

Post-Selection Estimation Theory

Thesis submitted in partial fulfillment
of the requirements for the degree of
"DOCTOR OF PHILOSOPHY"

by
Nadav Harel

Ben-Gurion University of the Negev

date

Beer Sheva

Post-Selection Estimation Theory

Thesis Submitted in partial fulfillment
of the requirements for the degree of
"DOCTOR OF PHILOSOPHY"

by
Nadav Harel

Submitted to the Senate of Ben-Gurion University of the Negev

Approved by the advisor, Prof. Tirza Routtenberg

Signature _____ Date _____

Approved by the Dean of the Kreitman School of Advanced Graduate Studies

Signature _____ Date _____

Beer Sheva

This work was carried out under the supervision of Prof. Tirza Routtenberg

In the School of Electrical and Computer Engineering

Faculty of Engineering

Acknowledgements

I am deeply indebted to Prof. Tirza Routtenberg, my supervisor, for her invaluable guidance, unwavering support, and constant encouragement throughout my M.Sc. and Ph.D. studies. Her relentless pursuit of excellence inspires me and shapes me into the researcher I am today.

As the Jewish Midrash, Genesis Rabbah, states, "A knife does not sharpen itself except on the thigh of its companion." I am sincerely grateful for your patience and steadfast determination during times of need.

I am grateful to the Kreitman School of Advanced Graduate Studies at Ben-Gurion University for its support through the Negev Scholarship. In addition, to the Israeli Science Foundation (grant No. 1173/16 and grant No. 1148/22) and the Pazi Foundation.

Lastly, I wish to dedicate this work to my wife, Dina, whose unwavering support has been my anchor throughout this journey. And to my son, Shaked, for the joy he brings to my life.

Abstract

Post-selection estimation refers to the scenario where a preliminary data-based selection stage determines the specific estimation problem. In this work, we are researching two fundamental problems of this framework: estimation after model selection and estimation after parameter selection. In estimation after model selection, the observation model is unknown. Therefore, prior to estimation, a model selection procedure is used to choose a model from a set of candidate models. Then, the parameters of the selected model are estimated. In estimation after parameter selection, the observations model is known. In this case, the selection refers to choosing the “parameters of interest” based on the data, while the rest of the unknown parameters are considered as nuisance parameters. In both problems, the selection stage impacts the subsequent estimation, for example, by introducing a selection bias. This research establishes a post-selection estimation theory for these two special cases, including estimation methods, appropriate unbiasedness, and performance bounds.

The contributions of this work are as follows. First, for the estimation after parameter selection problem, we consider the post-selection mean-squared error (PSMSE) as an appropriate performance measure that takes into account the selection procedure. We introduce the appropriate unbiasedness criterion, which is the unbiasedness in the Lehmann sense w.r.t. post-selection mean-squared error cost function. The post-selection maximum likelihood (PSML) estimator has been presented and has been shown to reduce the bias in Lehmann sense and the PSMSE compared with conventional estimators. Since, the PSML estimator often lacks an analytical form and has high computational complexity, we developed new low-complexity post-selection estimation methods for estimation after parameter selection architecture. In addition, we present an appropriate Cramér-Rao bound (CRB) for this problem, the Ψ -CRB, and develop a new algorithm for efficient computation of approximate this bound empirically. We generalize this model for unidentifiable model scenario where all the parameters cannot be estimated and a selection stage, that aims to identify the significant parameters is conducted prior to

the estimation. We present the coherent PSML estimator as an appropriate estimator for this problem and provided practical algorithm for implementation. We show that the presented estimator out-perform the outperforms other common solutions for this problem.

Second, we consider the estimation after model selection problem. The estimation after model selection is closely related to the concept of estimation under model misspecification. While the literature on estimation under misspecified models offers a framework for addressing model misspecification, it does not account for the selection process that led to the misspecified model. In estimation post-model selection, there are several candidate models, each with the potential to be selected. Consequently, the interpretation of the assumed model, in this case, is not straightforward. We present three different interpretations to address the non-Bayesian post-model-selection estimation problem as an estimation problem under model misspecification. Each of these interpretations induces a misspecified maximum likelihood estimator and a novel corresponding misspecified CRB. Finally, we consider a post-model-selection Bayesian parameter estimation approach of a random vector with an unknown deterministic support set, where this support set represents the model. We present different estimators and performance bounds. In particular, we develop the selective Bayesian CRB (BCRB) and selective tighter BCRB, lower bounds on the mean-squared error (MSE) for any coherent estimator.

Keywords:

Bayesian framework, Cramér-Rao bound, Lehmann-unbiasedness, lower bounds, mean-squared-error, model misspecification, model-selection, non-Bayesian framework, parameter estimation, parameter selection, selective inference

Contents

Contents	iv
List of Figures	viii
Glossary	x
Notations	xi
1 Introduction	1
1.1 Scientific Background	1
1.1.1 Non-Bayesian Estimation	2
1.1.2 Bayesian Estimation	3
1.2 Post-Selection Estimation	5
1.2.1 Estimation After Parameter Selection	5
1.2.2 Estimation After Model Selection	6
1.2.3 Estimation Under Model Misspecification	7
1.3 Contributions of the Dissertation	7
1.4 List of Publications	9
1.5 Outline	10
1.6 Notations	11
2 Low-Complexity Methods for Estimation After Parameter Selection	12
2.1 Observation Model	12
2.1.1 Special Cases	13
2.2 PSMSE and Ψ -Unbiasedness	14
2.3 PSML estimator	16
2.4 Low-complexity PSML	18
2.4.1 MBP-PSML	19
2.4.2 Second-best PSML	22

2.4.3	SA-PSML	24
2.5	Ψ -CRB	27
2.5.1	Empirical Ψ -CRB	28
2.6	Generalization To a Two-Stage Observation Model	30
2.6.1	Two-stage Ψ -unbiasedness	31
2.6.2	Two-stage PSML Estimator and MBP Algorithm	32
2.6.3	Two-stage Ψ -CRB	33
2.7	Simulations	33
2.7.1	Linear Gaussian model	34
2.7.2	Bernoulli model	37
2.7.3	Spectrum estimation after channel selection	41
2.7.4	Spectrum estimation with “black-box” selection rule	43
3	Post-Selection Estimation in Unidentifiable Models	46
3.1	Model and Problem Formulation	47
3.2	Post-parameter-selection Estimation Methods	49
3.2.1	ML Estimator	49
3.2.2	Coherent ML Estimator	49
3.2.3	PSML Estimator	50
3.2.4	Coherent PSML Estimator	50
3.3	Practical Implementation of the coherent PSML estimator	51
3.4	Simulations	52
4	Non-Bayesian Post-Model-Selection Estimation as Estimation Under Model Misspecification	55
4.1	Background: Estimation Under Model Misspecification	56
4.1.0.1	Misspecified CRB	57
4.2	Model: Estimation After Model Selection	58
4.3	Post-Selection Estimation as Estimation Under Model Misspecification	60
4.3.1	Naive Interpretation	60
4.3.2	Normalized Interpretation	61
4.3.3	Selective Inference Interpretation	62
4.4	Post-Model-Selection Estimators	64
4.4.1	Maximum Selected Likelihood (MSL) Estimator	64
4.4.2	Maximum Selected Normalized Likelihood (MSNL) Estimator	65
4.4.3	Post-Selection ML (PSML) Estimator	66
4.5	Post-Model-Selection Performance Analysis	67

4.5.1	Post-Model-Selection MSE	67
4.6	Post-Selection Pseudo-True Parameter Vectors	69
4.6.1	Naive Interpretation	69
4.6.2	Normalized Interpretation	70
4.6.3	Selective Inference Interpretation	71
4.7	Post-Model-Selection Misspecified Cramér-Rao-Type Lower Bounds	72
4.7.1	Post Selection Regularity Conditions	72
4.7.2	Marginal PS-MCRB	73
4.7.3	Global PS-MCRB	74
4.8	Interpretations of the PS-MCRB	76
4.8.1	Naive Interpretation	76
4.8.2	Normalized Interpretation	77
4.8.3	Selective Inference Interpretation	78
4.8.4	Remarks and discussion	79
4.9	Example: Estimation after channel selection	79
4.9.1	Estimators	81
4.9.2	Pseudo-True Parameter Vectors	82
4.9.3	PS-MCRBs	83
4.9.4	Simulation Results	85
5	Bayesian Post-Model-Selection Estimation	89
5.1	Bayesian Post-Model-Selection Estimators	91
5.1.1	Oracle MMSE	91
5.1.2	Oracle coherent MMSE (cMMSE)	91
5.1.3	Selected MMSE (sMMSE)	92
5.1.4	Full MMSE (fMMSE)	93
5.2	Bayesian Post-model-selection bounds	93
5.2.1	Coherent MMSE bound	93
5.2.2	Selective Bayesian Cramér-Rao Bound (BCRB)	94
5.3	Simulations	96
6	Summary and Future Research	99
A		102
A.i	Existence of Ψ -unbiased estimator	102

