

$$f(x|\theta) = \begin{cases} 
\frac{1}{\sqrt{2\pi}\sigma^2}\exp\left(-\frac{(x-1)^2}{2\sigma^2}\right), & \theta = 1 \\
\frac{1}{\sqrt{2\pi}\sigma^2}\exp\left(-\frac{(x+1)^2}{2\sigma^2}\right), & \theta = -1 
\end{cases}.$$

Hence

$$f(x) = f(x|\theta = 1)\frac{1}{2} + f(x|\theta = -1)\frac{1}{2}.$$

**Figure 14:** The posterior density $f(\theta|x)$ concentrates mass on the pair of points $\theta = \pm 1$.

From the following steps we discover that the MAP estimator is a minimum distance decision rule, i.e., it selects the value $\hat{\theta}$ as that value of $\theta$ which is closest to the measured value $X$:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta = 1, -1} f(X|\theta)$$
\[ = \arg\min_{\theta = -1, 1} \{(X - \theta)^2\} \]
\[ = \begin{cases} 
1, & X \geq 0 \\
-1, & X < 0 
\end{cases} \]

On the other hand, the CME estimator is

\[
\hat{\theta}_{CME} = (1) P(\theta = 1|X) + (-1) P(\theta = -1|X)
\]
\[
= \frac{\exp\left(-\frac{(X-1)^2}{2\sigma^2}\right) - \exp\left(-\frac{(X+1)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(X-1)^2}{2\sigma^2}\right) + \exp\left(-\frac{(X+1)^2}{2\sigma^2}\right)}.
\]

The MAP and CME estimators are illustrated in Fig. 15. Unfortunately, we cannot derive the CME since it is not well defined for discrete valued parameters \(\theta\) (why?).

Based on these above examples we make the summary remarks:

1. Different error criteria usually give different optimal estimators.
2. Optimal estimators of random parameters are not invariant to functional transformations. Specifically, if \(\hat{g}(\theta)\) is an optimal estimator of \(g(\theta)\) and \(\hat{\theta}\) is an optimal estimator of \(\theta\):

\[ \hat{g}(\theta) \neq g(\hat{\theta}) \]

in general.
3. When they exist, the CmE and MAP estimators always take values in the parameter space Θ. The values taken by CME may fall outside of Θ, e.g., if it is discrete or if it is not a convex set.

4. The “MAP equation” stationary point condition \( \partial f(\theta|x)/\partial \theta = 0 \) at \( \theta = \hat{\theta}_{MAP} \) is only useful for continuous densities that are differentiable and concave in continuous valued parameters θ (Fig. 16).

![Figure 16: Use of the stationary point MAP equation can fail to find the MAP estimator. In general there may exist no stationary points of the posterior density \( f_2, f_3 \). or there may be multiple stationary points of the posterior density \( f_1 \).](image)

### 4.3 ESTIMATION OF RANDOM VECTOR VALUED PARAMETERS

The problem of estimation of multiple unknown parameters is formulated as estimation of a vector valued parameter \( \theta \in \Theta \subseteq \mathbb{R}^p \). When the parameter vector is random we can define a Bayesian estimation criterion just like in the scalar case considered above. It suffices to optimize a generalization of the scalar criterion \( E[c(\hat{\theta}, \theta)] \) to handle vector parameter estimation. This turns out to be quite easy, at least for two of our proposed estimation criteria. Some possible generalizations of the previous three scalar criteria are (Figs. 17-20)

**Estimator total mean squared error (MSE):**

\[
\text{MSE}(\hat{\theta}) = E[\|\hat{\theta} - \theta\|^2] = \sum_{i=1}^{p} E[(\hat{\theta}_i - \theta_i)^2].
\]

**Estimator total mean absolute error (MAE):**

\[
\text{MAE}(\hat{\theta}) = E[\|\hat{\theta} - \theta\|_1] = \sum_{i=1}^{p} E[|\hat{\theta}_i - \theta_i|].
\]

**Estimator maximum error probability:**

\[
P_e(\hat{\theta}) = 1 - P(\|\hat{\theta} - \theta\|_\infty \leq \epsilon),
\]
where $\|\hat{\theta} - \theta\|_\infty = \max_{i=1,\ldots,p} |\hat{\theta}_i - \theta_i|$ is the $l_\infty$ norm of the error vector $\hat{\theta} - \theta$. Similarly to the scalar case, this error probability can be expressed as the statistical expectation of a uniform error criterion, taking value 0 inside a cube shaped region of edge length $2\epsilon$.

The MAE criterion, also known as total variation norm, does not often lead to unique optimal vector-valued estimators. Although the total variation norm has been of substantial recent interest, in our introductory treatment only MSE and $P_e$ will be discussed.

### 4.3.1 VECTOR SQUARED ERROR

As $\text{MSE}(\hat{\theta}) = \sum_{i=1}^p \text{MSE}(\hat{\theta}_i)$ is an additive function, the minimum MSE vector estimator attains the minimum of each component $\text{MSE}(\hat{\theta}_i), i = 1, \ldots, p$. Hence, we have the nice result that the vector minimum MSE estimator is simply the vector of scalar CME’s for each component:

$$
\hat{\theta}_{\text{CME}} = E[\hat{\theta}|X] = \begin{bmatrix}
E[\theta_1|X] \\
\vdots \\
E[\theta_p|X]
\end{bmatrix}
$$

As in the case of scalar estimation the minimum MSE estimator is the center of mass of the multivariate posterior density (Figs. 21-22).

### 4.3.2 VECTOR UNIFORM ERROR

For small $\epsilon$ the minimum mean uniform error ($P_e$) is attained by the vector MAP estimator which has form similar to the scalar MAP estimator

$$
\hat{\theta}_{\text{MAP}} = \arg\max_{\theta \in \Theta} f(\theta|x).
$$
Figure 18: Absolute error criterion $c(\hat{\theta}, \theta)$ for which $E[c(\hat{\theta}, \theta)]$ is the total mean absolute error.

Figure 19: Uniform error criterion for which $E[c(\hat{\theta}, \theta)]$ is the maximum probability of error.
Figure 20: *Constant contours of the three error criteria in 17-18.*

Figure 21: *Bivariate posterior density of two unknown parameters. Optimal estimates shown in Fig. 22.*
4.4 ESTIMATION OF NON-RANDOM PARAMETERS

To estimate random parameters one has a prior distribution and we can define a global estimation error criterion, the Bayes risk, which depends on the prior but not on any particular value of the parameter. In non-random parameter estimation there is no prior distribution. One can of course look at the problem of estimation of non-random parameters as estimation of random parameters conditioned on the value of the parameter, which we could call the true value. However, the formulation of optimal non-random parameter estimation requires a completely different approach. This is because if we do not have a prior distribution on the parameter virtually any reasonable estimation error criterion will be local, i.e., it will depend on the true parameter value. Thus we will need to define weaker properties than minimum risk, such as unbiasedness, that a good estimator of non-random parameters should have.

As before we first consider estimation of scalar non-random parameters $\theta$. In this case it does not make sense to use the conditional density notation $f(x|\theta)$ and we revert to the alternative notation for the model $f_{\theta}(x) = f(x; \theta)$.

So, what are some possible design criteria for estimators of scalar real $\theta$? One could try to minimize MSE, defined as

$$\text{MSE}_\theta = E_{\hat{\theta}}[(\hat{\theta} - \theta)^2].$$

Here we encounter a difficulty: if the true value $\theta$ is $\theta_0$, the constant estimator $\hat{\theta} = c$ attains 0 MSE when $\theta_0 = c$ (Fig. 23).

4.4.1 SCALAR ESTIMATION CRITERIA FOR NON-RANDOM PARAMETERS

Some possible scalar criteria for designing good estimators are the minimax criteria.
1. Minimize worst case MSE. Choose \( \hat{\theta} \) to minimize

\[
\max_{\theta} \text{MSE}(\hat{\theta}) = \max_{\theta} E_\theta[(\hat{\theta} - \theta)^2]
\]

2. Minimize worst case estimator error probability:

\[
\max_{\theta} P_e = \max_{\theta} P_\theta(|\hat{\theta} - \theta| > \epsilon)
\]

If we would be satisfied by minimizing an upper bound on \( P_e \), then we could invoke Tchebychev inequality

\[
P_\theta(|\hat{\theta} - \theta| \geq \epsilon) \leq \frac{E_\theta[(\hat{\theta} - \theta)^2]}{\epsilon^2}
\]  (35)

and focus on minimizing the worst case MSE. There is a large literature on minimax MSE estimation, see for example [47], but the mathematical level necessary to develop this theory is too advanced for an introductory treatment. We will not consider minimax estimation further in this book.

We next give several weaker conditions that a good estimator should satisfy, namely consistency and unbiasedness.

Definition: \( \hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n) \) is said to be (weakly) consistent if for all \( \theta \) and all \( \epsilon > 0 \)

\[
\lim_{n \to \infty} P_\theta(|\hat{\theta}_n - \theta| > \epsilon) = 0
\]

This means that \( \hat{\theta}_n \) converges in probability to the true parameter \( \theta \). It also means that the pdf of the estimator concentrates about \( \theta \) (Fig. 24). Furthermore, by the Tchebychev inequality (35), if MSE goes to zero as \( n \to \infty \) then \( \hat{\theta}_n \) is consistent. As the MSE is usually easier to derive than
shows that MSE converges to zero is the typical way that one shows that an estimator is consistent.

For an estimator $\hat{\theta}$ define the estimator bias at a point $\theta$ to be

$$b_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}] - \theta.$$  

Likewise the estimator variance is

$$\text{var}_{\theta}(\hat{\theta}) = E_{\theta}[(\hat{\theta} - E_{\theta}[\hat{\theta}])^2].$$

Here the reader should recall the definition of the expectation operator $E_{\theta}$: $E_{\theta}[g(X)] = \int_X g(x)f(x;\theta)dx$, where $X$ is a r.v. with density $f(x;\theta)$. As compared to the Bayes expectation $E[g(X)]$ used for random parameters, this expectation acts like a conditional expectation given a specific value of $\theta$.

It is natural to require that a good estimator be unbiased, i.e., $b_{\theta}(\hat{\theta}) = 0$ for all $\theta \in \Theta$. This suggests a reasonable design approach: constrain the class of admissible estimators to be unbiased and try to find one that minimizes variance over this class. In some cases such an approach leads to a really good, in fact optimal, unbiased estimator called a UMVU estimator (Fig. 25). A caveat to the reader is necessary however: there exist situations where unbiasedness is not a desirable property to impose on an estimator. For example there are models for which no unbiased estimator of the model parameter exists and others for which the biased estimator has unreasonably high MSE, see Exercises at the end of this chapter and [67, Sec. 7.11, 7.15].

Definition: $\hat{\theta}$ is said to be a uniform minimum variance unbiased (UMVU) estimator if for all $\theta \in \Theta$ it has less variance than any other unbiased estimator $\hat{\theta}$. Thus a UMVU estimator satisfies

$$\text{var}_{\theta}(\hat{\theta}) \leq \text{var}_{\theta}(\hat{\theta}), \quad \theta \in \Theta$$

Unfortunately, UMVU estimators only rarely exist for finite number $n$ of samples $X_1, \ldots, X_n$. Thus one is usually forced to sacrifice the unbiasedness constraint in order to develop good tractable
estimation procedures. For such estimators there exists an important relation between MSE, variance and bias:

$$\text{MSE}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2] = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}]) + (E_\theta[\hat{\theta}] - \theta)]^2$$

$$= \frac{E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]}{\text{var}_\theta(\hat{\theta})} + \frac{(E_\theta[\hat{\theta}] - \theta)^2}{b_\theta(\hat{\theta})} + \frac{2E_\theta[\hat{\theta} - E_\theta[\hat{\theta}]]}{b_\theta(\hat{\theta})} = 0$$

$$\Rightarrow \text{var}_\theta(\hat{\theta}) + b_\theta^2(\hat{\theta})$$

The above relation implies that in general, for specified MSE, there always exists a “bias-variance tradeoff,” at least for good estimators: any reduction in bias comes at the expense of an increase in variance.

We now get down to the business of defining some general procedures for designing good estimators of non-random parameters. Two important classes of estimation procedures we will consider are:

* method of moments
* maximum likelihood

4.4.2 METHOD OF MOMENTS (MOM) SCALAR ESTIMATORS

The method of moments is a very natural procedure which consists in finding the parameter that attains the best match between empirically computed moments and ensemble moments. Specifically, for positive integer k let $m_k = m_k(\theta)$ be the k-th order ensemble moment of $f(x; \theta)$:

$$m_k = E_\theta[X^k] = \int x^k f(x; \theta)dx.$$
What if we could find a set of $K$ moments such that some vector function $h$ could be found that satisfies

$$\theta = h(m_1(\theta), \ldots, m_K(\theta)).$$

For example, let’s say we could compute a closed form expression $g(\theta)$ for the $k$-th ensemble moment $E_\theta[X^k]$ and found that the function $g$ was invertible. Then if someone only reported the value $m_k$ of this ensemble moment without specifying the $\theta$ for which it was computed we could recover $\theta$ by applying the inverse function

$$\theta = g^{-1}(m_k).$$

Since $g^{-1}$ recovers $\theta$ from the ensemble moment of $X$, if we only have access to an i.i.d. sample $X_1, \ldots, X_n$ from $f(x; \theta)$ it makes sense to estimate $\theta$ by applying $g^{-1}$ to an estimated moment such as the empirical average

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^n X_i^k,$$

yielding the estimator

$$\hat{\theta} = g^{-1}(\hat{m}_k).$$

In many cases it is difficult to find a single ensemble moment that gives an invertible function of $\theta$. Indeed, using only the $k$-th moment we may only be able to find a constraint equation $g(\theta) = \hat{m}_k$ that gives several possible solutions $\hat{\theta}$. In these cases, one can sometimes compute other ensemble and empirical moments to construct more constraint equations and force a unique solution. We will explore this approach in the examples below. Next we give some important asymptotic optimality properties of MOM estimators (see Serfling [72] for proofs).

**IMPORTANT PROPERTIES OF MOM ESTIMATORS**

When the moments $m_k$ are smooth functions of the parameter $\theta$ and an inverse function $g^{-1}$, described above, exists:

1. MOM estimators are asymptotically unbiased as $n \to \infty$

2. MOM estimators are consistent

Note that MOM estimators are not always unbiased in the finite sample regime. There are, however, some inherent difficulties that one sometimes encounters with MOM which are summarized below.

1. MOM estimator is not unique, i.e., it depends on what order moment is used.

2. MOM is inapplicable in cases where moments do not exist (e.g. Cauchy p.d.f.) or are unstable.

An alternative to MOM which can sometimes circumvent the existence problem is to match sample and ensemble fractional moments $m_k$ where $k$ is a positive rational number less than one. Fractional moments can exist when integer moments do not exist and can be quite useful in these situations [73].

Let’s do some examples.

**Example 14** $X$ i.i.d. Bernoulli random variables
Bernoulli measurements arise anytime one deals with (binary) quantized versions of continuous variables, e.g., thresholded radar signals (“radar return is above or below a threshold”), failure data, or digital media, e.g., Internet measurements. In these cases the parameter of interest is typically the probability of success, i.e., the probability that the measured variable is a “logical 1.”

The model is that \( X = [X_1, \ldots, X_n] \) are i.i.d. with
\[
X_i \sim f(x; \theta) = \theta^x (1 - \theta)^{1-x}, \quad x = 0, 1.
\]

Here \( \theta \in [0, 1] \) or, more specifically, \( \theta = P(X_i = 1), 1 - \theta = P(X_i = 0) \).

Objective: find a MOM estimator of \( \theta \)

Note that for any \( k > 0 \) \( E[X_i^k] = P(X_i = 1) = \theta \) so that all moments are identical and the function \( g \) mapping moments to \( \theta \) is the identity map. Thus a MOM estimator of \( \theta \) is simply sample mean:
\[
\hat{\theta} = \bar{X}.
\]

It is obvious that \( \hat{\theta} \) is unbiased since \( E_\theta[\bar{X}] = m_1 = \theta \). Furthermore, it has variance taking a maximum at \( \theta = \frac{1}{2} \) (Fig. 26)
\[
\text{var}_\theta(\bar{X}) = \frac{(m_2 - m_1^2)}{n} = \frac{\theta (1 - \theta)}{n}.
\]

Reiterating, for this Bernoulli example the order of the moment used in the moment matching process leads to identical MOM estimators. This behavior of MOM is very unusual.

**Example 15** \( X \) i.i.d. Poisson random variables

Poisson measurements are ubiquitous in many scenarios where there are counting measurements. For example, in positron emission tomography (PET) the decay of an isotope in a particular spatial
location within a patient’s body produces a gamma ray which is registered as a single ”count” on a detector. The temporal record of the times at which these counts are registered on the detector forms a Poisson process\cite{75}. The total number of counts registered over a finite time interval is a Poisson random variable with rate parameter determined by the mean concentration of the isotope. The objective of a PET system is to reconstruct, i.e., estimate, the distribution of rates over the imaging volume. The Poisson distribution is also frequently used as a model for the number of components or degrees of freedom generating the measured values. For example, the number of molecules in a mass spectroscopy measurement, the number of atoms in a molecule, or the number of targets in a cell detected by a radar.

Again assuming i.i.d. measurements, the model for each data sample is

$$X_i \sim p(x; \theta) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 0, 1, 2, \ldots,$$

where $\theta > 0$ is the unknown rate. It is readily verified that the mean $m_1$ is equal to $\theta$. Therefore, like in the Bernoulli example a MOM estimator of $\theta$ is the sample mean

$$\hat{\theta}_1 = \bar{X}.$$

Alternatively, as the second moment satisfies $m_2 = \theta + \theta^2$, another MOM estimator is the (positive) value of $\hat{\theta}_2$ which satisfies the equation : $\hat{\theta}_2 + \hat{\theta}_2^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 := \bar{X}^2$, i.e.

$$\hat{\theta}_2 = \frac{-1 \pm \sqrt{1 + 4\bar{X}^2}}{2}.$$

As yet another example, we can express $m_2$ as $m_2 = \theta + m_1^2$ or $\theta = m_2 - m_1^2 = \text{var}(X_i)$. Hence, a MOM estimator is

$$\hat{\theta}_3 = \bar{X}^2 - \bar{X}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Among all of these MOM estimators only the sample mean estimator is unbiased for finite $n$:

$$E_\theta(\hat{\theta}_1) = \theta, \quad \text{var}_\theta(\hat{\theta}_1) = \theta/n,$$

$$E_\theta(\hat{\theta}_3) = \frac{n-1}{n} \theta, \quad \text{var}_\theta(\hat{\theta}_3) \approx (2\theta^2 + \theta)/n.$$

Closed form expressions for bias and variance of $\hat{\theta}_2$ do not exist.

You should notice that $\hat{\theta}_1$ compares favorably to $\hat{\theta}_3$ since it has both lower bias and lower variance. We make the following observations.

1. $\hat{\theta}_1$ is unbiased for all $n$.
2. $\hat{\theta}_2, \hat{\theta}_3$ are asymptotically unbiased as $n \to \infty$.
3. Consistency of $\hat{\theta}_1$ and $\hat{\theta}_3$ is directly verifiable from the above expressions for mean and variance and Thebychev’s inequality.
4.4.3 MAXIMUM LIKELIHOOD (ML) SCALAR ESTIMATORS

Maximum likelihood (ML) is arguably the most commonly adopted parametric estimation principle in signal processing. This is undoubtedly due to the fact that, unlike other methods, ML usually results in unique estimators and is straightforward to apply to almost all problems.

For a measurement $X = x$ we define the “likelihood function” for $\theta$

$$L(\theta) = f(x; \theta)$$

and the log-likelihood function

$$l(\theta) = \ln f(x; \theta).$$

These should be viewed as functions of $\theta$ for a fixed value of $x$ (Fig. 27). Readers may find it strange that the $x$-dependence of the functions $L(\theta)$ and $l(\theta)$ is not indicated explicitly. This convention of dropping such dependencies to clarify the “working” variable $\theta$ is common in statistics and signal processing.

![Figure 27: The likelihood function for $\theta$](image)

The ML estimator $\hat{\theta}$ is defined as the value of $\theta$ which causes the data $x$ to become "most likely," i.e., $\hat{\theta}$ makes it most likely that $x$ was generated from $f(x; \theta)$. Mathematically, we have the equivalent definitions

$$\hat{\theta} = \arg\max_{\theta \in \Theta} f(X; \theta)$$

$$= \arg\max_{\theta \in \Theta} L(\theta)$$

$$= \arg\max_{\theta \in \Theta} l(\theta).$$

In fact the ML estimate can be found by maximizing any monotone increasing function of $L(\theta)$. 
Important properties of ML estimators for smooth likelihoods (Ibragimov and Has’minskii [32], Serfling [72]) are

Property 1. MLE’s are asymptotically unbiased. The proof requires additional technical conditions.

Property 2. MLE’s are consistent. The proof requires additional technical conditions.

Property 3. Unlike many other estimators, e.g. MAP and UMVUE estimators, MLE’s are invariant to any transformation of the parameters, i.e.,

\[ \varphi = g(\theta) \Rightarrow \hat{\varphi} = g(\hat{\theta}). \]

This is easy to see for monotone transformations (Fig. 28) but in fact it applies to arbitrary transformations (See exercises).

![Figure 28: Invariance of MLE to functional transformation g](image)

Property 4: MLE’s are asymptotically UMVU in the sense that

\[ \lim_{n \to \infty} n \text{var}_{\theta}(\hat{\theta}) = \frac{1}{F_1(\theta)}, \]

where \( F_1 \) is a quantity known as the Fisher information, which will be introduced soon, and \( 1/F_1 \) specifies the fastest possible asymptotic rate of decay of any unbiased estimator’s variance. The proof requires additional technical conditions.

Property 5: MLE’s are asymptotically Gaussian in the sense

\[ \sqrt{n}(\hat{\theta}_n - \theta) \to Z, \quad (i.d.) \]

where \( Z \sim N(0, 1/F_1(\theta)) \). Here the notation i.d. denotes convergence in distribution. This means that the cumulative distribution function (cdf) of \( \sqrt{n}(\hat{\theta}_n - \theta) \) converges to the (standard normal) cdf of \( Z \). The proof requires additional technical conditions.
Property 6: The MLE is equivalent to the MAP estimator for a uniform prior $f(\theta) = c$.

Property 7: If the MLE is unique, the MLE is a function of the data only through the sufficient statistic.

Now let’s go back and revisit our MOM examples with the MLE in mind.

**Example 16** $X$ i.i.d. Bernoulli random variables

We can solve for the MLE in two ways: (1) considering the entire observation $X$; and (2) considering only a sufficient statistic $T(X)$.

1. With the entire observation $X = x$ the likelihood function is the product

$$L(\theta) = f(x; \theta) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i}.$$  

It is convenient to rewrite this in the form

$$L(\theta) = \theta^{\sum_{i=1}^{n} x_i} (1 - \theta)^{n - \sum_{i=1}^{n} x_i} = \theta^{n \overline{X} - \sum_{i=1}^{n} x_i} (1 - \theta)^{n - n \overline{X}}.$$  

(36)

As this function smooth and concave in $\theta$, differentiation with respect to $\theta$ yields a stationary point condition, the "ML equation," for the MLE $\hat{\theta}$

$$0 = \frac{\partial}{\partial \theta} f(x; \hat{\theta}) = n \left[ \frac{(1 - \hat{\theta}) \overline{X} - \hat{\theta} (1 - \overline{X})}{\hat{\theta} (1 - \hat{\theta})} \right] f(x; \hat{\theta}).$$

Solving the equation $(1 - \hat{\theta}) \overline{X} - \hat{\theta} (1 - \overline{X}) = 0$ we obtain the MLE

$$\hat{\theta} = \overline{X},$$

(37)

which is identical to the MOM estimator obtained above.

2. Using the Fisher factorization (24) on the p.d.f. (36) of $X$ it is easily seen that $T(X) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for $\theta$. The distribution of $T$ is binomial with parameter $\theta$:

$$f_T(t; \theta) = \binom{n}{t} \theta^t (1 - \theta)^{n-t}, \quad t = 0, \ldots, n,$$

where the subscript $T$ on the density of $T$ is to clarify that this is the p.d.f. of the r.v. $T$. Identification of $t = n \overline{X}$ reveals that this is of exactly the same form, except for a constant multiplication factor, as (36). The ML equation is therefore the same as before and we obtain the identical MLE estimator (37).

**Example 17** $X$ i.i.d. Poisson random variables

To find the MLE of the rate parameter $\theta$ express the density of the samples as:

$$f(x; \theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-\theta}.$$
The likelihood function \( L(\theta) = f(x; \theta) \) has to be maximized over \( \theta \) to produce the MLE. It is more convenient to deal with the log likelihood

\[
\hat{\theta}_{ml} = \arg\max_{\theta > 0} \ln L(\theta)
\]

and we have

\[
l(\theta) = \ln f(x; \theta) = \ln \prod_{k=1}^{n} \frac{\theta^{x_k}}{x_k!} e^{-\theta}
\]

\[
= \sum_{k=1}^{n} x_k \ln \theta - n\theta - \sum_{k=1}^{n} \ln x_k!
\]

\[
= \bar{x} \ln \theta - n\theta + c,
\]

where \( c \) is an irrelevant constant.

It is easily verified (look at second derivative) that the log-likelihood \( l(\theta) \) is a smooth strictly concave function of \( \theta \). Thus the MLE is the unique solution \( \theta = \hat{\theta} \) of the equation

\[
0 = \frac{\partial \ln f}{\partial \theta} = \frac{n \bar{x}}{\theta} - n.
\]

We find that the MLE is identical to the first MOM estimator we found for this problem:

\[
\hat{\theta} = \bar{X},
\]

which we know is unbiased with variance equal to \( \theta \).

Let’s check the asymptotic Gaussian property of the MLE for Examples 16 and 17. Write

\[
\sqrt{n}(\bar{X} - \theta) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta) \right)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \theta).
\]

By the central limit theorem (CLT), this converges in distribution to a Gaussian r.v.

\[
E_{\theta}[\sqrt{n}(\bar{X} - \theta)] = 0 \\
\text{var}_{\theta}(\sqrt{n}(\bar{X} - \theta)) = \theta.
\]

## 4.4.4 SCALAR CRAMÉR-RAO BOUND (CRB) ON ESTIMATOR VARIANCE

The CRB can be defined for both random and non-random parameters. However the CRB is more useful for non-random parameters as it can be used to establish optimality or near optimality of an unbiased candidate estimator. Unlike the non-random case, for random parameters the optimal estimator and its MSE are functions of the known joint density of \( \theta \) and \( \bar{X} \). Thus there exist more accurate alternatives to the CRB for approximating estimator MSE, most of which boil down to approximating an integral representation of the minimum mean squared error. We therefore focus
our energies on the CRB for non-random parameters - the interested reader can refer to [84] for the random case.

**The Cramér-Rao Lower Bound** Let $\theta \in \Theta$ be a non-random scalar and assume:

1. $\Theta$ is an open subset, e.g. $(a,b)$, of $\mathbb{R}$.
2. $f(x;\theta)$ is smooth (Ibragimov and Has’minskii [32]) and differentiable in $\theta$.

The following is the Cramér-Rao bound for scalar $\theta$

For any unbiased estimator $\hat{\theta}$ of $\theta$

\[
\text{var}_\theta(\hat{\theta}) \geq 1/F(\theta),
\]

where "=" is attained iff for some non-random scalar $k_\theta$

\[
\frac{\partial}{\partial \theta} \ln f(x;\theta) = k_\theta(\hat{\theta} - \theta).
\]

Here $k_\theta$ is a constant that can depend on $\theta$ but not on $x$. When the CRB is attainable it is said to be a tight bound and (39) is called the CRB tightness condition.

In the CRB $F(\theta)$ is the Fisher information which can be shown [84] to take on either of the following two equivalent forms:

\[
F(\theta) = E_\theta \left[ \left( \frac{\partial}{\partial \theta} \ln f(X;\theta) \right)^2 \right]
= -E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln f(X;\theta) \right]
\]

This latter second derivative form of the Fisher information can be used to show that the scalar $k_\theta$ in the tightness condition (39) is in fact equal to $F(\theta)$. To see this simply differentiate both sides of the equation (39), take expectations, and use the fact that $\hat{\theta}$ is unbiased.

Before going on to some examples, we provide a simple derivation of the scalar CRB here. A more detailed proof of the more general vector parameter CRB will be given later. There are three steps to the derivation of the scalar CRB - assuming that interchange of the order of integration and differentiation is valid. The first step is to notice that the mean of the derivative of the log-likelihood is equal to zero:

\[
E_\theta[\partial \ln f_\theta(X)/\partial \theta] = E_\theta \left[ \frac{\partial f_\theta(X)/\partial \theta}{f_\theta(X)} \right]
= \int \frac{\partial}{\partial \theta} f_\theta(x) dx
= \frac{\partial}{\partial \theta} \int f_\theta(x) dx
= 0
\]
The second step is to show that the correlation between the derivative of the log-likelihood and the estimator is a constant:

\[
E_{\theta}[(\hat{\theta}(X) - E_{\theta}[\hat{\theta}])(\partial \log f_{\theta}(X)/\partial \theta)] = \int (\hat{\theta}(x) - E_{\theta}[\hat{\theta}]) \frac{\partial}{\partial \theta} f_{\theta}(x) \, dx
\]

\[
= \frac{\partial}{\partial \theta} \int \hat{\theta}(x) f_{\theta}(x) \, dx = E_{\theta}[\hat{\theta}] = \frac{1}{E_{\theta}[\hat{\theta}] = \theta}
\]

Where we have used the result of step 1 in line 2 above. Finally, apply the Cauchy-Schwarz (CS) inequality \(E^2[UV] \leq E[U^2]E[V^2]\) to obtain:

\[
1 = E_{\theta}^2[(\hat{\theta}(X) - E_{\theta}[\hat{\theta}])(\partial \ln f_{\theta}(X)/\partial \theta)] \\
\leq E_{\theta}[(\hat{\theta}(X) - E_{\theta}[\hat{\theta}])^2] \cdot E_{\theta}[(\partial \ln f_{\theta}(X)/\partial \theta)^2] \\
= \text{var}_{\theta}(\hat{\theta}) \cdot F(\theta).
\]

Equality occurs in the CS inequality if and only if \(U = kV\) for some non-random constant \(k\). This gives (38) and completes the derivation of the CRB.