CONTENTS

B	103
B.i Proof of Proposition 4.1	103
B.ii Proof of Theorem 4.1	104
C	111
C.i Detailed developments for the example in (4.79)	111
C.i.1 Estimators	111
C.i.1.1 Naive Interpretation- MSL Estimator	111
C.i.1.2 Normalized Interpretation- MSNL estimator	112
C.i.1.3 Selective Inference Interpretation- PSML estimator	113
C.i.2 Post-Selection-Pseudo-true Parameters	114
C.i.2.1 Naive Interpretation	114
C.i.2.2 Normalized Interpretation	115
C.i.2.3 Selective Inference Interpretation	115
C.i.3 Post-selection PS-MCRB	116
C.i.3.1 Naive Interpretation	116
C.i.3.2 Normalized Interpretation	117
C.i.3.3 Selective Inference Interpretation	118
C.ii Conditional Expectations	119
C.ii.1 Conditional Expectations	120
C.ii.2 Conditional Covariance Matrices	121
D	122
D.i Proof of Theorem 5.1	122
D.ii Proof of Theorem 5.2	123
D.iii Proof of Remark 5.1 (order relation)	124
Bibliography	125

List of Figures

2.1	Post-selection estimation scheme: First, the parameters of interest is selected based on the observation vector, \mathbf{x} , by a known predetermined selection rule, Ψ . Second, the selected parameter, θ_Ψ , is estimated based on the same observation vector	13
2.2	Linear Gaussian model: The Ψ -bias (a) and PSMSE (b) of the SA-PSML, the second-best PSML, CS, split-the-data, and the ML estimator versus the number of observations, N	37
2.3	Linear Gaussian model: Comparison between the probability of selection and the pairwise probability of selection.	38
2.4	Linear Gaussian model: Run-time of the SA-PSML and the second-best PSML methods versus the number of parameters, M	38
2.5	Linear Gaussian model: The Ψ -bias (a) and PSMSE (b) and mean run-time (c) of the SA-PSML estimator Vs. the number of Monte-Carlo simulations , K where the number of observation is $N = 100, 500, 1200$	39
2.6	Bernoulli case: The Ψ -bias (a) and PSMSE (b) of the SA-PSML and second-best PSML versus Δ , the difference between θ_1 and the rest of the parameters, compared to split-the-data and the ML estimators.	41
2.7	Spectrum estimation: The Ψ -bias of the SA-PSML and second-best PSML versus the number of observations N , compared to split-the-data and the ML estimators.	44
2.8	Spectrum estimation with a “black-box” kNN selection rule: The Ψ -bias (a) and PSMSE (b) of the SA-PSML, split-the-data, and ML estimators versus the number of observations, N	45

LIST OF FIGURES

3.1	Post-selection estimation scheme: First, the set of significant parameters is selected based on the observation vector, \mathbf{x} , by a known predetermined selection rule, which results in a selected support set, $\hat{\Lambda}$. Second, the selected parameters, $\boldsymbol{\theta}_{\hat{\Lambda}}$, are estimated based on the same observation vector, while the unselected parameters, $\boldsymbol{\theta}_{\hat{\Lambda}^c}$, are estimated as zero.	48
3.2	The PSMSE of the SA-cPSML compared to the coherent-ML	54
4.1	The MSE of the MSL, MSNL, PSML, and the oracle ML estimators, and PS-MCRBs and the oracle CRB versus the threshold, γ , where the true hypothesis is \mathcal{H}_1	86
4.2	The bias (a) and MSE (b) of the MSL, MSNL, PSML, and the oracle ML estimators, and PS-MCRBs and the oracle CRB versus the threshold, γ , where the true hypothesis is \mathcal{H}_2	86
4.3	The MSE of the MSL, MSNL, PSML, and the oracle ML estimators, and PS-MCRBs and the oracle CRB versus M , where the true hypothesis is \mathcal{H}_2 for the BIC selection rule.	87
4.4	The MSE of the MSL, MSNL, PSML, and the oracle ML estimators, and PS-MCRBs and the oracle CRB versus the SNR, where the true hypothesis is \mathcal{H}_1 for the AIC selection rule.	88
5.1	The MSE of the oracle MMSE, cMMSE, sMMSE, and fMMSE estimators compared to the oracle BCRB, selective BCRB, and selective TBCRB, versus $\frac{1}{\sigma^2}$	98
5.2	The sMSE of the cMMSE and sMMSE estimators compared with the selective BCRB and the selective TBCRB versus the FA rate, α	98

Glossary

BCRB Bayesian Cramér-Rao bound
cdf Cumulative distribution function
CR Cognitive radio
CRB Cramér-Rao bound
DOA Direction-of-arrival
FIM Fisher information matrix
i.i.d. Independent identically distributed
KLD Kullback-Leibler divergence
MAP Maximum *a-posteriori* probability
MBP Maximization by parts
MCRB Misspecified Cramér-Rao bound
ML Maximum likelihood
MML Misspecified maximum likelihood
MMSE Minimum mean-squared-error
MSE Mean-squared-error
MSSE Misspecified squared-error
MSMSE Misspecified mean squared-error
MRU Minimum risk unbiased
MVU Minimum variance unbiased
OMP Orthogonal matching pursuit
pdf Probability density function
PSFIM Post-selection Fisher information matrix
PSLL Post-selection log-likelihood
PSML Post-selection maximum likelihood
PSMSE post-selection mean squared-error
PSSE post-selection squared-error
SA Stochastic approximation
SNR Signal-to-noise ratio

Notations

\mathbb{R}^n	The real coordinate n-dimension space
\mathbb{C}^n	The complex coordinate n-dimension space
\mathbf{a}	Vector
$a_m, [\mathbf{a}]_\alpha$	m th element of \mathbf{a} and subvector of \mathbf{a} with index set α
\mathbf{A}	Matrix
$[\mathbf{A}]_{\alpha,\beta}$	Submatrix of \mathbf{A} with rows from index set α and columns from index set β
$\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta})$	gradient of $g(\boldsymbol{\theta})$
\mathbf{I}_M	$M \times M$ identity matrix
\mathbf{e}_m	m th column of \mathbf{I}_M
$\mathbf{0}$	Vector/matrix of zeros
$\mathbf{A} \succeq \mathbf{0}, \mathbf{A} \succ \mathbf{0}$	Positive-semidefinite and positive-definite matrices
$\text{Tr}(\cdot)$	Trace
$\text{vec}(\cdot)$	Vectorization
$\log(\cdot)$	Natural logarithm
$ \cdot $	Absolute value, cardinality of set
$\mathcal{P}(\cdot)$	The power set
$(\cdot)^c$	The complement-set
$(\cdot)^T, (\cdot)^H$	Transpose and conjugate transpose
$\text{diag}(\cdot)$	Diagonal matrix
$(\cdot)^{-1}, (\cdot)^\dagger$	Inverse and Moore-Penrose pseudo-inverse
$\mathbf{P}_\mathbf{A}$	Orthogonal projection matrices onto the column space of \mathbf{A}
$\mathbf{P}_\mathbf{A}^\perp$	Orthogonal projection matrices onto the null of \mathbf{A}
$\ \cdot\ _p$	The ℓ_p norm
$\ \cdot\ $	The standard Euclidean ℓ_2 -norm or the induced ℓ_2 -norm
$\mathbb{1}_A$	The indicator function of an event A
$\text{E}[\cdot], \text{E}[\cdot Z]$	Expectation and conditional expectation given an event Z

Chapter 1

Introduction

In this chapter, we introduce the framework of post-selection estimation. To that end, we present the scientific background for parameter estimation. Then, we introduce the framework for estimation after selection - first, we introduce the concept of estimation after parameter selection and the concept of estimation after model selection. Within the discussion of estimation after model selection, we present the background for estimation under model misspecification, which emerges as a central concern in the context of estimation after model selection.

1.1 Scientific Background

Estimation is a procedure that aimed to approximate the values of certain parameters based on observations from random variables or random process. Parameter estimation problems are usually performed within one of two frameworks: the Bayesian and the non-Bayesian approaches. In the former, we assume that the parameters for estimation are random variables, while in the later, in the absence of prior knowledge regarding the statistical properties of the parameters, we assume unknown deterministic parameters.