To illustrate the CRB let’s go back and reconsider one of the previous examples.

**Example 18 CRB for the Poisson rate**

Assume again that \(X = [X_1, \ldots, X_n]\) is a vector of i.i.d. Poisson random variables

\[
X_i \sim f(x; \theta) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 0, 1, 2, \ldots
\]

To find the CRB we must first compute the Fisher information. Start with

\[
\ln f(x; \theta) = \sum_{k=1}^{n} x_k \ln \theta - n \theta - \sum_{k=1}^{n} \ln x_k!, \text{ constant in } \theta
\]

and differentiate twice

\[
\frac{\partial \ln f(x; \theta)}{\partial \theta} = \frac{1}{\theta} \sum_{k=1}^{n} x_k - n \quad (40)
\]

\[
\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} = -\frac{1}{\theta^2} \sum_{k=1}^{n} x_k. \quad (41)
\]

Therefore, as \(E[\sum_{k=1}^{n} X_k] = n\theta\), the Fisher information given the \(n\) i.i.d. samples is

\[
F_n(\theta) = \frac{n}{\theta}.
\]

The CRB asserts that for any unbiased estimator of the Poisson rate \(\theta\)

\[
\text{var}_{\theta}(\hat{\theta}) \geq \frac{\theta}{n}.
\]
It is useful to make the following key observations.

Observation 1: From example (15) we know that the sample mean $\overline{X}$ is unbiased and has $\text{var}_\theta(\overline{X}) = \theta/n$. This is equal to the CRB and we conclude the CRB is tight.

Observation 2: In fact we could have concluded by inspection that the unbiased estimator $\overline{X}$ achieves the CRB; i.e., without having to explicitly compute its variance and compare to one over the Fisher information. This follows from the fact that equation (40) implies that the CRB tightness condition (39) is satisfied:

$$\frac{\partial}{\partial \theta} \ln f(X; \theta) = \frac{1}{\theta} \sum_{k=1}^{n} X_k - n = \frac{n}{\theta} (\overline{X} - \theta).$$

Furthermore, once tightness is established in this fashion the variance of $\overline{X}$ can be computed by computing the CRB. This indirect method can sometimes be simpler than direct computation of estimator variance.

Observation 3: the expectation of the right hand side of (42) is zero since $\hat{\theta}$ is unbiased. This implies that

$$E_\theta [\partial \ln f(X; \theta) / \partial \theta] = 0.$$ 

The interpretation is that the gradient at $\theta$ of the log-likelihood is an unbiased estimator of zero when $\theta$ is the true parameter, i.e. the parameter appearing in the subscript of the expectation. This relation is generally true: it holds for any density satisfying the differentiability and smoothness conditions [32]) sufficient for existence of the CRB.

GENERAL PROPERTIES OF THE SCALAR CRB

Property 1. The Fisher information is a measure of the average (negative) curvature of the log likelihood function $\ln f(x; \theta)$ near the true $\theta$ (Kass and Voss [38]) (Fig. 30).

---

Figure 29: The curvature of the log likelihood function $\ln f(x; \theta)$ in the vicinity of true $\theta$. 

---
Property 2. Let $F_n(\theta)$ be the Fisher information for a sample of $n$ i.i.d. measurements $X_1, \ldots, X_n$. Then

$$F_n(\theta) = n F_1(\theta).$$

Hence, for smooth likelihood functions of continuous parameters, and unbiased estimators, the variance $\text{var}_\theta(\hat{\theta})$ cannot decay faster than order $1/n$.

Proof of Property 2:

Since $X = [X_1, \ldots, X_n]^T$ are i.i.d.

$$f(x; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

so that

$$F_n(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \ln f(X; \theta) \right]$$

$$= -E \left[ \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta) \right]$$

$$= \sum_{i=1}^{n} -E \left[ \frac{\partial^2}{\partial \theta^2} \ln f(X_i; \theta) \right]_{F_1(\theta)}$$

For unbiased estimators, the CRB specifies an unachievable region of variance as a function of $n$ (Fig. 30). Good unbiased estimators $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ of scalar continuous parameters have variance that behaves as $\text{var}_\theta(\hat{\theta}) = O(1/n)$.

Property 3. If $\hat{\theta}$ is unbiased and $\text{var}_\theta(\hat{\theta})$ attains the CRB for all $\theta$, $\hat{\theta}$ is said to be an efficient estimator. Efficient estimators are always UMVU (but not conversely, e.g., see counterexample in [67, Ch 9]). Furthermore, if an estimator is asymptotically unbiased and its variance decays with optimal rate constant

$$\lim_{n \to \infty} b_\theta(\hat{\theta}) = 0, \quad \lim_{n \to \infty} n \text{var}_\theta(\hat{\theta}) = 1/F_1(\theta),$$

where $F_1$ is the Fisher information given a single sample $X_i$, then $\hat{\theta}$ is said to be asymptotically efficient.

Exponential families play a special role with regard to efficiency. In particular, if $X$ is a sample from a density in the exponential family with scalar parameter $\theta$ having the mean value parameterization (recall discussion in Sec. 3.6.1) then (See exercise 4.32)

$$\theta = E_\theta[t(X)]$$

$$F(\theta) = 1/\text{var}_\theta(t(X)),$$

where $F(\theta)$ is the Fisher information given the sample $X$. Therefore, if one has an i.i.d. sample $X = [X_1, \ldots, X_n]^T$ from such a density then $\hat{\theta} = n^{-1} \sum_{i=1}^{n} t(X_i)$ is an unbiased and efficient estimator of $\theta$. 

Somewhat surprisingly, the next property states that efficient estimators can exist only when the sample comes from an exponential family with mean value parameterization.

Property 4. Efficient estimators for $\theta$ can only exist when the underlying model is in an exponential family, defined in Sec. 3.6.1:

$$f(x; \theta) = a(\theta)b(x)e^{-c(\theta)t(x)}.$$ 

and when $E_\theta[t(X)] = \theta$, i.e., the density is in its mean value parameterization.

Proof of Property 4:

Without loss of generality we specialize to the case of a single sample $n = 1$ and $\Theta = (-\infty, \infty)$. Recall the condition for equality in the CR bound to be achieved by an estimator $\hat{\theta}$ is that the p.d.f. be expressible as

$$\frac{\partial}{\partial \theta} \ln f(x; \theta) = k_\theta(\hat{\theta} - \theta).$$

For fixed $\theta_o$, integrate the LHS of condition (45) over $\theta \in [\theta_o, \theta']$

$$\int_{\theta_o}^{\theta'} \frac{\partial}{\partial \theta} \ln f(x; \theta)d\theta = \ln f(x; \theta') - \ln f(x; \theta_o).$$

On the other hand, integrating the RHS of the condition

$$\int_{\theta_o}^{\theta'} k_\theta(\hat{\theta} - \theta)d\theta = \hat{\theta} \int_{\theta_o}^{\theta'} k_\theta d\theta - \int_{\theta_o}^{\theta'} k_\theta \theta d\theta.$$
Or combining the integrals of RHS and LHS of (45)

\[ f(x; \theta) = \frac{e^{-\phi(x)}}{a(\theta)} \int_{b(x)}^{t(x)} e^{\phi(\theta)} \, d\theta. \]

We illustrate the above properties with two more examples.

**Example 19** Parameter estimation for the exponential density.

A non-negative random variable \( X \) has an exponential density with mean \( \theta \) if its p.d.f. is of the form \( f(x; \theta) = \frac{1}{\theta} \exp(-x/\theta) \) where \( \theta > 0 \). The exponential random variable is commonly used as a model for service time or waiting time in networks and other queuing systems. You can easily verify that this density is in the exponential family specified by \( a(\theta) = \theta^{-1} \), \( b(x) = 1_{[0, \infty)}(x) \), \( c(\theta) = -\theta^{-1} \) and \( t(x) = x \). As \( E_\theta[X] = \theta \) the p.d.f. \( f(x; \theta) \) is in its mean value parametrization and we conclude that the sample mean \( \bar{X} \) is an unbiased estimator of \( \theta \). Furthermore, it is efficient and therefore UMVU when \( n \) i.i.d. observations \( X = [X_1, \ldots, X_n]^T \) are available.

**Example 20** \( \bar{X} \) i.i.d., \( X_i \sim \mathcal{N}(\theta, \sigma^2) \)

The Gaussian "bell curve" distribution arises in so many applications that it has become a standard model. Use of this model is usually justified by invocation of the Central Limit Theorem as describing the measurements, or measurement noise, as the sum of many small contributions, e.g. random atomic collisions, scattered light, aggregation of repeated measurements.

Our first objective will be to find the MLE and CRB for estimating the mean \( \theta \) of univariate Gaussian with known variance \( \sigma^2 \). As the Gaussian with unknown mean is in the exponential family we could take the same approach as above to find efficient estimators. But let’s spice things up and follow an alternative route of trying to tease an efficient estimator out of the tightness condition in the CRB.

\[ f(x; \theta) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \theta)^2 \right\}. \]

Or

\[ \ln f(x; \theta) = -\frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \theta)^2 + c, \]

where \( c \) is constant. Compute the first derivative

\[ \frac{\partial \ln f}{\partial \theta} = \frac{1}{\sigma^2} \sum_{k=1}^{n} (x_k - \theta) \]

\[ = \frac{n}{\sigma^2} \left( \frac{1}{n} \sum_{k=1}^{n} x_k - \theta \right). \]
Thus the CRB tightness condition (39) is satisfied and we can identify, once again, the sample mean \( \bar{x} \) as the optimal estimator of the common mean of a Gaussian sample.

We take another derivative of the log-likelihood with respect to \( \theta \) and invert it to verify what we already knew about the variance of the sample mean

\[
\text{var}_\theta(\bar{X}) = 1/F_n(\theta) = \sigma^2/n.
\]

The first inequality is only true since we know that \( \bar{X} \) is efficient.

Note that the leading factor in the tight CRB condition (46) is: \( k_\theta = \text{var}_\theta^{-1}(\bar{X}) \). This is always true for efficient estimators when \( k_\theta \) does not depend on \( \theta \).

### 4.5 ESTIMATION OF MULTIPLE NON-RANDOM PARAMETERS

We now turn the more general problem of many unknown deterministic parameters. This problem is quite different from the previously studied case of multiple random parameters since there is no joint posterior density to marginalize. First we arrange all unknown parameters in a vector:

\[
\theta = [\theta_1, \ldots, \theta_p]^T,
\]

and state the problem as finding a vector valued estimator \( \hat{\theta} \) of \( \theta \).

The joint density for the measurements \( X \) is written as:

\[
f(x; \theta_1, \ldots, \theta_p) = f(x; \theta).
\]

#### POSSIBLE ESTIMATOR PERFORMANCE CRITERIA

As for a scalar estimator we define the vector estimator bias vector:

\[
b_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta,
\]

and the symmetric estimator covariance matrix:

\[
\text{cov}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E[\hat{\theta}])(\hat{\theta} - E[\hat{\theta}])^T] = \\
\begin{bmatrix}
\text{var}_\theta(\hat{\theta}_1) & \text{cov}_\theta(\hat{\theta}_1, \hat{\theta}_2) & \cdots & \text{cov}_\theta(\hat{\theta}_1, \hat{\theta}_p) \\
\text{cov}_\theta(\hat{\theta}_2, \hat{\theta}_1) & \text{var}_\theta(\hat{\theta}_2) & \cdots & \cdots \\
\vdots & \ddots & \ddots & \cdots \\
\text{cov}_\theta(\hat{\theta}_p, \hat{\theta}_1) & \cdots & \cdots & \text{var}_\theta(\hat{\theta}_p)
\end{bmatrix}.
\]

This matrix is often referred to as the variance-covariance matrix.

In many cases, only the diagonal entries of the estimator covariance matrix, i.e. the component estimator variances, will be of interest. However, as we will soon see, the entire estimator covariance matrix is very useful for generalizing the scalar parameter CRB.

We can also define the estimator concentration:

\[
P_\theta(||\hat{\theta} - \theta|| > \epsilon) = \int_{||\hat{\theta} - \theta|| > \epsilon} f(\hat{\theta}; \theta) d\hat{\theta}
\]

\[
= \int_{\{x: ||\hat{\theta}(x) - \theta|| > \epsilon\}} f(x; \theta) dx
\]

The first order of business is to extend the CRB to vector parameters, called the matrix CRB.
4.5.1  MATRIX CRAMÉR-RAO BOUND (CRB) ON COVARIANCE MATRIX

Let \( \theta \in \Theta \) be a \( p \times 1 \) vector and assume:

1. \( \Theta \) is an open subset of \( \mathbb{R}^p \)
2. \( f(x; \theta) \) is smooth \([32]\) and differentiable in \( \theta \)
3. \( \text{cov}_\theta(\hat{\theta}) \) and \( F(\theta) \) (defined below) are non-singular matrices

The matrix CRB for vector valued parameters is the following. For any unbiased estimator \( \hat{\theta} \) of \( \theta \)

\[
\text{cov}_\theta(\hat{\theta}) \geq F^{-1}(\theta),
\]  

(47)

where “=” is attained iff the following is satisfied for some non-random matrix \( K_\theta \)

\[
K_\theta \nabla_\theta \ln f(X; \theta) = \hat{\theta} - \theta.
\]  

In the case that this tightness condition (48) is satisfied \( \hat{\theta} \) is said to be an efficient vector estimator.

In the matrix CRB (47) \( F(\theta) \) is the Fisher information matrix, which takes either of two equivalent forms,

\[
F(\theta) = E \left[ (\nabla_\theta \ln f(X; \theta)) (\nabla_\theta \ln f(X; \theta))^T \right],
\]

\[
= -E \left[ \nabla^2_\theta \ln f(X; \theta) \right].
\]

where we have defined the gradient operator

\[
\nabla_\theta = \left[ \frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_p} \right]^T,
\]

and the symmetric Hessian (curvature) operator

\[
\nabla^2_\theta = \begin{bmatrix}
\frac{\partial^2}{\partial \theta_1^2} & \frac{\partial^2}{\partial \theta_1 \theta_2} & \cdots & \frac{\partial^2}{\partial \theta_1 \theta_p} \\
\frac{\partial^2}{\partial \theta_2 \theta_1} & \frac{\partial^2}{\partial \theta_2^2} & \cdots & \frac{\partial^2}{\partial \theta_2 \theta_p} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2}{\partial \theta_p \theta_1} & \frac{\partial^2}{\partial \theta_p \theta_2} & \cdots & \frac{\partial^2}{\partial \theta_p^2}
\end{bmatrix}.
\]

The matrix CR Bound (47) has a few more properties than the scalar CRB.

Property 1: The inequality in the matrix bound should be interpreted in the sense of positive definiteness. Specifically if \( A, B \) are \( p \times p \) matrices

\[
A \geq B \iff A - B \succeq 0,
\]

where \( A - B \succeq 0 \) means \( A - B \) is non-negative definite. This means that, in particular,

\[
z^T (A - B) z \geq 0
\]

for any vector \( z \in \mathbb{R}^p \), and all eigenvalues of \( A - B \) are non-negative. For example, choosing \( z = [1, 0, \ldots, 0]^T \) and \( z = [1, 1, \ldots, 1]^T \), respectively, \( A \geq B \), \( A \geq B \) implies both

\[
a_{ii} \geq b_{ii}, \quad \text{and} \quad \sum_{i,j} a_{ij} \geq \sum_{i,j} b_{ij}.
\]
However, $A \succeq B$ does NOT mean $a_{ij} \geq b_{ij}$ in general. A simple counterexample is constructed as follows. Let $0 < \rho < 1$ and consider

$$
\begin{bmatrix}
2 & 0 \\
0 & 2
\end{bmatrix} - \begin{bmatrix}
1 & \rho \\
\rho & 1
\end{bmatrix} = \begin{bmatrix}
1 & -\rho \\
-\rho & 1
\end{bmatrix},
$$

which has two eigenvalues $1 - \rho > 0$ and $1 + \rho > 0$. Hence $A - B > 0$ while clearly $a_{12} = 0 \neq \rho$.

Property 2: The matrix inequality (47) implies a scalar CRB on the variance of the $i$-th component of an unbiased vector estimator $\hat{\theta}$

$$\text{var}_\theta(\hat{\theta}_i) \geq [F^{-1}(\theta)]_{ii},$$

where the right hand side (RHS) denotes the $i$-th element along the diagonal of the inverse Fisher information matrix.

Property 3. Fisher information matrix is a measure of the average curvature profile of the log likelihood near $\theta$.

Property 4. Let $F_n(\theta)$ be the Fisher information for a sample of $n$ i.i.d. measurements $X_1, \ldots, X_n$. Then, as in the scalar parameter case,

$$F_n(\theta) = nF_1(\theta).$$

Hence $\text{var}_\theta(\hat{\theta}) = O(1/n)$ is also expected for good estimators of multiple unknown continuous valued parameters.

Property 5. Efficient vector estimators only exist for multiparameter exponential families with mean value parameterization

$$f(x; \theta) = a(\theta)b(x)e^{-\frac{1}{2}c(\theta)^Tc(x)},$$

and

$$E_{\tilde{\theta}}[t(X)] = \tilde{\theta}.$$

Furthermore, in this case $E[n^{-1}\sum_{i=1}^{n} t(X_i)] = \theta$, $\hat{\theta} = n^{-1}\sum_{i=1}^{n} t(X_i)$ is an unbiased efficient estimator of $\theta$.

Property 6. If an estimator $\hat{\theta}$ satisfies

$$\nabla_{\hat{\theta}} \ln f = K_{\hat{\theta}}(\hat{\theta} - \theta),$$

for some non-random matrix $K_{\hat{\theta}}$ then we can immediately conclude:

1. $\hat{\theta}$ is unbiased since, as shown in proof of the multiple parameter CRB;

$$E_{\theta}[\nabla_{\hat{\theta}} \ln f(X; \theta)] = 0,$$

2. $\hat{\theta}$ is efficient and thus its components are UMVU estimators;

3. The covariance of $\hat{\theta}$ is given by the inverse Fisher information $F(\theta)$;

4. $K_{\hat{\theta}}$ is the Fisher information $F(\theta)$ since

$$E_{\theta}[\nabla_{\hat{\theta}} \ln f(X; \theta)] = E_{\theta}[\nabla_{\hat{\theta}} \nabla_{\hat{\theta}} \ln f(X; \theta)] = E_{\theta}[\nabla_{\hat{\theta}} (K_{\hat{\theta}}(\hat{\theta} - \theta))]$$

and, by the chain rule and the unbiasedness of $\hat{\theta}$

$$E_{\theta}[\nabla_{\hat{\theta}} (K_{\hat{\theta}}(\hat{\theta} - \theta))] = \nabla_{\theta}(K_{\hat{\theta}})E_{\theta}[(\hat{\theta} - \theta)] + K_{\hat{\theta}}E_{\theta}[\nabla_{\hat{\theta}} ((\hat{\theta} - \theta))] = -K_{\hat{\theta}}.$$
5. The estimator covariance is
\[
\text{cov}_{\hat{\theta}}(\hat{\theta}) = K_{\theta}^{-1}.
\]

Proof of Matrix CR bound:
There are 3 steps in our derivation, which, with one exception, is a direct generalization of the proof of the scalar CRB: (1) show that the gradient of the log-likelihood is zero mean; (2) the correlation between the gradient of the log-likelihood and estimator is constant; (3) the covariance matrix of the concatenated gradient and estimator error gives a relation between Fisher info and estimator covariance.

Step 1. Show
\[
E_{\hat{\theta}}[\nabla_{\hat{\theta}} \ln f(X; \theta)] = 0.
\]

\[
\Rightarrow \quad E_{\hat{\theta}} \left[ \frac{1}{f(X; \theta)} \nabla_{\theta} f(X; \theta) \right] = \int_{\mathcal{X}} \nabla_{\theta} f(x; \theta) dx = 0.
\]

Step 2. \[
E_{\hat{\theta}} \left[ \nabla_{\hat{\theta}} \ln f(X; \theta) (\hat{\theta} - \theta)^T \right] = I.
\]

First observe
\[
E_{\hat{\theta}} \left[ \nabla_{\hat{\theta}} \ln f(X; \theta) \hat{\theta}^T \right] = E_{\hat{\theta}} \left[ \frac{1}{f(X; \theta)} \nabla_{\theta} f(X; \theta) \hat{\theta}^T \right]
\]
\[
= \int_{\mathcal{X}} \nabla_{\theta} f(x; \theta) \hat{\theta}^T (x) dx
\]
\[
= \nabla_{\theta} \int_{\mathcal{X}} f(x; \theta) \hat{\theta}^T (x) dx
\]
\[
= \int_{\mathcal{X}} f(x; \theta) \hat{\theta}^T (x) dx
\]
\[
= I.
\]

Now putting this together with result of the previous step
\[
E_{\hat{\theta}} \left[ \nabla_{\hat{\theta}} \ln f(X; \theta) (\hat{\theta} - \theta)^T \right]
\]
\[
= E_{\hat{\theta}} \left[ \nabla_{\hat{\theta}} \ln f(X; \theta) \hat{\theta}^T \right] - E_{\hat{\theta}} \left[ \nabla_{\hat{\theta}} \ln f(X; \theta) \right] \theta^T.
\]

Step 3. Define a \(2p \times 1\) random vector \(U\):
\[
U = \begin{bmatrix} \hat{\theta} - \theta \\ \nabla_{\hat{\theta}} \ln f(X; \theta) \end{bmatrix}.
\] (49)

Since any matrix expressed as an outer product of two vectors is non-negative definite
\[
E_{\hat{\theta}}[UU^T] \geq 0.
\]
Using the results of steps 1 and 2, we have

\[ E_2[UU^T] = \begin{bmatrix} \text{cov}_2(\hat{\theta}) & \mathbf{I} \\ \mathbf{I} & \mathbf{F}(\hat{\theta}) \end{bmatrix} \geq 0. \]

It only remains to apply the result of Sec. 2.4 to the above partitioned matrix to see that this implies that

\[ \text{cov}_2(\hat{\theta}) - \mathbf{F}^{-1}(\hat{\theta}) \geq 0. \]

An alternative, and more direct, way to show this is to let \( w \) and \( y \) be arbitrary \( p \)-vectors and define \( v = \begin{bmatrix} w \\ y \end{bmatrix} \). Then, as \( v^T E_2[UU^T] v \geq 0 \),

\[ w^T \text{cov}_2(\hat{\theta}) w + 2w^T y + y^T \mathbf{F}(\hat{\theta}) y \geq 0. \]

Taking \( y = -\mathbf{F}^{-1}(\hat{\theta}) w \) in the above we obtain

\[ w^T \text{cov}_2(\hat{\theta}) w \geq 0. \]

It remains to obtain the tightness condition ensuring equality in the CRB. Note first that if \( \text{cov}_2(\hat{\theta}) = \mathbf{F}^{-1}(\hat{\theta}) \) then \( E_2[UU^T] \) necessarily has rank \( p \) (see exercises at end of chapter). This can only happen if the random vector \( U \) (49) has \( p \) linearly independent components. As \( \text{cov}_2(\hat{\theta}) \) and \( \mathbf{F}(\hat{\theta}) \) have been assumed non-singular, \( \hat{\theta} - \theta \) can have no linear dependencies and neither does \( \nabla_2 \ln f \). Hence it can only be that

\[ \mathbf{K}_2 \nabla_2 \ln f = \hat{\theta} - \theta \]

for some non-random matrix \( \mathbf{K}_2 \). In other words the gradient of the log likelihood lies in the span of the estimator errors.

We move on to generalizations of MOM and ML estimators to the vector parameter case.

### 4.5.2 METHODS OF MOMENTS (MOM) VECTOR ESTIMATION

Let \( m_k = m_k(\theta) \) be the \( k \)-th order moment of \( f(x; \theta) \). The vector MOM estimation procedure involves finding \( K \) moments such that the vector function of \( \hat{\theta} \in \mathbb{R}^p \)

\[ g(\theta) = [m_1(\theta), \ldots, m_K(\theta)] \]

can be inverted, i.e., there exists a unique value \( \hat{\theta} \) satisfying

\[ \hat{\theta} = g^{-1}(m_1, \ldots, m_K). \]

As in the scalar case, the MOM estimator is constructed by replacing \( m_k \) with its empirical estimate

\[ \hat{\theta} = g^{-1}(\hat{m}_1, \ldots, \hat{m}_K), \]

where \( \hat{m}_k = \frac{1}{n} \sum_{i=1}^n X_i^k \).
4.5.3 MAXIMUM LIKELIHOOD (ML) VECTOR ESTIMATION

The vector MLE is an obvious generalization of the scalar MLE
\[ \hat{\theta} = \arg\max_{\theta \in \Theta} f(\mathbf{X}; \theta). \]

For smooth likelihood functions, vector MLEs have several key properties ([32]):
1. Vector MLE’s are asymptotically unbiased;
2. Vector MLE’s are consistent;
3. Vector MLE’s are invariant to arbitrary vector transformations;
\[ \varphi = g(\theta) \Rightarrow \hat{\varphi} = g(\hat{\theta}); \]
4. Vector MLE’s are asymptotically efficient and thus their component estimators are asymptotically UMVU;
5. Vector MLE’s are asymptotically Gaussian in the sense
\[ \sqrt{n}(\hat{\theta}_n - \theta) \to \mathbf{z}, \quad (i.d.) \]
where \( \mathbf{z} \sim \mathcal{N}_p(0, \mathbf{F}_1^{-1}(\theta)) \) and \( \mathbf{F}_1(\theta) \) is the single sample Fisher information matrix
\[ \mathbf{F}_1(\theta) = -E_\theta \left[ \nabla_\theta^2 \log f(\mathbf{X}_1; \theta) \right]. \]

A couple of examples will illustrate these estimators.

**Example 21** Joint estimation of mean and variance in a Gaussian sample

This is an extension of Example 20 to the case where both the mean and the variance are unknown. Assume an i.i.d. sample \( \mathbf{X} = [X_1, \ldots, X_n] \) of Gaussian r.v.s \( X_i \sim \mathcal{N}(\mu, \sigma^2) \). The unknowns are \( \theta = [\mu, \sigma^2] \).

The log-likelihood function is
\[ l(\theta) = \ln f(\mathbf{x}; \theta) = -\frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2 + c. \] (50)

A. MOM approach to estimation:

We know that \( m_1 = \mu, \ m_2 = \sigma^2 + \mu^2 \) and thus
\[ \mu = m_1, \quad \sigma^2 = m_2 - m_1^2. \]

Hence a MOM estimator of \( \theta \) is:
\[ \hat{\theta} = [\hat{\mu}, \hat{\sigma}^2] = [\hat{m}_1, \hat{m}_2 - \hat{m}_1^2] = \left[ \mathbf{X}, \mathbf{X}^2 - \mathbf{X}^2 \right] \]
\[ = \left[ \mathbf{X}, (\mathbf{X} - \overline{\mathbf{X}})^2 \right]. \]
As usual we denote
\[
\bar{X} = n^{-1} \sum_{k=1}^{n} X_k
\]
\[
(X - \bar{X})^2 = n^{-1} \sum_{k=1}^{n} (X_k - \bar{X})^2 = \frac{n - 1}{n} s^2,
\]
and
\[
s^2 = (n - 1)^{-1} \sum_{k=1}^{n} (X_k - \bar{X})^2
\]
is the sample variance.

B. ML approach.

As \( l(\theta) \) (50) is a concave function (verify that \( -\nabla^2_{\theta} \ln f \) is positive definite) we can use the likelihood equation (stationary point condition) for finding \( \theta = \hat{\theta} \)
\[
0 = \nabla_{\theta} \ln f(x; \theta) = \begin{bmatrix}
\frac{1}{\sigma^2} \sum_{k=1}^{n} (x_k - \theta_1) \\
-\frac{n/2}{\sigma^2} + \frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \theta_1)^2
\end{bmatrix}.
\]
Therefore,
\[
\hat{\theta}_1 = \hat{\mu} = \bar{X}, \quad \hat{\theta}_2 = \hat{\sigma^2} = \frac{n-1}{n} s^2,
\]
so that the MLE and MOM estimators are identical.

Let’s consider the performance of the ML/MOM estimator. The bias and covariance are simple enough to compute (recall that in Sec. 3.4 we showed that \( (n-1)s^2/\sigma^2 \) is Chi square distributed with \( n-1 \) degrees of freedom):
\[
\text{E}_\theta[\hat{\mu}] = \mu, \quad \text{E}_\theta[\hat{\sigma^2}] = \left( \frac{n-1}{n} \right) \sigma^2; \quad \text{var}_\theta(\bar{X}) = \sigma^2/n;
\]
and
\[
\text{var}_\theta(\hat{\sigma^2}) = \left( \frac{n-1}{n} \right)^2 \text{var}_\theta(s^2) = 2\sigma^4/n \left( \frac{n-1}{n} \right).
\]
Since the sample mean and sample variance are uncorrelated (recall Sec. 3.4)
\[
\text{cov}_\theta(\hat{\mu}, \hat{\sigma^2}) = \begin{bmatrix}
\sigma^2/n & 0 \\
0 & 2\sigma^4/n \left( \frac{n-1}{n} \right)
\end{bmatrix}, \quad (51)
\]
Next we compute the Fisher information matrix by taking the expectation of the Hessian \( -\nabla^2_{\theta} \ln f(X; \theta) \)
\[
F(\theta) = \begin{bmatrix}
\frac{n}{\sigma^2} & 0 \\
0 & n/(2\sigma^4)
\end{bmatrix}, \quad (52)
\]
giving the CR bound

$$\begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}. \quad (53)$$

Some interesting observations are the following:

Observation 1. MOM and ML estimators derived above have covariances which violate the CR bound (compare the (2,2) elements of matrices (51) and the RHS of (53)). This is not a contradiction since the ML variance estimator is not unbiased!