Let $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$ be a parameter vector of interest, where $\Omega_{\boldsymbol{\theta}}$ represents the parameter space. We aim to recover the vector $\boldsymbol{\theta}$ based on a set of observations vectors $\mathbf{x} \in \Omega_{\mathbf{x}}$, where $\Omega_{\mathbf{x}}$ is the the observation space. An estimator, $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \rightarrow \Omega_{\boldsymbol{\theta}}$, is a mapping from the observation space onto the parameter space that attempts to recover the value of $\boldsymbol{\theta}$. We define a cost function, $\mathcal{C} : \Omega_{\boldsymbol{\theta}} \times \Omega_{\boldsymbol{\theta}} \rightarrow \mathbb{R}$, where $\mathcal{C}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ is the cost of estimating $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}$. The risk is defined as the expectation of the cost-function,

$$\mathcal{R}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \triangleq \text{E}[\mathcal{C}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})]. \quad (1.1)$$

1.1.1 Non-Bayesian Estimation

The non-Bayesian estimation approach, also known as frequentist estimation, assumes that the parameters of interest are unknown deterministic parameters. Generally, non-Bayesian risks depend on the parameters to be estimated; thus, unrestricted minimization of the risk, as in the Bayesian approach, yields the trivial estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. This is not realizable since $\boldsymbol{\theta}$ is the unknown vector we are seeking. Therefore, we need to consider a restriction on the potential estimators. A common approach to avoid the trivial estimator involves implementing bias restriction. In particular, it is common to impose *unbiasedness* of the estimator [1–5]. However, it is well known that the conventional mean-unbiasedness criterion is not always the most appropriate criterion. Therefore, a generalization of the concept of unbiasedness that takes into account the cost-function was suggested in [4, 6] and is named as “Lehmann-unbiasedness”. This definition for unbiasedness requires that a uniformly unbiased estimator w.r.t. a cost-function is “closer” to the value of the parameter than to any other value in the parameter space, where the “closeness” is measured by the expectation of the cost-function.

Definition 1.1 (Lehmann-Unbiasedness). *An estimator, $\hat{\boldsymbol{\theta}}$, is said to be an unbiased estimator of $\boldsymbol{\theta}$ in the Lehmann sense w.r.t. the cost function $\mathcal{C}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ if*

$$\mathbb{E}_{\boldsymbol{\theta}}[\mathcal{C}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})] \leq \mathbb{E}_{\boldsymbol{\theta}}[\mathcal{C}(\hat{\boldsymbol{\theta}}, \boldsymbol{\eta})], \quad \forall \boldsymbol{\theta}, \boldsymbol{\eta} \in \Omega_{\boldsymbol{\theta}}, \quad (1.2)$$

where the expectation is parameterized by $\boldsymbol{\theta}$.

For example, for the squared-error cost function that produces the mean squared-error (MSE) risk, we obtain that the Lehmann-unbiasedness definition coincides with the mean-unbiasedness criterion [6]. Different examples for additional cost functions were developed in [7–13]. In [14, 15] the Lehmann-unbiasedness was defined for hybrid estimation problems, where the set of parameters was composed of both deterministic and random parameters [16].

Ideally, in the non-Bayesian parameter estimation approach we aim to find an estimator that is uniformly minimum risk unbiased (MRU), i.e. an estimator that is uniformly unbiased and achieves minimum risk. Thus, the MRU is the solution of the following optimization problem:

$$\hat{\boldsymbol{\theta}}^{(\text{MRU})} = \arg \min_{\hat{\boldsymbol{\theta}} \in \Omega_{\boldsymbol{\theta}}^{(UB)}} \mathcal{R}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}), \quad (1.3)$$

where $\Omega_{\boldsymbol{\theta}}^{(UB)} \subset \Omega_{\boldsymbol{\theta}}$ is the set of all unbiased-estimators in the Lehmann sense w.r.t. \mathcal{R} . Unfortunately, the uniformly MRU estimator may not always be attainable, and even if it is, it might be intractable to find [1].

The Cramér-Rao bound (CRB) [17, 18] provides a lower bound on the MSE of any mean-unbiased estimator. It is commonly used as a performance benchmark in non-Bayesian estimation due to its simplicity compared to other lower bounds. An estimator that uniformly achieves the CRB is named an *efficient estimator*; in particular, it is a uniformly minimum variance unbiased (MVU) estimator, the MRU definition for the MSE risk. Cramér-Rao-type bounds for several risks were developed in [8–13]. The Bhattacharyya bound [19] is a tighter generalization of the CRB based on high-order derivatives of the likelihood function. Additional lower bounds include the Barankin bound [20], which is the tightest lower bound on the MSE of unbiased estimators. However, the Barankin bound is usually impractical due to its complexity. Therefore, several bounds were developed as approximations of the Barankin bound, such as the Hammersley-Chapman-Robbins bound [21, 22], and others [23–26]. In [27], a new class of non-Bayesian bounds was presented, which is based on a generalization of the derivative, and sampling operators appear in [23–26] with an the integral transform.

Since the MRU estimator is frequently impractical, and may not exist, a commonly used method is the maximum-likelihood (ML) estimator. Intuitively, the motivation behind this estimator is to find the values of the parameters that are most likely to “explain” the observations. Therefore, the ML estimator is given by

$$\hat{\boldsymbol{\theta}}^{(\text{ML})} \triangleq \arg \max_{\boldsymbol{\theta}} f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}), \quad (1.4)$$

where $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})$ is the probability density function (pdf) of the observation set, \mathbf{x} , parameterized by the deterministic parameter vector $\boldsymbol{\theta}$. The ML estimator is popular due to its asymptotic properties. It can be shown that (at least) asymptotically, the ML estimator is unbiased and achieves the CRB; thus, it is asymptotically optimal. Furthermore, if an efficient estimator exists, it coincides with the ML estimator. Another important property is the invariance property of the ML estimator: that if $\boldsymbol{\mu} = \mathbf{g}(\boldsymbol{\theta})$ then the ML estimation of $\boldsymbol{\mu}$ is obtained by $\mathbf{g}(\hat{\boldsymbol{\theta}}^{(\text{ML})})$.

1.1.2 Bayesian Estimation

The Bayesian approach is based on modeling the parameter vector, $\boldsymbol{\theta}$, as a random parameter vector based on a random observation vector \mathbf{x} . This approach is based on the assumption of available prior knowledge about the statistical properties of $\boldsymbol{\theta}$, that can be utilized in the estimation process [1, 2]. A risk-optimal estimation is obtained by the

estimator that minimizes the risk function, i.e.

$$\hat{\boldsymbol{\theta}}^{(\text{opt})} \triangleq \arg \min_{\boldsymbol{\theta}} \mathcal{R}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}). \quad (1.5)$$

For the commonly used MSE risk, the optimal estimator, which achieves the minimum mean squared-error (MMSE), is obtained by the conditional expectation of the parameter vector, $\boldsymbol{\theta}$, given the observation vector, \mathbf{x} :

$$\hat{\boldsymbol{\theta}}^{(\text{MMSE})} \triangleq \text{E}[\boldsymbol{\theta}|\mathbf{x}]. \quad (1.6)$$

Another common estimation method in the Bayesian framework is the maximum *a-posteriori* probability (MAP) estimator [1, 2], defined by

$$\hat{\boldsymbol{\theta}}^{(\text{MAP})} \triangleq \arg \max_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}|\mathbf{x}}(\boldsymbol{\theta}|\mathbf{x}), \quad (1.7)$$

where $f_{\boldsymbol{\theta}|\mathbf{x}}(\boldsymbol{\theta}|\mathbf{x})$ is the conditional pdf of $\boldsymbol{\theta}$ given \mathbf{x} .

Bayesian lower bounds on the MSE of any estimator can be partitioned into three main as detailed below.

1. The Weiss-Weinstein class: These bounds [28] are based on covariance inequalities (e.g. [2, p. 33]). This class include the Bayesian CRB (BCRB) [2], Bayesian Bhattacharyya bound [19], Bobrovsky-Zakai bound [29], and Bayesian Abel bound [30]. In [31], a generalization of the Weiss-Weinstein class is proposed. This generalization is based on integral transforms that generalize the traditional derivative and sampling operators of functions used on the Weiss-Weinstein bounds. The Reuven-Messer [32] bound on the MSE of a hybrid vector, that is some of its entries are deterministic while other are random. Nevertheless, its Bayesian version, where all the parameters are random, can be included in the Weiss-Weinstein class.
2. The Ziv-Zakai class: This class [33] relates the MSE in the estimation problems to the probability of error in a hypothesis testing problem. This class includes the Ziv-Zakai bound [33], Bellini-Tartara bound [34], the Chazan-Ziv-Zakai bound [35], the Weinstein bound [36], the extended-Ziv-Zakai bound [37], and the Bell bound [38].
3. The MMSE class: By disregarding practical constraints or incorporating additional knowledge, estimators can be devised that may not be practical for actual estimation purposes. However, their MSE values serve as lower bounds when analyzing the performance of practical estimators. In particular, , we refer to the oracle or

clairvoyant estimators that leverage knowledge of latent or unknown parameters (e.g. [39–41]).