Observation 2. Consider the bias-corrected estimator of $\mu, \sigma^2$ T

$$\hat{\theta} = [\bar{X}, s^2]^T.$$ This estimator is unbiased. Now, as $s^2 = \left( \frac{n}{n-1} \right) \hat{\sigma}^2$

$$\text{var}_\theta(s^2) = \left( \frac{n}{n-1} \right)^2 \text{var}_\theta(\hat{\sigma}^2),$$

$$\text{cov}_\theta(\hat{\theta}) = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \left( \frac{n}{n-1} \right) \end{bmatrix} \geq \mathbf{F}^{-1}(\theta).$$

We conclude that the bias-corrected estimator’s covariance no longer violates the CRB. Indeed, $\bar{X}$ is efficient estimator of $\mu$ since

$$\text{var}_\theta(\hat{\mu}) = [\mathbf{F}^{-1}]_{11} = \sigma^2/n.$$ 

However, $s^2$ is not an efficient estimator of $\sigma^2$ since

$$\text{var}_\theta(s^2) > [\mathbf{F}^{-1}]_{22}.$$ 

Observation 3. as predicted, the MLE is asymptotically efficient as $n \to \infty$.

$$n\text{cov}_\theta(\hat{\theta}) = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4 \left( \frac{n-1}{n} \right) \end{bmatrix} \rightarrow \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4 \end{bmatrix} = \mathbf{F}^{-1}(\theta).$$

Observation 4. We can also verify that, as predicted, $[\hat{\mu}, \hat{\sigma}^2]$ is asymptotically Gaussian. It suffices to consider the following results:

a) $\mu$ and $\hat{\sigma}^2$ are independent r.v.s;

b) $\sqrt{n}(\hat{\mu} - \mu) = \mathcal{N}(0, \sigma^2)$;

c) $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \sigma^2 \sqrt{n}(\chi^2_{n-1}/(n-1) - 1)$;

d) $\chi^2_{\nu} \sim \mathcal{N}(\nu, 2\nu), \nu \to \infty$.

Observation 5. We can easily manipulate the condition for equality in the CR bound to find an efficient vector estimator (but not of $\theta$ as originally specified!):

$$\nabla_\theta \ln f(X; \theta) = \mathbf{K}_\theta \begin{bmatrix} \bar{X} - \mu \\ \bar{X}^2 - (\sigma^2 + \mu^2) \end{bmatrix},$$
where

\[ K_2 := \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2\mu & 1 \end{bmatrix}^{-1}. \]

As the sample moments are unbiased estimates of the ensemble moments, we conclude that \( \overline{X}, \overline{X^2} \) are efficient estimators of the first moment \( E[X] = \mu \) and second (non-central) moment \( E[X^2] = \sigma^2 + \mu^2 \), respectively.

**Example 22** Joint estimation of mean vector and covariance matrix in a multivariate Gaussian sample

This example is a generalization of the univariate Gaussian example, Example 21, to the multivariate Gaussian case. We only give the final results here, the detailed derivations are given in Ch. 12.

The multivariate Gaussian distribution arises in many problems of signal processing, communications, and machine learning. Sensor array processing was one of the earliest applications of this distribution in signal processing [33]. In the multivariate Gaussian model the measurements are a set of \( n \) i.i.d. \( p \)-dimensional Gaussian vectors, each having mean vector \( \mu \) and \( p \times p \) covariance matrix \( \mathbf{R} \). In sensor array processing each of these Gaussian random vectors is a single snapshot of the output of a \( p \)-element sensor array and information about the directions of signals propagating across the array is encoded in \( \mu \) and \( \mathbf{R} \). As usual we jointly refer to the unknown parameters in \( \mu \) and \( \mathbf{R} \) by the parameter vector \( \theta \). We assume that \( \mathbf{R} \) is positive definite.

Thus the multivariate Gaussian measurements can be considered as a random \( p \times n \) matrix formed from the concatenation of \( n \) i.i.d. columns:

\[ \mathbf{X} = [X_1, \ldots, X_n] \]

where

\[ X_i = \begin{bmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{bmatrix}, \quad i = 1, \ldots, n. \]

In terms of \( X_i \) the mean vector is \( \mu = E[\theta[X_i]] \) and the covariance matrix is \( \mathbf{R} = \text{cov}_{\theta}(X_i) \). Since the columns of \( \mathbf{X} \) are independent its joint density is

\[
 f(\mathbf{X}; \mu, \mathbf{R}) = \left( \frac{1}{(2\pi)^p |\mathbf{R}|} \right)^{n/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^T \mathbf{R}^{-1} (X_i - \mu) \right). \quad (54)
\]

The objective is to estimate the mean and covariance using the measurement matrix \( \mathbf{X} \). The maximum likelihood estimator of the mean is simply the sample mean

\[ \hat{\mu} = n^{-1} \sum_{i=1}^{n} X_i \]

and the maximum likelihood estimator of \( \mathbf{R} \) is

\[ \hat{\mathbf{R}} = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^T. \]
These expressions for the ML estimators are derived in Ch. 12 using trace identities and matrix eigendecompositions to simplify the maximization of (54).

Similarly to the case of univariate Gaussian ML estimates of the mean and variance, these ML estimators are also method of moments estimators. It can also be shown that both estimators are asymptotically unbiased and consistent. However, while $\hat{\mu}$ is unbiased, $\hat{R}$ is biased. A bias corrected version of the ML covariance estimator is the sample covariance matrix

$$S = (n - 1)^{-1} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^T.$$  

Derivation of the CR bound on estimator covariance is more difficult than for the univariate Gaussian case. The principal difficulty is that the elements of the covariance matrix $R$ are redundant and have non-linear dependencies since $R$ is a symmetric positive definite matrix. Nonetheless the CR bound has been derived. It is known as Bang’s formula [3] in sensor array processing.

**Example 23 Joint estimation of class probabilities in a multinomial sample**

Consider an experiment where we measure a discrete valued random variable that can take on one of a number $K$ of possible labels or categories. For such a categorical random variable the actual labels are arbitrary and only the number $K$ of labels is important. Thus we often map the label to the integers $\{1, \ldots, K\}$. The multinomial model specifies the probability distribution of different combinations of labels that are observed in $n$ i.i.d. draws of this random variable.

A common signal processing example where the multinomial model arises is the analog-to-digital (A/D) converter. An A/D converter takes continuous valued input random variable $X$ and quantizes it to one of $K$ levels $a_1, \ldots, a_K$, producing a discrete output $Q(X) \in \{a_1, \ldots, a_K\}$ whose value is the level closest to $X$. When $n$ i.i.d. samples $\{X_i\}_{i=1}^{n}$ are processed through the A/D converter the empirical histogram of the outputs $\{Q(X_i)\}_{i=1}^{n}$ is multinomial distributed.

Another example of the multinomial model arises in the monitoring of computer networks in which a set of routers and terminals are connected by $K$ links. Over a period of time each link in the network may intermittently fail and generate a number of dropped packets. If a packet is dropped independently of other packets the multinomial distribution is a good model for the joint distribution of the vector recording the number dropped packets over each of the $K$ links, the so-called count vector [44].

Yet another example of the multinomial model arises in document indexing and retrieval of text databases. In this context, a document may contain words or other items falling into $K$ possible word classes or categories. Let the number of words from class $k$ be denoted $n_k$. The bag of words model summarizes the document by the word count vector and models this vector as multinomial distributed [50]. A more sophisticated hierarchical model for topically diverse document databases is described in the next example.

Continuing with the computer network example consider performing an experiment where packets are transmitted from a source terminal to a destination terminal in a packet switched network like TCP-IP over the Internet. Assume that for successful transmission the packet must pass through a fixed set of $K$ links along the source-destination path and that each link drops the packet randomly. For the $i$-th transmission define the $K$-element random indicator vector $Z_i$ taking on a single non-zero value equal to “1” in the $k$-th place if the packet was dropped and it was the $k$-th link that dropped it. Assume that each of the links drops the packet with probability $\theta_1, \ldots, \theta_K$. 
respectively, with \( \sum_{k=1}^{K} \theta_k = 1 \). The number of packets dropped by each link is the vector, called the empirical histogram,

\[
N = [N_1, \ldots, N_K]^T = \sum_{i=1}^{n} Z_i.
\]

Assume that the number \( n = \sum_{k=1}^{K} N_k \) of dropped packets is fixed and assume that the \( Z_i \)'s are i.i.d. Under these assumptions the \( N \) is multinomial distributed with parameters \( \theta = [\theta_1, \ldots, \theta_K] \) with probability mass function:

\[
p(N; \theta) = P_{\theta}(N_1 = k_1, \ldots, N_K = k_K) = \frac{n!}{k_1! \cdots k_K!} \theta_1^{k_1} \cdots \theta_K^{k_K},
\]

where \( k_i \geq 0 \) are integers satisfying \( \sum_{i=1}^{K} k_i = n \) and \( \theta_i \in [0, 1] \) are cell probabilities satisfying \( \sum_{i=1}^{K} \theta_i = 1 \).

A MOM estimator of \( \theta \) is obtained by matching the first empirical moment \( N \) to the first ensemble moment \( E_{\theta}[N] = \theta n \). This yields the estimator \( \hat{\theta} = \frac{N}{n} \), or more explicitly

\[
\hat{\theta} = \left[ \frac{N_1}{n}, \ldots, \frac{N_K}{n} \right].
\]

To find the MLE of \( \theta \) we need to proceed with caution. The \( K \) parameters \( \theta \) live in a \( K-1 \) subspace of \( \mathbb{R}^{K} \) due to the constraint \( \sum_{i=1}^{K} \theta_i = 1 \). We can find the MLE either by reparameterization of the problem (see comment at end of this example) or by using Lagrange multipliers. The Lagrange multiplier method will be adopted here.

To account for the constraint we replace the log-likelihood function with the penalized log-likelihood function

\[
J(\theta) = \ln p(N; \theta) - \lambda \left( \sum_{i=1}^{K} \theta_i - 1 \right),
\]

where \( \lambda \) is a Lagrange multiplier which will be selected in order to satisfy the constraint.

Now as \( J \) is smooth and concave we set the gradient of \( J(\hat{\theta}) \) to zero to find the MLE:

\[
0 = \nabla_{\theta} J(\theta) = \nabla_{\theta} \left[ \sum_{i=1}^{K} N_i \ln \theta_i - \lambda \left( \sum_{i=1}^{K} \theta_i - 1 \right) \right] = \left[ \frac{N_1}{\theta_1} - \lambda, \ldots, \frac{N_K}{\theta_K} - \lambda \right].
\]

Thus

\[
\hat{\theta}_i = \frac{N_i}{\lambda}, \quad i = 1, \ldots, K
\]

Finally, we find \( \lambda \) by forcing \( \hat{\theta} \) to satisfy constraint

\[
\sum_{i=1}^{K} N_i / \lambda = 1 \Rightarrow \lambda = \sum_{i=1}^{K} N_i / n.
\]

The solution to this equation gives the MLE and it is identical to the MOM estimator.

Similarly to the previous example the derivation of the CRB is more difficult due to parameter dependencies; recall that the \( \theta_i \)'s sum to one. The CRB can be derived reparameterizing the multinomial probability mass function by the \( K-1 \) linearly independent parameters \( \theta_1, \ldots, \theta_{K-1} \), which determine the remaining parameter by \( \theta_K = 1 - \sum_{i=1}^{K-1} \theta_i \), or by using the theory of constrained CR bounds [23].
Example 24 Multinomial-Dirichlet models for bag-of-words document processing

This type of distribution is commonly used to model categorical variables such as those that occur in document indexing and retrieval of text databases that self-organize into hierarchies of topics. In this context, a document may contain words or other items falling into $K$ possible word classes or categories, $K \geq 2$. Let the number of words from class $k$ in a given document be denoted as $N_k$ and the total number of words in the document as $n = \sum_{k=1}^{K} N_k$. The multinominal bag of words model summarizes the document by the word count vector $N = [N_1, \ldots, N_K]$ and assumes that this vector is multinomial distributed with parameter vector $p = [p_1, \ldots, p_K]$ (denoted by $\theta$ in Example 23).

In a database of $M$ documents, each document will be governed by a different multinomial parameter vector, e.g., $p_l$ for the $l$-th document. Hence, the population of parameter vectors $\{p_l\}_{l=1}^{M}$ might itself be modeled as a set of i.i.d. realizations from a prior distribution $f(p; \alpha)$, where $\alpha = [\alpha_1, \ldots, \alpha_p]$ are hyperparameters that specify the prior. The Dirichlet prior distribution has a particularly simple form

$$f(p; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k-1}$$

where $B(\alpha) = \left(\prod_{k=1}^{K} \Gamma(\alpha_k)\right) / \Gamma\left(\sum_{k=1}^{K} \alpha_k\right)$ is the Beta function and the $\alpha_k$’s are positive. The Multinomial-Dirichlet model is specified by the joint distribution of $N$ and $p$

$$P(N_1 = n_1, \ldots, N_K = n_K | p, \alpha) f(p; \alpha) = \frac{n!}{N_1! \ldots N_K!} \prod_{k=1}^{K} p_k^{N_k + \alpha_k - 1}.$$

The marginal distribution $P(N; \alpha)$ of the word count vector parameterized by $\alpha$ is obtained by integrating the right hand side over $p$. The results takes on a remarkably simple closed form for the marginal due to the fact that the Dirichlet distribution is conjugate to the multinomial distribution:

$$P(N; \alpha) = \frac{n!}{\prod_{k=1}^{K} N_k! \Gamma(n + a)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)},$$

with $a = \sum_{k=1}^{K} \alpha_k$.

For this model the hyperparameters $\alpha$ are assumed known and are generally application dependent, e.g., scientific documents, web pages, and news media documents will each have different $\alpha$’s. The hyperparameters could also be estimated empirically from the entire database, as we show next.

Assuming there are $M$ documents in the database, let $N_i$ denote the word count vector of the $i$-th document and $n_i$ are the total number of words in this document. Assume that, conditioned on $p_l$, $N_i$ is multinomial distributed with parameter $p_l$. Then, marginalizing the joint distribution of all the documents over the $p_l$’s, the likelihood function for $\alpha$ is:

$$P(N_1, \ldots, N_M | \alpha) = \prod_{i=1}^{M} \left( \frac{n_i!}{\prod_{k=1}^{K} N_{ik}! \Gamma(n_i + a)} \prod_{k=1}^{K} \frac{\Gamma(N_{ik} + \alpha_k)}{\Gamma(\alpha_k)} \right),$$

(55)

where $n_i = \sum_{k=1}^{K} N_{ik}$ is the total word count in the $i$’th document.

The multinomial-Dirichlet model is an example of a probabilistic graphical model known as a latent Dirichlet process (LDP) [45]. A topic model takes the LDP approach one step further and models
the $\alpha$ parameters themselves as being random, e.g., from a mixture of Dirichlet distributions to capture clusters of documents into different and unknown topic classes. This approach of putting a prior on hyperparameters can be repeatedly nested to represent deep hierarchies, resulting in a hierarchical Dirichlet processes (HDP) [82]. For example, a database may contain several general topics like science, sports, political news, etc, that each subdivide into subtopics and subsubtopics.

### 4.6 Handling Nuisance Parameters

In many cases only a single parameter $\theta_1$ is of direct interest while the other unknowns $\theta_2, \ldots, \theta_p$ are nuisance parameters which are not of interest. For example, in the Gaussian example with both unknown mean and variance, Example 21, the variance may not be of intrinsic interest. In this example, we found that the estimator covariance is diagonal, which implies that there is no correlation between the mean parameter estimation errors and the variance parameter estimation errors. As we will see below, this means that the variance is a rather benign nuisance parameter since knowledge or lack of knowledge of the variance does not affect the variance of the ML mean estimator. We divide the discussion of nuisance parameters into the cases of random and non-random parameters.

**CASE I: HANDLING RANDOM NUISANCE PARAMETERS:**

For random nuisance parameters the average loss only penalizes $\hat{\theta}_1$’s estimation errors. When all the parameters including $\theta_1$ are random the average loss is:

$$E[c(\hat{\theta}_1, \theta_1)] = \int_{\theta_1} d\theta_1 \int_X dx \ c(\hat{\theta}_1(x), \theta_1) f(x|\theta_1) f(\theta_1).$$

The prior on $\theta_1$ is computed from the prior on $\theta$

$$f(\theta) = \int d\theta_2 \ldots \int d\theta_p \ f(\theta_1, \theta_2, \ldots, \theta_p).$$

The conditional density of $X$ given $\theta_1$ is therefore

$$f(x|\theta_1) = \int d\theta_2 \ldots \int d\theta_p \ f(x|\theta_1, \theta_2, \ldots, \theta_p) f(\theta_2, \ldots, \theta_p|\theta_1),$$

yielding the posterior on $\theta_1$

$$f(\theta_1|x) = \int d\theta_2 \ldots \int d\theta_p \ f(\theta_1, \ldots, \theta_p|x).$$

The maximization of $f(\theta_1|x)$ over $\theta_1$ yields the MAP estimator for random nuisance parameters. When $\theta_1$ is not random then maximizing $f(x|\theta_1)$ in (56) over $\theta_1$ yields the maximum likelihood estimator for random nuisance parameters.

Observe that explicit estimates of the nuisance parameters $\theta_2, \ldots, \theta_p$ are not required to implement the marginalized likelihood (56) or the posterior distribution (57) of $\theta_1$. However, integration (marginalization) of the conditional density over $\theta_2, \ldots, \theta_p$ is required and this may be quite difficult especially when $p$ is large. An exception is when the prior distribution of the nuisance parameters is conjugate to the likelihood function in which case the marginalization yields a closed form expression for (56). Example 24 provides a good illustration for the case that the multinomial parameters $\underline{\theta}$ are nuisance parameters and the Dirichlet parameters $\underline{\alpha}$, governing the population of
document word frequency distributions, are the parameters of interest. In this case the marginal likelihood function (56) for \( \alpha \) has the closed form expression (55).

**CASE II: HANDLING NON-RANDOM NUISANCE PARAMETERS:**

The case of non-random parameters is quite different. The average loss still only penalizes for \( \hat{\theta}_1 \) estimation errors but nonetheless depends on all unknowns:

\[
E_{\theta}[C] = \int_X c(\hat{\theta}_1(x), \theta_1) f(x; \theta) \, dx.
\]

The maximum Likelihood Estimator of \( \theta_1 \) is simply

\[
\hat{\theta}_1 = \arg\max_{\theta_1} \left( \max_{\theta_2, \ldots, \theta_p} \log f(X | \theta_1, \theta_2, \ldots, \theta_p) \right).
\]

As compared to the case of random nuisance parameters, which required integration of the likelihood function over the nuisance parameters, here we require maximization of the likelihood over nuisance parameters. In some but not all cases maximization may be easier than integration. There are also cases where the nuisance parameters do not affect the estimator of the parameter of interest. Sometimes the maximum likelihood estimator of the parameter of interest is not a function of the nuisance parameters and thus no estimation or marginalization of these latter parameters is necessary. The CR bound can be used to explore the effect of nuisance parameters on estimation performance.

**CR BOUND PREDICTIONS FOR NON-RANDOM NUISANCE PARAMETERS**

As before let’s say we are interested in unbiased estimation of only the first entry \( \theta_1 \) in the vector of unknown parameters \( \theta \). Our derivation of the matrix CRB (47) made the explicit assumption that there existed unbiased estimators of all of the parameters. It turns out that this restriction is unnecessary when only \( \theta_1 \) is of interest (see exercises).

Assume that \( \theta = [\theta_1, \ldots, \theta_p]^T \) is an unknown parameter vector. The variance of any unbiased estimator \( \hat{\theta}_1 \) of \( \theta_1 \) obeys the lower bound:

\[
\text{var}_\theta(\hat{\theta}_1) \geq [\text{F}^{-1}(\theta)]_{11},
\]

where equality occurs iff there exists a nonrandom vector \( h_\theta \) such that

\[
h_\theta^T \nabla_\theta \ln f(X; \theta) = (\hat{\theta}_1 - \theta_1).
\]

In (58) \([A]_{ij}\) denotes the \( ij \) entry of matrix \( A \), and as before

\[
\text{F}(\theta) = -E \begin{bmatrix}
\frac{\partial^2 l(\theta)}{\partial \theta_1^2} & \frac{\partial^2 l(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l(\theta)}{\partial \theta_1 \partial \theta_p} \\
\frac{\partial^2 l(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l(\theta)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l(\theta)}{\partial \theta_2 \partial \theta_p} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 l(\theta)}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 l(\theta)}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 l(\theta)}{\partial \theta_p^2}
\end{bmatrix},
\]

and \( l(\theta) = \ln f(X; \theta) \).

Let the Fisher matrix be partitioned as

\[
\text{F}(\theta) = \begin{bmatrix}
a & b^T \\
b & C
\end{bmatrix},
\]
where
\* \( a = -E_2[\partial^2 \ln f(X; \theta)/\partial \theta_1^2] \) = Fisher info for \( \theta_1 \) without nuisance parameters,
\* \( b = -E_2[\partial^\nabla_{\theta_2, \ldots, \theta_p} \ln f(X; \theta)/\partial \theta_1] \) = Fisher coupling of \( \theta_1 \) to nuisance parameters,
\* \( C = -E_2[\nabla^2_{\theta_2, \ldots, \theta_p} \ln f(X; \theta)] \) = Fisher info for nuisance parameters.

Using the partitioned matrix inverse identity (2) the RHS of CRB (58) can be expressed as

\[
\left[ (F^{-1}(\theta)) \right]_{11} = \frac{1}{a - b^T C^{-1} b}.
\]

This gives several insights:

Observation 1: \( \left[ (F^{-1}(\theta)) \right]_{11} \geq 1/a = 1/\left[ (F(\theta)) \right]_{11} \). Thus presence of nuisance parameters can only degrade estimator performance;

Observation 2: the amount of degradation is directly proportional to the amount of information coupling between \( \theta_1 \) and \( \theta_2, \ldots, \theta_p \);

Observation 3: no degradation occurs when the Fisher matrix is block diagonal;

**Example 25** *Estimation of the mean of a Gaussian when the variance is a nuisance parameter.*

As in Example 21 assume that \( n \) i.i.d. samples \( \{X_i\}_{i=1}^n \) from a \( \mathcal{N}(\mu, \sigma^2) \) are available for estimating the mean \( \mu \), where the variance \( \sigma^2 \) is an unknown non-random nuisance parameter. We saw that, for any fixed value of \( \sigma^2 \), the ML estimator \( \hat{\mu} \) of \( \mu \) is the sample mean which does not depend on \( \sigma^2 \). Furthermore, the \( 2 \times 2 \) Fisher information matrix was determined to be diagonal, indicating that there is no information coupling between \( \mu \) and \( \sigma^2 \) and therefore lack of knowledge of \( \sigma^2 \) does not cause any performance degradation in \( \hat{\mu} \). In other words, for the Gaussian model ML estimation of the mean for unknown non-random \( \sigma^2 \) is easy.

It will be instructive to consider the case of a random nuisance parameter \( \sigma^2 \). There are many possible choices for the prior on \( \sigma^2 \) that could be postulated. A natural choice is the inverse-Gamma prior which is conjugate to the Gaussian distribution with random \( \sigma^2 \). A simpler choice, but one that leads to the same type of marginal distribution, is the improper prior: \( f(\sigma^2) \) proportional to \( \sigma^{-2} \) over the range \( \sigma^2 > 0 \). This prior is improper since it is not integrable. However, the marginalization integral \( f(X; \mu) = \int_0^\infty f(X|\sigma^2, \mu) f(\sigma^2) d\sigma^2 \) exists and is equal to the non-standardized student-t density

\[
f(X; \mu) = \frac{\kappa_{n-1}}{s} \left( 1 + \frac{n}{n-1} \frac{(\mu - X)^2}{s^2} \right)^{-n/2},
\]

where \( s^2 \) is the sample variance and \( \kappa_{n-1} \) is a normalizing constant depending only on \( n \) (see [35, Eq. 28.70]). The marginalized ML estimator of \( \mu \), obtained by maximizing \( f(X; \mu) \) over \( \mu \) is again the sample mean \( \bar{X} \), just like in the case of non-random \( \sigma^2 \).

In summary, the random and non-random approaches to nuisance parameters give the same answer. This is not always the case.
4.7 BACKGROUND REFERENCES

One of my favorite introductory texts covering estimation theory is the book on mathematical statistics by Mood, Graybill and Boes [56], mentioned before, which is concise, easy to read, and has many interesting examples and exercises. Nice books on this subject that focus on the Bayesian point of view are Ferguson and [19] and DeGroot [17]. A good survey of Bayesian tools for statistical inference, and estimation in particular, is the book by Tanner [81]. Texts which have more of an engineering flavor are the now classic book by Van Trees [84], and the more recent books by Kay [40], Srinath, Rajasekaran and Viswanathan [77], and Scharf [69]. For a more advanced treatment, requiring some background in real analysis, I like Bickel and Doksum [9], Lehmann [47], and Ibragimov and Has’minskii [32], and Poor [64].

4.8 EXERCISES

4.1 Prove the formula \(|a + \Delta| = |a| + \text{sgn}(a)\Delta + [\text{sgn}(a + \Delta) - \text{sgn}(a)](a + \Delta)\) in Sec. 4.2.2.

4.2 Show the equivalence of the two expressions (29) and (30).

4.3 Let \(X = [X_1, \ldots, X_n]^T\) be a vector of i.i.d. r.v.s \(X_i\) which are uniformly distributed over the interval \((\theta_1, \theta_2), \theta_1 < \theta_2\). Find the maximum likelihood estimator of \(\theta\).

4.4 Let \(Z_i, i = 1, \ldots, n,\) be a set of i.i.d. random variables each with the alpha density

\[
f(z; \beta) = \frac{\beta}{\sqrt{2\pi}\Phi(\alpha)z^2} \exp\left(-\frac{1}{2} [\alpha - \beta/z]^2\right),
\]

where \(\beta > 0\) is unknown, \(\alpha\) is known and \(\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du\) is the standard normal CDF. Assuming that \(\alpha = 0\) and that \(\beta\) has an exponential prior density: \(f(\beta) = \frac{1}{\sigma_\beta} e^{-\beta/\sigma_\beta}\), where \(\sigma_\beta > 0\) is known. Find an expression for the MAP estimate of \(\beta\). What does the MAP estimate reduce to as \(\sigma_\beta \to \infty\) (least informative prior)?

4.5 Let \(W_i, i = 1, \ldots, n\), be a set of zero mean i.i.d. Gaussian random variables with variance \(\sigma_w^2\). Let \(a\) be a zero mean Gaussian random variable with variance \(\sigma_a^2\) which is independent of \(W_i\). The objective is to estimate the value of \(a\) given the observation

\[X_i = a + W_i, \quad i = 1, \ldots, n\]

(a) Find the MMSE estimator of \(a\). How does this estimator compare to the MAP and MMAE estimators of \(a\)?

(b) Compute the MSE of the MMSE estimator (Hint: express error as a sum of two independent r.v.’s to simplify algebra). What happens to the MSE as \(n \to \infty\) or as SNR = \(\sigma_a^2/\sigma_w^2\) \(\to \infty\)?

4.6 Let \(X = [X_1, \ldots, X_n]^T\) be a vector of i.i.d. Gaussian r.v.s with mean \(\mu\) and variance \(\sigma^2 = \mu^2\)

\(X_i \sim \mathcal{N}(\mu, \mu^2)\).

(a) Find a method of moments (MOM) estimator of \(\mu\) based on the first moment.

(b) Find the maximum likelihood estimate of \(\mu\).

4.7 Let \(X_i, i = 1, \ldots, n,\) be an i.i.d. sample from the shifted exponential density \(f(x; \theta) = e^{-(x-\theta)}, x \geq \theta,\) where \(\theta\) is an unknown parameter \(-\infty < \theta < \infty\). Assume that \(n \geq 1\).

(a) Find a MOM estimator of \(\theta\).

(b) Find the ML estimator of \(\theta\).
(c) Assuming the exponential prior for $\theta$, $f(\theta) = e^{-\theta}$, $\theta \geq 0$, find the MAP estimator, the MMSE estimator, and the MMAE estimator of $\theta$ given the i.i.d. sample (be careful with your limits of integration in computing $f(\theta|x)!$). What happens to these estimators as $n \to \infty$?

(d) Calculate the MSE of each of the estimators derived in part (c) (assume large $n$). Verify that the MMSE estimator has the lowest MSE.

4.8 The mean square error of a certain unbiased estimator $\hat{\theta}(x)$ of the mean of a measured random variable is equal to $\sigma^2/2$ where $\sigma^2 = \text{var}(x)$. What if anything does this tell you about the distribution of $x$ (Hint: what does the CR bound say about distributions that are impossible)?