1.2 Post-Selection Estimation

Classic estimation theory relies on the assumption that the observation model is correctly known and that the parameters of interest are predetermined. However, this assumption does not hold in many practical applications of signal processing, communication, and data science. In practice, the exact model is unknown in many cases, and before the estimation procedure, there is a preliminary data-based decision to select a model, a.k.a. a *model selection* procedure. Even if the model is known, in many scenarios, only some of the unknown parameters are of interest, and some may be considered nuisance parameters. Nevertheless, in practice, the decision about which parameters are of interest and which parameters are nuisances is a data-based selection procedure made by a predetermined selection rule. In both cases, the selection stage impacts subsequent estimation, for example, by introducing selection bias through over-optimistic inferences, non-covering confidence intervals, and invalid performance bounds.

1.2.1 Estimation After Parameter Selection

Parameter selection, as a preliminary step in estimation problems, plays an essential role in modern signal processing and data analysis. For example, in cognitive radio communications [42, 43], a selection of parameters of interest may be based on the signal-to-noise ratio (SNR), energy, or transmission rate, and the parameters of interest can be channel gain and noise variance of only the selected channel. In neuroimaging analysis [44–46], a subset of voxels in the brain are selected for further analysis based on their activity pattern in functional scans. In medical trials [47–52], preliminary tests select the best treatments for a large-scale clinical trial, and then, only the parameters of the selected treatment are essential. As a result, the problem of estimation after parameter selection, i.e., estimating a subset of parameters after they are selected based on a data-based selection rule, is of great interest in signal processing, communication systems, and statistical analysis.

In post-selection inference, it is well known that the selection stage has an impact on subsequent estimation by introducing a selection-bias [45, 49, 53–55]. Usually, in such cases, there is no uniformly unbiased estimator [56]. Bias-correction methods for specific parametric models and specific estimators have been suggested in [49, 50, 54, 55, 57–61].

It has been shown in various cases [47–50, 52, 59] that a two-stage scheme, in which the parameters are estimated based on a two-stage data model, where only the first-stage data is used for the selection, enables the derivation of uniformly unbiased estimators. In contrast, for the single-stage approach, a uniformly unbiased estimator does not exist. Therefore, using sequential multistage schemes the selection bias can be reduced [62], and substantial estimation performance gains are achievable.

1.2.2 Estimation After Model Selection

Traditional estimation methods and performance analysis are based on the assumption that the observation model is correctly specified. However, this assumption does not hold in many signal processing, communication, and data science applications. In many practical parameter estimation problems, the exact model is unknown. Therefore, before estimation, a model-selection stage is performed by a predetermined data-based selection rule, where the selected model may be different from the true one, which affects the consequent estimation approach. This problem arises, for example, in direction-of-arrival (DOA) estimation where there is a detection stage before the DOA estimation [63] and in post-model-selection estimation of rain level from commercial microwave links [64]. Post-model-selection estimation can be described as a two-stage approach: in the first stage, the model is selected from the candidate models based on the observations; in the second stage, the parameters of the selected model are estimated based on the same observations. The selection stage is conducted according to a predetermined data-based selection rule .

Performance bounds and estimation methods have been developed for post-parameter-selection estimation [9, 65, 66], where a subset of parameters of interest is selected based on the data before the estimation. In [67–69], the bounds and estimators were developed for cases where the informative data region is selected before the estimation.

Post-model-selection estimation has been discussed as a vital component of the framework of selective inference. In [70–79], it was shown that ignoring the model-selection procedure may cause overoptimistic inferences and non-covering confidence intervals, and introduces selection bias. In the context of signal processing, [63, 80] presented the effect of a preliminary detection step on the estimation procedure. In [13, 81], post-model-selection Cramér-Rao-type lower bounds on the MSE were developed for non-Bayesian and Bayesian parameter estimation, respectively.

1.2.3 Estimation Under Model Misspecification

Estimation under model misspecification refers to the scenario where the considered, assumed model observation model may differ from the actual model that generate the observations, a.k.a. the true model. Estimation under misspecified models has been discussed in the literature and garnered renewed attention in recent years. In particular, [82–84] discuss the asymptotic properties of the ML estimator under a misspecified model, also known as the misspecified ML (MML) estimator. For Bayesian parameter estimation, [85, 86] discuss the properties of Bayesian estimators under misspecified models. A Cramér-Rao-type bound that accounts for model misspecification, the misspecified CRB (MCRB), was developed in [87, 88] and was discussed in and derived for several scenarios in [89–92]. Modifications of the MCRB, such as a constrained MCRB for problems involving equality constraints [93] and a generalized MCRB for estimation problems in which the Hessian matrix is singular [94], were introduced. In [95], a cyclic MCRB, which is a lower bound on the mean cyclic error for periodic estimation problems under model misspecification, was derived. In [96, 97], the MCRB was used to design a model-selection procedure. In [98], a bilateral bound on the MSE under model mismatch that is applicable for Bayesian and non-Bayesian approaches was developed. Despite the elegant and useful theory presented in these works, none of the existing works deals with post-model-selection estimation, i.e. selection that is data-dependent, and these works do not consider the procedure that led to the selection of the misspecified model. That is, the problem of post-model-selection estimation has not yet been explored from the viewpoint of estimation under a misspecified model. In conventional estimation under model misspecification, there is a clear definition of the assumed pdf. However, in post-model-selection estimation, there are several candidate models.

1.3 Contributions of the Dissertation

The main objective of this research is to establish new Bayesian and non-Bayesian frameworks for estimation after selection. In particular, we consider the problems of estimation after parameter selection and estimation after model selection. The contribution of this research is manifested in developing new estimation methods that are aware of the selection procedure. We present both theoretical and practical estimators and develop tractable algorithms for low-complexity implementation of these estimators. In addition, for performance analysis, we establish appropriate performance bounds that account for the characteristics of post-selection estimation.

- **Estimation after parameter selection:** We develop practical, low-complexity estimation methods for multivariate cases where the Post-selection maximum likelihood (PSML) estimator is intractable. We implement the maximization by parts (MBP) algorithm, which is based on decomposing the likelihood function into “easily optimized” and complicated parts, adapted to the specific setting of post-selection estimation. We show the convergence of the MBP-PSML algorithm based on the “information dominance” of the Fisher information matrix (FIM) over the information contained in the selection approach. The MBP-PSML algorithm was used to develop two low-complexity estimation methods. The second-best PSML method uses the probability of selection between the two highest-ranked parameters in terms of the selection rule. The stochastic approximation PSML (SA-PSML) method is based on a Monte Carlo approximation of the intractable gradient of the probability of selection in the post-selection log-likelihood maximization, which we then plug into the MBP-PSML algorithm. We develop the empirical post-selection FIM (PSFIM) and the empirical CRB-type lower bound for low-complexity performance analysis. The research related to this topic is presented in Chapter 2 and was published in [65].

- **Estimation in Unidentifiable Models:** We consider the case of estimation in problems where estimation of all the unknown parameters under a given model is time-consuming, not economical, or even impractical. In these cases, a selection stage that aims to identify the significant parameters subset is usually conducted prior to the estimation. A common practice is to tackle these two tasks separately: first, a “parameters of interest” selection stage is conducted, and then, in the second stage, the selected unknown parameters are estimated by conventional estimation methods and the unselected parameters are usually set to zero. The use of the same data for the selection step and for the estimation may affect the performance. We develop a practical algorithm, the stochastic approximation coherent PSML (SA-cPSML), for the implementation of the coherent PSML estimator, which accounts for the selection approach.

The research related to this topic is presented in Chapter 3 and was published in [66].

- **Non Bayesian estimation after model selection:** We address the problem of estimation after model selection via the theory and methods developed for estimation under model misspecification. We present three interpretations of post-model-selection estimation as an estimation under model misspecification problem: 1) the commonly-used naive interpretation that defines the assumed pdf as the pdf of the

selected model, which results in a non-valid pdf; 2) the normalized interpretation, which is obtained by normalizing the naive pdf to a valid pdf, but creates coupling between the selected and the unselected parameters; 3) a selective inference interpretation, which is both valid (normalized) and without unnecessary coupling between the parameters, and is consistent with the use of conditional likelihood in selective inference. We derive the corresponding MML estimator for each interpretation and discuss the relations between the estimators. For performance analysis, we develop a novel MCRB, the post-model-selection MCRB (PS-MCRB), that considers both the model-selection procedure and the model misspecification. We derive the PS-MCRB for each of the interpretations and discuss their properties. The research related to this topic is presented in Chapter 4 and was published in [99].