4.9 Available are $n$ i.i.d. samples of a random variable $X$ with density

$$f(x; \theta) = \frac{1 + 3\theta x^2}{1 + \theta}$$

where $-1 \leq x \leq 1$ and $\theta \geq 0$.

(a) Is this density in the exponential family?

(b) Is the sample mean a sufficient statistic? If so, prove it for general $n$. If not, give a counterexample, e.g. specialize to $n = 2$.

(c) Find a MOM estimator of $\theta$.

(d) Find the CR bound on estimator variance for any unbiased estimator of $\theta$.

(e) Using either numerical integration (MATLAB) or analysis find the bias and variance of the MOM estimator and compare to the CR bound for large $n$ (e.g. $n = 100$).

4.10 Let the observation $X$ have conditionally uniform density

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 < x \leq \theta \\ 0, & \text{o.w.} \end{cases}$$

where $\theta$ is a random variable with density

$$f_\theta(\theta) = \begin{cases} \theta \exp(-\theta), & \theta \geq 0 \\ 0, & \text{o.w.} \end{cases}$$

A useful formula ($v \geq 0$): $\int_v^\infty u e^{-u} du = (v + 1)e^{-v}$

(a) Find the MAP estimator of $\theta$.

(b) Find the minimum mean squared error estimator of $\theta$.

(c) Find the minimum mean absolute error estimator of $\theta$.

4.11 Let $Z$ be a single observation having density function

$$f_\theta(z) = (2\theta z + 1 - \theta), \quad 0 \leq z \leq 1$$

where $-1 \leq \theta \leq 1$.

(a) Assuming that $\theta$ is a nonrandom parameter, find and plot the maximum likelihood estimator of $\theta$ as a function of $Z$.

(b) Is the ML estimator unbiased? If so does it achieve the CR bound?

(c) Now assume that $\theta$ is a random variable with uniform prior density: $f(\theta) = \frac{1}{2}$, $\theta \in [-1, 1]$. Find and plot the minimum mean square error estimator of $\theta$ as a function of $Z$. 
(d) Compute the bias and MSE for the estimator in part a and the conditional bias $E[\hat{\theta}|\theta] - \theta$ and the conditional MSE $E[(\hat{\theta} - \theta)^2|\theta]$ given $\theta$ for the estimator in part c. Plot the two conditional MSE functions obtained and compare the MSE’s of the two estimators. Does one estimator perform uniformly better than the other?

4.12 $X = [X_1, \ldots, X_n]^T$ is an i.i.d. observation from the Gamma density

$$X_i \sim f(x|\theta) = \frac{1}{\Gamma(\theta)} x^{\theta-1}e^{-x}, \quad x \geq 0$$

where $\theta$ is an unknown non-negative parameter and $\Gamma(\theta)$ is the Gamma function. You should note the useful formulae

$$\Gamma(\theta) = \int_0^\infty x^{\theta-1}e^{-x}dx \quad \text{and} \quad \frac{\Gamma(\theta + k)}{\Gamma(\theta)} = \theta(\theta + 1)\cdots(\theta + k - 1)$$

(a) Find the CR bound on unbiased estimators of $\theta$.

(b) Find the first order MOM estimator of $\theta$ by matching ensemble mean to sample mean.

Is your estimator unbiased? Compute the variance of your estimator.

4.13 In this exercise you will establish that UMVUE’s do not always exist. Let $Z$ be a r.v. with probability mass function

$$p_\theta(z) = \begin{cases} \theta, & z = -1 \\ (1 - \theta)^2z, & z = 0, 1, 2, \ldots \end{cases}$$

where $\theta \in (0, 1)$.

(a) Define the estimator

$$\hat{\theta}_o(z) = \begin{cases} 1, & z = -1 \\ 0, & z = 0, 1, 2, \ldots \end{cases}$$

Show that $\hat{\theta}_o$ is an unbiased estimator of $\theta$.

(b) Note that any unbiased estimator $\hat{\theta}$ can be expressed in the form $\hat{\theta} = \hat{\theta}_o + U$ where $U = U(Z)$ is a statistic satisfying $E_\theta[U] = 0$ (any $U$ satisfying this condition is called an ancillary statistic). Using this condition and the form for the pmf of $Z$ given above, establish that $U$ must be of the form $U(Z) = aZ$ for some non-random constant $a$ (Hint: Z-transform tables may be helpful).

(c) Now find an expression for the variance of an unbiased $\hat{\theta}$ and show that the value $a$ which minimizes the variance is a function of $\theta$. Hence no single unbiased estimator can achieve minimum variance for all $\theta \in (0, 1)$ and therefore no UMVUE for $\theta$ exists.

(d) Show that a UMVUE for $\phi = (1 - \theta)^2$ does exist even though a UMVUE for $\theta$ does not exist (Hint: define $\hat{\phi}_o(z) = 1$ for $z = 0$ and $\hat{\phi}_o(z) = 0$, otherwise and repeat the steps in part a through c).

4.14 The observation consists of $x_1, \ldots, x_n$ i.i.d. samples where $x_i \sim f(x|\theta)$ and

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} x^{\frac{1}{\theta} - 1}, & 0 \leq x \leq 1 \\ 0, & \text{o.w.} \end{cases}$$

where $\theta, \, 0 < \theta < \infty$ is an unknown parameter.
(a) Compute the CR bound on unbiased estimators of $\theta$. Is there an estimator that achieves the bound?

(b) Find the maximum likelihood estimator of $\theta$.

(c) Compute the mean and variance of the maximum likelihood estimator. Specify a function $\varphi = g(\theta)$ for which the maximum likelihood estimator of $\varphi$ is efficient.

(d) From one of your answers to parts a-c you should be able to derive the following formula

$$\int_0^1 u^\beta \ln \left( \frac{1}{u} \right) du = \frac{1}{(1 + \beta)^2}, \quad \beta > -1.$$  

4.15 The measurement $\underline{x} = [x_1, \ldots, x_n]^T$ is i.i.d. Gaussian with unknown mean $\mu$ and variance $\sigma^2$.

(a) Show that the sample mean $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ and sample variance $s^2 = (n-1)^{-1} \sum_{k=1}^n (x_k - \bar{x})^2$ are unbiased estimators and that they are uncorrelated and independent random variables. (Hint: show that the Gaussian random variables $x_i - \bar{x}$ and $\bar{x}$ are uncorrelated for $i = 1, \ldots, n$).

(b) Using the results of part (a) derive the covariance matrix for the estimator $\hat{\theta} = [\bar{x}, s^2]^T$. (Hint: to save yourself lots of algebra you should represent $s^2 = s^2(\underline{x})$ in terms of $\sigma^2$ and the sample variance $s^2(\underline{z})$ for $\underline{z}$ a vector of $n$ i.i.d. zero mean unit variance Gaussian variables. Then use the representation (ch. 3 of course notes) $s^2(\underline{z}) = \frac{1}{n-1} \chi_{n-1}$ and properties of the Chi square r.v. to find the expression for variance of $s^2$).

(c) Derive the CR bound on the covariance matrix of any unbiased estimator $\hat{\theta}$ of $\theta = [\theta_1, \theta_2]^T = [\mu, \sigma^2]^T$. Compare to the result of part (b).

4.16 Show that if the CR bound is attained with equality then $E[UU^T]$ has rank $p$, where $U$ is given by (49). (Hint: show that the matrix

$$E[UU^T] = \begin{bmatrix} F^{-1}(\theta) & I \\ I & \theta(\theta) \end{bmatrix}$$

has rank $p$.)

4.17 An alternative approach to parameter estimation is called the "quantile matching method" and you will explore this method here. Let $f(x; \theta)$ be a density of the continuous r.v. $X$ parameterized by the scalar parameter $\theta$ and define the theoretical cdf $F(x; \theta) = \int_{-\infty}^x f(u; \theta)du$. For $n$ i.i.d. realizations $\{X_i\}_{i=1}^n$ from $f(x; \theta)$ define the empirical cdf as the fraction of $X_i$’s which are less than or equal to $x$:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty,x]}(X_i)$$

where $I_A(y)$ equals 1 if $y \in A$ and zero otherwise (the indicator function of set $A$).

(a) Derive the mean $E[\hat{F}(x)]$ and covariance $\text{cov}(\hat{F}(x), \hat{F}(y))$ of $\hat{F}$. Show that $\hat{F}(x)$ is an asymptotically consistent estimator of $F(x; \theta)$.

(b) The quantile matching estimate (QME) $\hat{\theta}$ is defined as that value of $t$ which minimizes

$$\int_{-\infty}^{\infty} |F(x;t) - \hat{F}(x)|^2 dx$$

(59)
Let $\theta$ be a location parameter: $f(x; \theta) = f(x - \theta)$. Using the definition (59), show that $\hat{\theta}$ must satisfy the following equation (Hint: use integration by parts):

$$\int_{-\infty}^{\infty} f(x - \hat{\theta}) \hat{F}(x) dx - 1/2 = 0. \quad (60)$$

Show that if $\hat{\theta}$ is the unique solution to (60) it is an asymptotically consistent estimator of $\theta$ (Hint: for $\hat{\theta} = t$ fixed and non-random, compute mean square value of left hand side of (60) and show that as $n \to \infty$ it goes to a function of $t$ which equals zero at $t = \theta$).

(c) Using matlab, or other software application of your choice, simulate the QME and the MLE for the following cases:

i. $f(x; \theta)$ Gaussian with variance 1 and mean $\theta$.

ii. $f(x; \theta) = \alpha e^{-\alpha(x-\theta)} I_{[\theta, \infty)}(x)$ (shifted exponential) with $\alpha = 1$.

Run the above simulations 50-100 times each for the cases of $n = 1, 5, 10, 15, 20, 25$ observations, respectively. Using the results of your simulations find and plot as a function of $n$: 1) the average mean-squared error for MLE and QME estimators; 2) the average quantile squared error (59) evaluated at $t = \hat{\theta}$ (you should show 4 different plots). Also generate a couple of representative plots of the objective function (59) as a function of $t$ for the Gaussian and shifted exponential cases above. Comment on what can be concluded from your simulation study.

4.18 Available are $n$ i.i.d. samples of a discrete random variable $X$ with probability mass function $P(X = k) = p(k; \theta)$, given by

$$p(k; \theta) = \begin{cases} \left( \frac{\theta}{1+\theta} \right)^{k-k_0} \frac{1}{1+\theta}, & k = k_0, k_0 + 1, \ldots, \\ 0, & \text{a.w.} \end{cases}$$

where $k_0$ is a known non-negative integer and $\theta$ is unknown with $0 \leq \theta < \infty$. (A potentially useful identity: $\sum_{k=0}^{\infty} ka^k = a/(1-a)^2$).

(a) Is this density in the exponential family with mean value parameterization? Find a one dimensional sufficient statistic for $\theta$.

(b) Find a MOM estimator of $\theta$.

(c) Find the ML estimator of $\theta$.

(d) Find the Fisher information on estimator variance for any unbiased estimator of $\theta$. Are either of the estimators of part (b) or part (c) efficient?

4.19 Available is a single measurement of a random variable $W$. The model for $W$ is

$$W = (1 - Z)X + ZY,$$

where $Z$ is Bernoulli with $P(Z = 0) = P(Z = 1) = 1/2$, $X$ is Gaussian with zero mean and variance $\sigma^2$, and $Y$ is Gaussian with mean $\mu$ and variance $\sigma^2$. Assume that $\mu$ and $\sigma^2$ are known and that $X, Y, Z$ are independent.

(a) Find the posterior distribution of $Z$.

(b) Find the minimum mean squared error estimator of $Z$. Plot the estimator as a function of $W$.

(c) Find the MAP estimator of $Z$. Plot the estimator as a function of $W$. 
4.20 Let $X_1, X_2, \ldots, X_n$ be i.i.d. variables with the standard Pareto density:

$$f(x; \theta) = \begin{cases} \theta e^{\theta x} - (\theta + 1), & x \geq c \\ 0, & \text{o.w.} \end{cases}$$

where $c > 0$ is known and $\theta > 0$ is unknown.

(a) Is $f(x; \theta)$ a member of the exponential family? Why or why not?

(b) Find a one dimensional sufficient statistic for $\theta$ given $X_1, X_2, \ldots, X_n$.

(c) Find the Fisher information and state the CR bound for unbiased estimators of $\theta$.

(d) Derive the maximum likelihood estimator $\hat{\theta}$ of $\theta$.

(e) Is your estimator efficient?

4.21 Let $X_1, X_2, \ldots, X_n$ be i.i.d. variables with the generalized Pareto density:

$$f(x; \theta) = \begin{cases} c^{c} x^{-\theta (c+1)}, & x \geq \theta \\ 0, & \text{o.w.} \end{cases}$$

where $c > 0$ is known and $\theta > 0$ is unknown.

(a) Is $f(x; \theta)$ a member of the exponential family? Why or why not?

(b) Find a one dimensional sufficient statistic for $\theta$ given $X_1, X_2, \ldots, X_n$.

(c) Derive the maximum likelihood estimator $\hat{\theta}$ of $\theta$.

4.22 The posterior density of a scalar parameter $\theta$ given an observation $\underline{x} = [x_1, \ldots, x_n]^T$ is a function of the form $f(\theta|\underline{x}) = g(x_{\text{bar}} - \theta)$ where $x_{\text{bar}}$ is the sample mean and $g$ is an integrable function satisfying $g(-u) = g(u)$ and $g(0) > g(u), u \neq 0$. Derive the MAP, CME and CmE estimators of $\theta$.

4.23 The CRB has several generalizations that we explore in this problem for scalar parameters $\theta$ of a density $f_\theta(x)$.

(a) Define the finite difference $\delta f = (f_{\theta+\Delta} - f_{\theta})/\Delta$. Show that for any unbiased estimator $\hat{\theta}$ of non-random $\theta$

$$\text{var}_\theta(\hat{\theta}) \geq \frac{1}{E_\theta \left[ (\delta f_{\theta}/f_{\theta})^2 \right]}$$

with equality iff $\delta f_{\theta}/f_{\theta} = k_{\theta}(\hat{\theta} - \theta)$ for non-random constant $k_{\theta}$. The above bound is called the Chapman Robbins version of the Barankin bound.

(b) Show that the bound of part (a) implies the CRB in the case that $\theta$ is a non-random continuous parameter and $f_{\theta}$ is smooth (Hint: take limit as $\Delta \to 0$).

(c) When $\theta$ is a random variable with prior density $p(\theta)$ show that

$$E[(\hat{\theta} - \theta)^2] \geq \frac{1}{J}$$

where

$$J = E \left[ (\delta p(\theta|X)/p(\theta|X))^2 \right]$$

and $\delta p(\theta|X) = (p(\theta + \Delta|X) - p(\theta|X))/\Delta$. Here the expectation $E$ is taken over both $X$ and $\theta$. 
4.24 Let \( g(x; \phi_1) \) and \( h(x; \phi_2) \) be densities where \( \phi_1, \phi_2 \) are unknown scalar parameters. The arithmetic epsilon mixture model for \( X \) is:

\[
    f_A(x; \theta) = (1 - \epsilon)g(x; \phi_1) + \epsilon h(x; \phi_2)
\]

where \( 0 \leq \epsilon \leq 1 \) and \( \theta = [\phi_1, \phi_2, \epsilon]^T \). The geometric epsilon mixture model for \( X \) is:

\[
    f_G(x; \theta) = \frac{1}{d(\theta)} g^{1-\epsilon}(x; \phi_1)h^\epsilon(x; \phi_2),
\]

where

\[
    d(\theta) = \int g^{1-\epsilon}(x; \phi_1)h^\epsilon(x; \phi_2)dx
\]
is a normalizing constant (related to the Rényi \( \epsilon \)-divergence between \( g \) and \( h \)). From this exercise you will appreciate that the mixture \( f_G \) is easier to deal with than \( f_A \) for the purposes of investigating CR bounds, detectors and estimators. Assume that \( g \) and \( h \) are members of the exponential family of densities.

(a) Show that the three parameter density \( f_G(x; \theta) \) is a member of the exponential family. Show that \( f_A(x; \theta) \) is not a member of this family.

(b) Derive expressions for the six distinct entries of the Fisher information matrix (FIM) for jointly estimating the parameters \( \theta \) from \( n \) i.i.d. observations from \( f_G \). An explicit expression for the FIM does not generally exist for the standard mixture model \( f_A \).

(c) For \( n \) i.i.d. observations from \( f_G \) give a condition on the parameter vector \( \theta \) which guarantees that an efficient estimator exist for \( \theta \), i.e. for which the inverse FIM is an achievable lower bound on the covariance of unbiased estimators of \( \theta \) (Hint: what is the mean value parameterization as defined by (28)?)

(d) In the sequel of this exercise we specialize \( f_G \) to the case of a geometric mixture of two exponential densities

\[
    g(x; \theta) = \phi_1 \exp(-x\phi_1), \quad h(x; \theta) = \phi_2 \exp(-x\phi_2),
\]

where \( x, \phi_1, \phi_2 > 0 \). Derive an expression for \( d(\theta) \). Is the CR bound achievable for this model?

(e) Let \( n \) i.i.d. realizations be available from the geometric mixture \( f_G \) specified by (61) and (62). By evaluating the gradient of the likelihood function, find a set of (non-linear) equations which must be satisfied by the MLE of \( \theta \). Using these equations, and assuming that \( \phi_1, \phi_2 \) are known, find an explicit expression for the MLE of \( \epsilon \).

4.25 Let \( S \) and \( X \) be jointly Gaussian distributed with means and variances

\[
    \begin{align*}
        E[S] &= \mu_S, & E[X] &= \mu_X, \\
        \text{var}(S) &= \sigma_S^2, & \text{var}(X) &= \sigma_X^2, \\
        \text{cov}(S, X) &= \rho \sigma_S \sigma_X.
    \end{align*}
\]

Specifically the joint density is bivariate Gaussian

\[
    f_{S,X}(s, x) = \frac{1}{2\pi \sigma_S \sigma_X \sqrt{1 - \rho^2}} \exp\left( -\frac{1}{2(1 - \rho^2)} \left[ \frac{(s - \mu_S)^2}{\sigma_S^2} - 2\rho \frac{(s - \mu_S)(x - \mu_X)}{\sigma_S \sigma_X} + \frac{(x - \mu_X)^2}{\sigma_X^2} \right] \right).
\]

(a) By integrating the joint density over \( s \), show that the marginal density \( f_X \) of \( X \) is a univariate Gaussian density with mean parameter \( \mu_X \) and variance parameter \( \sigma_X^2 \).
(b) Using the above to show that the conditional density $f_{S|X}(s|x)$ of $S$ given $X$ is univariate Gaussian with mean and variance parameters

$$
\mu_{S|X}(x) = \mu_S + \rho \frac{\sigma_S}{\sigma_X}(x - \mu_X),
\sigma^2_{S|X} = (1 - \rho^2) \sigma^2_S.
$$

Note that while the mean parameter depends on $x$ the variance parameter is independent of $x$.

(c) Using this form for the conditional density show the mean and variance parameters are precisely the conditional mean and variance of $S$ given $X = x$, respectively.

4.26 A charity box is placed in a mall. The box can only accept quarters. With probability $p$ (a deterministic quantity), a (good) person would come and place a quarter in the box, thus incrementing the number of quarters in the box by one. With probability $1 - p$, a (bad) person would come and empty the box, thus setting the number of quarters in the box to zero.

Assuming stationarity, it can be shown that the probability that $k$ quarters will be observed at the end of the $d$-th day is

$$
P(T(d) = k) = p^k (1 - p).
$$

(Notation: $T(d)$ is the random variable representing the number of quarters in the box at the end of the $d$-th day.) In the following you should assume that $T(1), T(2), \ldots$, are independent identically distributed (i.i.d) random variables.

(a) Maximum Likelihood and Efficiency: To estimate the percentage of good people $p$, the box monitor counts the number of quarters in the box at the end of each day, $D$ days in a row.

- Write down the joint PDF of the vector of number of quarters observed $[T(1), T(2), \ldots, T(D)]$.
- Find the ML estimator of $p$ given $T(1) = k_1, T(2) = k_2, \ldots, T(D) = k_D$.
- Is the ML estimator $\hat{p}_{ML}$ efficient?

(b) Method of Moments: Define the the average number of quarters observed as $\bar{k} = \frac{1}{D} \sum_{d=1}^{D} k_d$.

- Find the expected value of the average number of quarters observed $E[\bar{k}]$ (hint: $\sum_{n=0}^{\infty} np^n = \frac{p}{(1-p)^2}$).
- Based on this result, suggest a method of moments estimator for $p$.

(c) Efficiency and the CRB: To investigate how well the charity box is doing, a new measure is considered $\gamma = \frac{p}{1-p}$, the ratio of the percentage of good people to the percentage of bad people, otherwise known as the good-to-bad ratio (GBR).

- Is the ML estimator of the GBR $\hat{\gamma}_{ML}$ efficient?
- Find the ML estimator of the GBR $\hat{\gamma}_{ML}$.
- Find the Cramér-Rao bound (CRB) on the MSE of an unbiased estimator for the GBR.
- Find the MSE of the ML estimator of the GBR.

4.27 Here you will show that the MLE is invariant to arbitrary functional transformations of the parameter. Let $\theta$ be a scalar parameter with range $\Theta = (-\infty, \infty)$, assume the sample $X$ has j.p.d.f $f(x; \theta)$, and that there exists a unique MLE $\hat{\theta}$. Given a transformation $g$ define the new parameter $\varphi = g(\theta)$.
(a) Assume that $g$ is monotone, i.e. $g(\theta)$ is 1-1 invertible over all $\Theta$. Show that the MLE of $\varphi$ is

$$\hat{\varphi} = g(\hat{\theta}).$$

(b) Next assume that $g$ is smooth in the sense of piecewise monotonicity, i.e., there exists a partition of $\Theta$ into intervals $(-\infty, \theta_1], (\theta_1, \theta_2], \ldots, (\theta_M, \infty)$ such that $g$ is monotone over each of these intervals ($M$ may not be finite). Define the integer function $h$ by: $h(\theta) = k$, if $\theta$ is in the $k$-th interval, $k = 1, \ldots, M + 1$. Show that the scalar-to-vector mapping $\theta \rightarrow [g(\theta), h(\theta)]$ is 1-1 invertible.

(c) Using result of (b) show that the MLE is invariant to piecewise monotone functional transformation.

4.28 Derive the CR bound (58) on the variance of an unbiased scalar estimator $\hat{\theta}_1$ of $\theta_1$ when the rest of the parameters $\theta_2, \ldots, \theta_p$ in $\theta$ are unknown nuisance parameters. Do not assume that the nuisance parameters have unbiased estimators (Hint: define $U = [\hat{\theta}_1 - \theta_1, \nabla_2^T \ln f(X; \theta)]^T$ and proceed as in the proof of the matrix CRB).

4.29 A sequence of measurements $X_1, \ldots, X_n$ are i.i.d. with marginal density

$$f_{X_1}(x; \theta) = \frac{\theta}{x^2} e^{-\frac{x}{\theta}}, \ x > 0$$

where $\theta > 0$ is an unknown parameter.

(a) For part (a) and (b) assume that $\theta$ is non-random. Is this density a member of the exponential family? Find a one dimensional sufficient statistic for $\theta$.

(b) Find the maximum likelihood estimator of $\theta$.

(c) For part (c) and (d) assume that $\theta$ is a random variable having density

$$f(\theta) = e^{-\theta}, \ \theta > 0.$$ 

Find the MAP estimator of $\theta$.

(d) Find the minimum mean squared error estimator of $\theta$ and compare to your result in part (c). Hint: $\int_0^\infty \alpha^\alpha e^{-\alpha} d\alpha = n!$.

4.30 Show that the vector conditional mean estimator $\hat{\theta}_{CME}$ of a random vector parameter $\theta$ satisfies the property that, for any other estimator $\hat{\theta}$

$$E[(\varphi - \hat{\varphi})(\theta - \hat{\theta})^T] \geq E[(\theta - \hat{\theta}_{CME})(\theta - \hat{\theta}_{CME})^T],$$

where the matrix inequality $A \geq B$ is interpreted in terms of non-negative definiteness of $A - B$.

4.31 Let $\theta$ be a nonrandom vector parameter of some smooth (in $\theta$) density function $f(x; \theta)$. Show that $E_\theta [\nabla_{\theta} \ln f(X; \theta)(\nabla_{\theta} \ln f(X; \theta))^T] = E_\theta [-\nabla_\theta^2 \ln f(X; \theta)].$

4.32 Assume that $X$ is a sample from a density in an exponential family with scalar parameter $\theta$ having the mean value parameterization $\{E_\theta[t(X)] = \theta, \text{recall discussion in Sec. 3.6.1.}\}$ Assuming the Fisher information $F(\theta)$ exists show that

$$F(\theta) = 1/\text{var}_\theta(t(X)).$$

Now show that if one has an i.i.d. sample $X = [X_1, \ldots, X_n]^T$ from such a density then $\hat{\theta} = n^{-1} \sum_{i=1}^n t(x_i)$ is an unbiased and efficient estimator of $\theta$. 

\[ \text{(63)} \]
4.33 In this problem you will investigate estimation of the transition probability of an observed binary sequence called a Markov chain. Available for measurement is a binary sequence $X_0, X_1, \ldots, X_n$ whose joint probability mass function satisfies

$$p_\theta(x_0, x_1, \ldots, x_n) = p(x_0) \prod_{i=1}^{n} p_\theta(x_i|x_{i-1}), \ x_i \in \{0, 1\}$$

where $p(x_0) = P(X_0 = x_0) = 1/2$, and the conditional probability $p_\theta(x_i|x_{i-1}) = P(X_i = x_i|X_{i-1} = x_{i-1})$ is given by

$$p_\theta(x_i|x_{i-1}) = \begin{cases} \theta, & (x_i, x_{i-1}) \in \{(1,1), (0,0)\} \\ 1 - \theta, & o.w. \end{cases}$$

The quantity $1 - \theta$ is the transition probability of the Markov chain (note that it is only an i.i.d. process when $\theta = 1/2$). The problem of estimating $\theta$ from a realization $x_0, x_1, \ldots, x_n$ arises in (BSC) channel identification and sequence dependency estimation.

(a) Find a sufficient statistic for $\theta$ and show that the likelihood function is in the exponential family. (Hint: express $p_\theta(x_i|x_{i-1})$ as an exponential function of $\theta$ and $1-\theta$ with exponent dependent on products of $x_k$’s).

(b) Find a method of moments estimator of $\theta$. Is your estimator unbiased?

(c) Find a maximum likelihood estimator of $\theta$. Is your estimator unbiased?

(d) Compute the Cramér-Rao lower bound on the variance of unbiased estimators of $\theta$. Is the CR bound achievable by the ML estimator?

4.34 Available are $n$ i.i.d. samples $\{X_i\}_{i=1}^{n}$ of a binary random variable $X$ with probability mass function $P(X = x) = p(x; \theta)$, given by

$$p(x; \theta) = \begin{cases} \theta^x \frac{1}{1+\theta}, & x = 0, 1 \\ 0, & o.w. \end{cases}$$

where $\theta > 0$ is an unknown non-random parameter.

(a) Find the MLE of $\theta$. Show that your estimator is not unbiased (Hint: specialize to the case $n = 1$ first.)

(b) Show that in fact no unbiased estimator can exist for this estimation problem (Use same hint as in (a)).

(c) Now assume that $\theta$ is a uniform random variable. Find the MAP and CME estimators of $\theta$ (to obtain a closed form expression for the CME you may specialize to the case of $n = 1$).

4.35 You measure $n$ i.i.d. samples $\{X_i\}_{i=1}^{n}$ of a discrete random variable $X$ with probability mass function $P(X = x) = p(x; \theta)$, given by

$$p(x; \theta) = \begin{cases} (1-\theta)^x \theta, & x = 0, 1 \ldots \\ 0, & o.w. \end{cases}$$

where $\theta$ is unknown with $0 < \theta < 1$. (A potentially useful identity: $\sum_{k=0}^{\infty} k a^k = a/(1-a)^2$).

(a) Is this density in the exponential family with mean value parameterization? Find a one dimensional sufficient statistic for $\theta$.

(b) Find a MOM estimator of $\theta$. 
(c) Find the ML estimator of $\theta$.
(d) Find the Fisher information on estimator variance for any unbiased estimator of $\theta$. Are either of the estimators of part (b) or part (c) efficient?

4.36 The negative binomial distribution is often used in survival analysis as a model for the waiting time $Y = X + k$ until the $k$-th occurrence of a “1” in a set of Bernoulli trials, where $X$ is a random variable with distribution

$$P_{\theta}(X = x) = \binom{k - 1 + x}{k - 1} \theta^x (1 - \theta)^k, \quad x = 0, 1, 2, \ldots \tag{64}$$

Here $\theta \in [0, 1]$ and $k$ is a positive integer. The moment generating function of this distribution is $M(s) = E[e^{sX}] = (1 - \theta)^k / (1 - \theta e^s)^k$ from which you can show that $E_{\theta}[X] = k\theta / (1 - \theta)$ and $\text{var}_{\theta}(X) = k\theta / (1 - \theta)^2$.

The objective is to estimate $\theta$, or related parameters, based on $n$ i.i.d. samples $X_1, \ldots, X_n$. You should assume that $k$ is fixed and known in answering following.

(a) Is the distribution (64) in the exponential family? If so express the distribution in terms of its natural parameterization and in terms of its mean parameterization, respectively.
(b) Find the ML estimator of $\theta$.
(c) Find the CRB on the variance of unbiased estimators of $\theta$.
(d) Now assume that the parameter to be estimated is $\phi = \theta / (1 - \theta)$. Find the ML estimator and find its bias and variance.
(e) Find the CRB on the variance of unbiased estimators of $\phi$. Is the CRB achieved by the ML estimator of (d)?

End of chapter