- **Bayesian estimation after model selection:** We investigate the problem of post-model-selection estimation for the case of *random* parameters with a *deterministic* unknown support set, where the support set represents the unknown model. We present different estimators: coherent/non-coherent and practical/theoretical. We develop three Bayesian bounds on the sMSE of any coherent estimator: the coherent MMSE bound, the selective BCRB, and the selective TBCRB. We discuss the concept of coherent estimators that estimate the unselected parameters (by a given model-selection rule) to their prior mean. Then, we develop performance bounds for coherent estimators. We evaluate the performance of the various estimators and compare them to the proposed bounds. We show that the proposed bounds are tighter than the oracle BCRB, which ignores the model-selection stage. Thus, these bounds are useful for practical performance analysis. The research related to this topic is presented in Chapter 5 and was published in [81].

1.4 List of Publications

The following publications are based on the research presented in this dissertation.

1. N. Harel and T. Routtenberg, “Non-Bayesian Post-Model-Selection Estimation as Estimation Under Model Misspecification”, *in IEEE Transactions on Signal Processing*, vol. 72, pp. 3641-3657, July 2024 [99]
2. N. Harel and T. Routtenberg, “Bayesian Estimation After Model Selection”, *in IEEE Signal Processing Letters* vol. 28, pp. 175-179, Jan. 2021. [81]

-
3. N. Harel and T. Routtenberg, “Low-Complexity Methods for Estimation After Parameter Selection”, in *IEEE Transactions on Signal Processing vol. 68*, pp. 1152-1167, Jan. 2020. [65]
 4. N. Harel and T. Routtenberg, “Post-Parameter-Selection Maximum-Likelihood Estimation”, in *2021 IEEE Statistical Signal Processing Workshop (SSP), Rio de Janeiro, Brazil, July, 2021*. [66]
 5. M. Khatib, N. Harel, Y. Ben-Horin, Y. Radzyner, and T. Routtenberg, “Cyclic Misspecified Cramér-Rao Bound for Periodic Parameter Estimation,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, April, 2024*. [95]
 6. M. Khatib, N. Harel, T. Routtenberg, and J Tabrikian, “New Derivation of the Misspecified Cramér-Rao Bound,” *In preparation*.

1.5 Outline

The dissertation is organized as follows: in Chapters 2–4, we address post-selection estimation problems in a non-Bayesian framework. In Chapter 5, we address post-selection estimation problems in a Bayesian framework. The research in Chapters 2 and 3 refers to the problem of estimation after parameter selection. In Chapters 4 and 5, we deal with estimation after model selection problem. Specifically, in Chapter 2, we derive new low-complexity estimation methods for non-Bayesian estimation after the parameter selection framework. We generalized the model to a two-stage estimation after parameter selection architecture and adapted the results for this model. In addition, For low-complexity performance analysis, we developed an empirical version of the Ψ -CRB, which is the appropriate Cramér-Rao-type bound for estimation after parameter selection. In Chapter 3, we discuss non-Bayesian estimation in non-identifiable models. In such a scenario, a subset of parameters is selected for estimation from the full unknown parameter vector by a data-based selection rule. We developed a low-complexity estimation method for this scenario. In Chapter 4, we present three interpretations to address the problem of non-Bayesian post-model-selection estimation as a problem of estimation under model misspecification. We developed the corresponding misspecified ML estimator and the misspecified CRB for each of these interpretations. Finally, in Chapter 5, we investigate the post-model-selection Bayesian parameter estimation of a random vector with an unknown deterministic support set that represents the model. We present different estimators, both theoretical and practical. For low-complexity performance analysis, we

developed the selective BCRB and selective tighter BCRB, which are lower bounds on the MSE for any coherent estimator.

1.6 Notations

In the following, vectors are denoted by boldface lowercase letters and matrices by boldface uppercase letters. The indicator function of an event A is denoted by $\mathbb{1}_A$ and the identity matrix is denoted by \mathbf{I} . The notations Λ^c , $|\Lambda|$, and $\mathcal{P}(\Lambda)$ denote the complement-set, cardinality, and power set of a set, Λ , respectively. The operators $(\cdot)^T$, $(\cdot)^H$, and $(\cdot)^{-1}$ denote the transpose, Hermitian transpose, and inverse, respectively. The operator $\|\cdot\|$ applied to a vector denotes the standard Euclidean l_2 -norm, while applied to a matrix denotes the induced l_2 -norm. The (m, k) th element and the m th column of the matrix \mathbf{A} are denoted by $[\mathbf{A}]_{m,k}$ and $[\mathbf{A}]_{:,m}$, respectively. The matrix \mathbf{A}_Λ denotes the submatrix of \mathbf{A} consisting of the columns indexed by Λ . Similarly, \mathbf{a}_Λ is the vector consisting of the elements of \mathbf{a} indexed by Λ . The notation $\mathbf{A} \succeq \mathbf{B}$ implies that $\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix, where \mathbf{A} and \mathbf{B} are Hermitian matrices of the same size. The m th element of the gradient vector $\nabla_{\boldsymbol{\theta}} c$ is given by $\frac{\partial c}{\partial \theta_m}$, where $\boldsymbol{\theta} = [\theta_1, \dots, \theta_M]^T$, c is a scalar function of $\boldsymbol{\theta}$, $\nabla_{\boldsymbol{\theta}} c \triangleq (\nabla_{\boldsymbol{\theta}} c)$, and $\nabla_{\boldsymbol{\theta}}^2 c \triangleq \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^H c$. The notations $E_p[\cdot]$ and $E_p[\cdot|A]$ represent the expectation and conditional expectation w.r.t. given event A , respectively.

Chapter 2

Low-Complexity Methods for Estimation After Parameter Selection

In this chapter, we consider estimation after parameter selection framework. In Section 2.1 we present the model scheme for estimation after parameter selection. In Section 2.2, we present the post-selection mean squared error (PSMSE) as an appropriate performance criterion for estimation after parameter selection and present the correspondence unbiasedness definition in the Lehmann sense. Then, in Section 2.3 we present the post-selection maximum likelihood (PSML) estimator. In Section 2.4, we develop low complexity algorithm for implementation of the PSML estimator. In Section 2.5, we present a Cramér-Rao-type bound on the PSMSE and derive an empirical algorithm to compute it in case it is intractable. Then, in Section 2.6, we extend the observation acquisition scheme into a two-stage model and show that all the results are valid for the extension. Finally, in Section 2.7 the proposed methods are evaluated via simulation.

2.1 Observation Model

Let $(\Omega, \mathcal{F}, P_\theta)$ denote a probability space, where Ω is the observation space, \mathcal{F} is the σ -algebra, and P_θ is a probability measure on \mathcal{F} that is parameterized by a real deterministic parameter vector $\boldsymbol{\theta} = [\theta_1, \dots, \theta_M]^T \in \mathbb{R}^M$. This probability space is assumed to be in the Hilbert space of absolutely integrable functions w.r.t. the corresponding probability measure.

We consider the problem of estimating the unknown parameter vector, $\boldsymbol{\theta}$, based on observations from Ω , gathered in two stages. Let $\mathbf{x} \in \Omega_{\mathbf{x}}$ be the observation vector with

the pdf, $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})$. A data-based selection rule, $\Psi_{\mathbf{x}} : \Omega_{\mathbf{x}} \rightarrow \{1, \dots, M\}$, is a deterministic function that selects a parameter of interest based on the observation vector, \mathbf{x} . That is, if $\Psi_{\mathbf{x}} = m$, then the estimation goal is to estimate the selected parameter, θ_m . We denote by $\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$ the probability that θ_m is the selected parameter $\forall m = 1, \dots, M$, where it is assumed that the deterministic sets

$$\mathcal{A}_m \triangleq \{\mathbf{x} \in \Omega_{\mathbf{x}} : \Psi_{\mathbf{x}} = m\} \quad (2.1)$$

are partitions of $\Omega_{\mathbf{x}}$. Thus, $\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) = \Pr(\mathbf{x} \in \mathcal{A}_m)$. We assume a non-redundant setting in which $\Psi_{\mathbf{x}}$ is not a sufficient statistic for estimating $\boldsymbol{\theta}$ based on \mathbf{x} , thus, hierarchical Bayesian model [6] perspective will not simplify our model. Finally, we denote by $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \rightarrow \mathbb{R}^M$ an estimator of $\boldsymbol{\theta}$ based on the observation vector, \mathbf{x} . It should be noted that in this work we take the selection rule for granted and discuss the estimation of the selected parameter that emerged from this given selection. The proposed estimation after parameter selection architecture is presented schematically, by in Fig. 2.1.

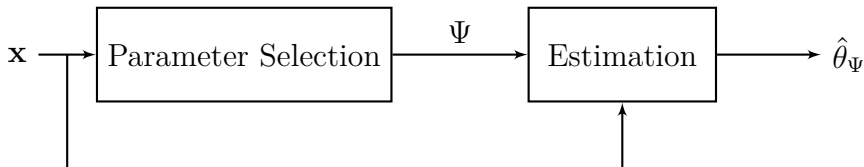


Figure 2.1: Post-selection estimation scheme: First, the parameters of interest is selected based on the observation vector, \mathbf{x} , by a known predetermined selection rule, Ψ . Second, the selected parameter, θ_{Ψ} , is estimated based on the same observation vector

2.1.1 Special Cases

Some special cases of the considered two-stage model from Section 2.1 are described in the following.

1. *Independent populations:* In many practical situations, it is common to compare several populations, select the desired one, and estimate the parameters associated with the selected population. The model of two-stage estimation after selection with independent populations, which we presented in our earlier work [62], is a classic model in mathematical statistics (see e.g. [47, 49–52, 59]). In this model, a given set of M independent populations is assumed, where each population has an associated observation vector, \mathbf{x}_m , with a marginal pdf, $f_m(\mathbf{x}_m; \theta_m)$, parameterized by a single unknown parameter, θ_m , $\forall m = 1, \dots, M$. Thus, the selection of a parameter θ_m is equivalent to the selection of the m th population. In this setting, only samples from

the selected population are acquired in the second observation stage. Thus, in this case, the observation pdf from (2.60) is the joint pdf of all populations:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \prod_{k=1}^M f_k(\mathbf{x}_k; \theta_k), \quad \forall \mathbf{x} \in \mathcal{A}_m. \quad (2.2)$$

In adaptive clinical trials [47, 49–52], the populations may represent different medical treatments and the selection rule may select the treatment with the highest estimated life expectancy; in this case, the variance and the mean of the selected treatment are usually the parameters to estimate and the populations are usually assumed to be Gaussian. In the context of signal processing, the populations may represent independent channels for CR communications, as described in Subsection 2.7.3, or for speech recognition [100]. Fast multistage processing is vital for rapid wide-band sensing of channels; therefore, the model for estimation after parameter selection may provide great benefits in estimation performance.

2. *Data-independent selection rule:* The randomized selection rule satisfies $\Pr(\Psi_{\mathbf{x}}^{(\text{rand})} = m; \boldsymbol{\theta}) = p_m$, where $\{p_m\}_{m=1}^M \in [0, 1] \forall m = 1, \dots, M$, are constant. That is, the selection of the parameter of interest is independent of the data. In particular, for $p_m = \mathbb{1}_{m=m_0}$, where θ_{m_0} is the parameter of interest, we obtain the well-known problem of non-Bayesian estimation in the presence of additional deterministic nuisance parameters [101–103].

2.2 PSMSE and Ψ -Unbiasedness

In this subsection, we present the PSMSE risk and its corresponding unbiasedness definition in the Lehmann sense. These definitions are an extension of similar results that we developed in [9, 10] for the single-stage model.

The problem of estimation after parameter selection can be interpreted as the problem of estimation of a parameter of interest in the presence of nuisance parameters, where it is unknown in advance which is the parameter of interest; this decision is made based on the data by the selection rule. In the presence of nuisance parameters, only estimation errors of the parameter of interest should be taken into consideration via the marginal squared-error cost of this parameter [101, 103, Eq. (1)]. Therefore, in post-selection estimation, the appropriate cost function is the squared error of the *data-based* selected parameter of interest, $\theta_{\Psi_{\mathbf{x}}}$:

$$C^{(\Psi)}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \triangleq (\hat{\theta}_{\Psi_{\mathbf{x}}} - \theta_{\Psi_{\mathbf{x}}})^2, \quad (2.3)$$

for a given selection rule, $\Psi_{\mathbf{x}}$. By using the properties of the indicator function, the post-selection squared-error (PSSE) from (2.3) can be rewritten as

$$C^{(\Psi)}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \triangleq \sum_{m=1}^M (\hat{\theta}_m - \theta_m)^2 \mathbb{1}_{\{\Psi_{\mathbf{x}}=m\}}. \quad (2.4)$$

The corresponding PSMSE, which is the expected cost function, is obtained by using (2.4) and the law of total expectation:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} [C^{(\Psi)}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] &= \sum_{m=1}^M \mathbb{E}_{\boldsymbol{\theta}} [(\hat{\theta}_m - \theta_m)^2 \mathbb{1}_{\{\Psi_{\mathbf{x}}=m\}}] \\ &= \sum_{m=1}^M \mathbb{E}_{\boldsymbol{\theta}} [(\hat{\theta}_m - \theta_m)^2 | \Psi_{\mathbf{x}} = m] \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}). \end{aligned} \quad (2.5)$$

The PSMSE in (2.5), which is the MSE over the selected parameter, is widely used in the mathematical statistics literature and in practical experiment design [50, 57, 58, 60, 61, 104, 105].

In non-Bayesian estimation, an unbiasedness restriction is usually imposed to exclude trivial estimators. The Lehmann unbiasedness definition [7], as presented in Subsection 1.1.1, generalizes the concept of mean-unbiasedness to unbiasedness w.r.t. the considered cost function (see e.g. [8, 12, 13, 103]). The Lehmann unbiasedness for estimation after parameter selection w.r.t. the PSSE cost-function in (2.4), named Ψ -unbiasedness, is defined as follows.

Proposition 2.1. (*Ψ -unbiasedness*) *The estimator $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \rightarrow \mathbb{R}^M$ is an Ψ -unbiased estimator in the Lehmann sense w.r.t. the PSSE cost function if*

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} [\hat{\theta}_m - \theta_m | \Psi_{\mathbf{x}} = m] &= 0, \\ \forall m = 1, \dots, M \text{ such that } \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) &\neq 0. \end{aligned} \quad (2.6)$$

Proof. Substitution of the PSSE cost-function from (2.4) in the Lehmann unbiasedness definition from (1.2) results in

$$\mathbb{E}_{\boldsymbol{\theta}} [C^{(\Psi)}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] \leq \mathbb{E}_{\boldsymbol{\theta}} [C^{(\Psi)}(\boldsymbol{\eta}, \hat{\boldsymbol{\theta}})], \quad \forall \boldsymbol{\eta}, \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}. \quad (2.7)$$

Substitution of (2.5) in (2.7) results in the following

$$\begin{aligned} & \sum_{m=1}^M \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_m - \theta_m)^2 | \Psi_{\mathbf{x}} = m \right] \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) \\ & \leq \sum_{m=1}^M \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_m - \eta_m)^2 | \Psi_{\mathbf{x}} = m \right] \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}), \quad \forall \boldsymbol{\eta}, \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}. \end{aligned} \quad (2.8)$$

The expectation in the r.h.s. of (2.8) can be rewritten as

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_m - \eta_m)^2 | \Psi_{\mathbf{x}} = m \right] &= \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_m - \theta_m + \theta_m - \eta_m)^2 | \Psi_{\mathbf{x}} = m \right] \\ &= \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_m - \theta_m)^2 | \Psi_{\mathbf{x}} = m \right] + 2(\theta_m - \eta_m) \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_m - \theta_m) | \Psi_{\mathbf{x}} = m \right] + (\theta_m - \eta_m)^2 \end{aligned} \quad (2.9)$$

Substitution of (2.9) in (2.8) results in the following equivalent inequality

$$\begin{aligned} & \sum_{m=1}^M 2(\eta_m - \theta_m) \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_m - \theta_m) | \Psi_{\mathbf{x}} = m \right] \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) \\ & \leq \sum_{m=1}^M (\theta_m - \eta_m)^2 \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}), \quad \forall \boldsymbol{\eta}, \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}. \end{aligned} \quad (2.10)$$

We can notice that the condition in (2.6) is sufficient for (2.10) to be satisfied. In the following we will show that this is also necessary. The inequality needs to be satisfied $\forall \boldsymbol{\eta} \in \Omega_{\boldsymbol{\theta}}$, in particular, $\boldsymbol{\eta} = \boldsymbol{\theta} + \epsilon \mathbf{e}_k$ for some $\epsilon \in \mathbb{R}$ and \mathbf{e}_k is the k th elementary unit vector. Substitution of that $\boldsymbol{\eta}$ in (2.10) results in

$$2\epsilon \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_k - \theta_k) | \Psi_{\mathbf{x}} = k \right] \Pr(\Psi_{\mathbf{x}} = k; \boldsymbol{\theta}) \leq \epsilon^2 \Pr(\Psi_{\mathbf{x}} = k; \boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}. \quad (2.11)$$

If $\Pr(\Psi_{\mathbf{x}} = k; \boldsymbol{\theta}) = 0$ then the inequality in (2.11) is satisfied. However, if $\Pr(\Psi_{\mathbf{x}} = k; \boldsymbol{\theta}) > 0$, the l.h.s. of (2.11) is positive for any $\epsilon \neq 0$. Since ϵ can be either positive or negative, to satisfy (2.11), (2.6) is a necessary condition. \square

2.3 PSML estimator

Similar to the uniformly minimum variance unbiased estimator [1, p. 20], the uniformly minimum risk unbiased estimator is an estimator that is uniformly Ψ -unbiased and achieves minimum PSMSE, does not always exist and may be intractable. Therefore, similarly to the commonly-used ML estimator,

$$\hat{\boldsymbol{\theta}}^{(\text{ML})} \triangleq \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{x}; \boldsymbol{\theta}), \quad \forall \mathbf{x} \in \Omega_{\mathbf{x}}. \quad (2.12)$$

We define the PSML estimator as

$$\hat{\boldsymbol{\theta}}^{(\text{PSML})} \triangleq \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{x} | \Psi_{\mathbf{x}} = m; \boldsymbol{\theta}), \quad \forall \mathbf{x} \in \mathcal{A}_m. \quad (2.13)$$

By substituting (2.62) in (2.13) we obtain that the PSML can be decomposed as follows:

$$\hat{\boldsymbol{\theta}}^{(\text{PSML})} = \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{x}; \boldsymbol{\theta}) - \log \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}), \quad \forall \mathbf{x} \in \mathcal{A}_m. \quad (2.14)$$

The PSML estimator from (2.14) can be interpreted as a penalized ML estimator [106,107], where the penalty term is $-\log \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$, i.e. the penalty term is specifically designed to compensate for the selection approach. However, since the penalty term is not a probability density w.r.t. $\boldsymbol{\theta}$, and since we do not have any additional prior information, the PSML does not have a Bayesian interpretation. The maximization on the r.h.s. on (2.13) can be interpreted as the maximization step in the expectation-maximization algorithm [108]. However, in contrast to EM, in the considered post-selection scheme, the likelihood is a tractable function. The PSML estimator has been shown to have better performance than the ML estimator, in terms of Ψ -bias and PSMSE, in various scenarios [109,110]. Moreover, it has been shown in [9], similarly to the ML estimator and the conventional efficiency, that if an Ψ -efficient estimator exists, then it coincides with the PSML estimator for the selected parameter. An Ψ -efficient estimator, as defined in Definition 3 in [9], is an Ψ -unbiased estimator that achieves the Ψ -CRB on the PSMSE, which is given in Section 2.5. Thus, in this case, the PSML estimator is the minimum PSMSE unbiased estimator. In addition, it was shown in [111,112] that under mild conditions the conditional ML estimator is a consistent estimator w.r.t. the conditional pdf. Thus, we can conclude that the PSML estimator from (2.14) is a consistent estimator w.r.t. the conditional pdf from (2.62). If the selection rule is a consistent rule, then asymptotically, the influence of the selection process on the PSML decreases, since the probability $\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$ on the r.h.s. of (2.14) converges to a specific value. In this case, the PSML from (2.14) coincides with the ML from (2.12). The contribution of the probability of selection, $\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$, is more significant as there is more ambiguity in the selection process, such as in the case of close hypotheses.

We define the following regularity conditions:

- C.1. The PSL function, $\log f(\mathbf{x} | \Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$, is a concave function w.r.t. $\boldsymbol{\theta}$.
- C.2. The gradient vector of the PSL function, $\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x} | \Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$, exists and is finite, $\forall \boldsymbol{\theta} \in \mathbb{R}^M, \mathbf{x} \in \mathcal{A}_m$.

Under the regularity Conditions C.1 and C.2, the PSML estimator from (2.14) can be

obtained as the solution to the following score equation:

$$\nabla_{\theta} \log f(\mathbf{x}; \theta) - \mathbf{g}(\theta) = \mathbf{0}, \quad \forall \mathbf{x} \in \mathcal{A}_m, \quad (2.15)$$

where the gradient of the probability of selection is defined as

$$\mathbf{g}(\theta) \triangleq \nabla_{\theta} \log \Pr(\Psi_{\mathbf{x}} = m; \theta). \quad (2.16)$$

In general, an analytical solution of (2.15) is intractable. Yet, in many cases the unconditional log-likelihood function, $\log f(\mathbf{x}; \theta)$, is tractable, and may even be separable w.r.t. to the unknown parameters. Thus, the conventional ML can be easily found and the intractability of (2.15) stems from:

- A. the computation of $\nabla_{\theta} \log \Pr(\Psi_{\mathbf{x}} = m; \theta)$, which involves high-dimensional integration that does not have a closed form expression.
- B. the maximization may require a multi-dimensional grid search, where the computational complexity increases with the dimension of θ .

Hence, there is a need for practical, low-complexity estimation methods that use the special structure of the PSSL, as well as the tractability of the conventional log-likelihood part, $f(\mathbf{x}; \theta)$, to approximate the solution for the score equation from (2.15).

2.4 Low-complexity PSML

In this section, we develop low-complexity methods for estimation after parameter selection. We assume that the conventional ML estimator from (2.12), which ignores the selection, is tractable and develop low-complexity methods for maximizing the PSSL. In Subsection 2.4.1 we apply the MBP algorithm from [113] to solve iteratively the optimization problem on the r.h.s. of (2.14). Since the proposed MBP-PSML algorithm requires the evaluation of the gradient of the probability of selection from (2.16) at any iteration point, we propose low-complexity methods that are based on the MBP algorithm: the second-best PSML method and the SA-PSML method in Subsection 2.4.2 and Subsection 2.4.3, respectively.

2.4.1 MBP-PSML

The MBP algorithm [113] is an iterative optimization technique that divides a general log-likelihood function, $\ell(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$, with the observations, \mathbf{x} and \mathbf{y} , into two parts as follows:

$$\ell(\mathbf{x}; \boldsymbol{\theta}) = \ell_s(\mathbf{x}; \boldsymbol{\theta}) + \ell_c(\mathbf{x}; \boldsymbol{\theta}), \quad (2.17)$$

where ℓ_c is the complicated, intractable part and ℓ_s is the simple part, in the sense that solving the scoring equation, $\nabla_{\boldsymbol{\theta}} \ell_s(\mathbf{x}; \boldsymbol{\theta}) = 0$, is simple. The MBP algorithm is usually initialized by the solution of this scoring equation of the simple part, ℓ_s . Then, at the i th iteration, it evaluates the gradient of the complicated part, ℓ_c , at the previous point and updates the solution by solving

$$\nabla_{\boldsymbol{\theta}} \ell_s(\hat{\boldsymbol{\theta}}^{(i)}) = -\nabla_{\boldsymbol{\theta}} \ell_c(\hat{\boldsymbol{\theta}}^{(i-1)}). \quad (2.18)$$

This procedure is repeated until convergence. Unlike other numerical methods, such as Newton-Raphson and Fisher scoring, the MBP algorithm does not require the second order derivatives of the objective function.

We apply the MBP algorithm to solve the maximization of the PSSL by using the decomposition in (2.14), where the joint log-likelihood function, is the simple part, i.e. $\ell_s(\mathbf{x}; \boldsymbol{\theta}) = \log f(\mathbf{x}; \boldsymbol{\theta})$ and the log of the probability of selection, is set to be the complicated part, i.e. $\ell_c(\mathbf{x}; \boldsymbol{\theta}) = -\log \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$. Therefore, according to (2.18), the i th iteration of the MBP-PSML procedure updates the estimator, $\hat{\boldsymbol{\theta}}^{(i)}$, to be the solution of

$$\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \hat{\boldsymbol{\theta}}^{(i)}) = \nabla_{\boldsymbol{\theta}} \log \Pr(\Psi_{\mathbf{x}} = m; \hat{\boldsymbol{\theta}}^{(i-1)}), \quad \forall \mathbf{x} \in \mathcal{A}_m. \quad (2.19)$$

By substituting (2.16) evaluated at $\hat{\boldsymbol{\theta}}^{(i-1)}$ in (2.19), the i th iteration estimator, $\hat{\boldsymbol{\theta}}^{(i)}$, can be written as the solution of:

$$\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \hat{\boldsymbol{\theta}}^{(i)}) = \mathbf{g}(\hat{\boldsymbol{\theta}}^{(i-1)}), \quad \forall \mathbf{x} \in \mathcal{A}_m. \quad (2.20)$$

The initial estimator at $i = 0$ is set to be the ML estimator, $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}^{(\text{ML})}$.

It is well known that if an efficient estimator of $\boldsymbol{\theta}$ exists, then the gradient of the log-likelihood function can be written as [114]:

$$\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta}) (\hat{\boldsymbol{\theta}}^{(\text{ML})} - \boldsymbol{\theta}), \quad (2.21)$$

where

$$\mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{\theta}} \left[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}; \boldsymbol{\theta}) \right] \quad (2.22)$$

is the conventional FIM, which is assumed to be a non-singular matrix throughout this chapter. By substituting the tractable term from (2.21) in the MBP-PSML iteration from (2.20) and by replacing $\mathbf{J}_{\mathbf{x}}^{-1}(\hat{\boldsymbol{\theta}}^{(i)})$ with, $\mathbf{J}_{\mathbf{x}}^{-1}(\hat{\boldsymbol{\theta}}^{(i-1)})$, we obtain that the MBP-PSML update iteration is linear for this special case with the following update equation:

$$\hat{\boldsymbol{\theta}}^{(i)} = \hat{\boldsymbol{\theta}}^{(\text{ML})} - \mathbf{J}_{\mathbf{x}}^{-1}(\hat{\boldsymbol{\theta}}^{(i-1)}) \mathbf{g}(\hat{\boldsymbol{\theta}}^{(i-1)}). \quad (2.23)$$

If an efficient estimator does not exist, the iteration update in (2.23) can still be used as an approximation to (2.20), which is obtained by using a Taylor series, similarly to the development of the Fisher scoring method for conventional likelihood [1, Ch. 7.7]. It should be noted that under our assumption that the conventional log-likelihood is simple, the conventional FIM in (2.23) is usually tractable as well, in contrast to the post-selection FIM, which is discussed in Section 2.5. Thus, the proposed iteration in (2.23) is tractable, while developing a Fisher-scoring method for the PSL is usually intractable. The MBP-PSML procedure is described in Algorithm 1.

Algorithm 1 MBP-PSML

Require: observation vector, \mathbf{x} , convergence parameter, δ .

- 1: set $m = \Psi_{\mathbf{x}}$
- 2: initialize: $i = 0$, $\hat{\boldsymbol{\theta}}^{(0)}(\mathbf{x}) = \hat{\boldsymbol{\theta}}^{(\text{ML})}$
- 3: **repeat**
- 4: set $i = i + 1$
- 5: solve (2.20) or its approximation in (2.23) to obtain the next iteration: $\hat{\boldsymbol{\theta}}^{(i)}$
- 6: **until** $\left\| \hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}}^{(i-1)} \right\| \leq \delta$

Ensure: MBP-PSML estimator, $\hat{\boldsymbol{\theta}}^{(\text{MBP-PSML})} = \hat{\boldsymbol{\theta}}^{(i)}$

In the following, we establish the convergence of the MBP-PSML method, where the PSL function is analytically known. To this end, we define additional regularity conditions:

- C.3. The Hessian matrix of the PSL, $\nabla_{\boldsymbol{\theta}}^2 \log f(\mathbf{x} | \Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$, exists and is finite, $\forall \boldsymbol{\theta} \in \mathbb{R}^M$, $\mathbf{x} \in \mathcal{A}_m$.
- C.4. The operations of integration w.r.t. \mathbf{x} , and differentiation w.r.t. $\boldsymbol{\theta}$ can be inter-

changed $\forall \boldsymbol{\theta} \in \mathbb{R}^M$ for any differentiable and measurable function, $q(\mathbf{x}, \boldsymbol{\theta})$:

$$\int_{\Omega_{\mathbf{x}}} \nabla_{\boldsymbol{\theta}} q(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = \nabla_{\boldsymbol{\theta}} \int_{\Omega_{\mathbf{x}}} q(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x}. \quad (2.24)$$

Theorem 2.1. (*MBP-PSML convergence*) Under regularity Conditions C.1–C.4, the MBP-PSML from Algorithm 1 converges to the PSML estimator from (2.14).

Proof. The convergence of the MBP algorithm for the general case is discussed in [113, Sec. 4]. It is shown that the MBP algorithm converges asymptotically to the PSML estimator under regularity Conditions C.1–C.4 and under a certain “information dominance” condition. By substituting our two-stage estimation after parameter selection model, the information dominance condition for the MBP-PSML can be written as

$$\left\| \mathbf{J}_{\mathbf{x}}^{-1}(\boldsymbol{\theta}) \mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}(\boldsymbol{\theta}) \right\| < 1, \quad (2.25)$$

where $\mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta})$ is defined in (2.22) and

$$\mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)} \triangleq \nabla_{\boldsymbol{\theta}} \log \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}), \quad (2.26)$$

$\forall m = 1, \dots, M$. The matrix $\mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}$ in (2.26) can be interpreted as the Fisher information content of the selection stage. We assume that the selection rule, $\Psi_{\mathbf{x}}$, is not a sufficient statistic for the estimation of $\boldsymbol{\theta}$ from \mathbf{x} . Thus, the extension of the data processing inequality for Fisher information [115] implies that

$$\mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)} \prec \mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta}). \quad (2.27)$$

The inequality in (2.27) implies that the information content of the selection step, which is based on the observation vector \mathbf{x} , is less than the whole information contained in the observation vector, \mathbf{x} . Since $\mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta})$ is assumed to be a positive definite matrix and $\mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}$ is a positive semidefinite matrix, (2.27) implies that [116, Th. 7.7.3]

$$\mathbf{J}_{\mathbf{x}}^{-1}(\boldsymbol{\theta}) \mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}(\boldsymbol{\theta}) \prec \mathbf{I}, \quad (2.28)$$

and that (2.25) is satisfied, which guarantees the convergence of the MBP-PSML algorithm to the PSML estimator. \square

The proposed MBP-PSML algorithm requires the evaluation of $\mathbf{g}(\boldsymbol{\theta})$ from (2.16) at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(i)}$ for each iteration. Usually, high-dimensional integration is required to compute

the probability of selection. In the following, we develop low-complexity methods that use the MBP-PSML algorithm but do not require the analytical form of $\mathbf{g}(\boldsymbol{\theta})$.

2.4.2 Second-best PSML

In this subsection, we implement the MBP-PSML algorithm from Subsection 2.4.1, where we replace the probability of selection by the probability of the selection between the two highly-ranked parameters in $\boldsymbol{\theta}$ in terms of $\Psi_{\mathbf{x}}$. Thus, using the second-best scheme, the computation of the probability for the challenging M -parameter selection problem reduces to a much simpler task of a selection between the best two parameters. For example, for the population model from Subsection 2.1.1, if the selection rule selects the population with the largest mean, then the second-best parameter is the parameter which is associated with the second largest sample mean.

For a specific observation vector, \mathbf{x} , let $\theta_{\tilde{m}}$ denotes the second-best parameter, i.e. the parameter that would be selected in the absence of the selected parameter. That is, $\mathbf{x} \in \tilde{\mathcal{A}}_{m, \tilde{m}}$, where $\tilde{\mathcal{A}}_{m, \tilde{m}}$ is the subset of \mathcal{A}_m such that the second best selection is $\theta_{\tilde{m}}$. Thus, $\mathcal{A}_m = \bigcup_{k=1, k \neq m}^M \tilde{\mathcal{A}}_{m, k}$. We consider a pairwise selection between θ_m and $\theta_{\tilde{m}}$ by Ψ , where the selection of other parameters in $\boldsymbol{\theta}$ is prohibited. We suggest replacing the probability $\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$ in the PSML from (2.14) by the pairwise probability

$$\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \boldsymbol{\theta}) \triangleq \Pr(\Psi_{\mathbf{x}} = m | \Psi_{\mathbf{x}} \in \{m, \tilde{m}\}; \boldsymbol{\theta}). \quad (2.29)$$

That is, we suggest the following second-best PSML estimator:

$$\hat{\boldsymbol{\theta}}^{(2\text{B-PSML})} \triangleq \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{x}; \boldsymbol{\theta}) - \log \Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \boldsymbol{\theta}), \quad \forall \mathbf{x} \in \tilde{\mathcal{A}}_{m, \tilde{m}}. \quad (2.30)$$

Similarly to in the case of the PSML estimator from (2.14), under the regularity Conditions C.1 and C.2, the second-best PSML estimator from (2.30) can be obtained as the solution to the following score equation:

$$\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\theta}) - \tilde{\mathbf{g}}(\boldsymbol{\theta}) = \mathbf{0}, \quad \forall \mathbf{x} \in \tilde{\mathcal{A}}_{m, \tilde{m}}, \quad (2.31)$$

where the gradient of the pairwise probability of selection is defined as

$$\tilde{\mathbf{g}}(\boldsymbol{\theta}) \triangleq \nabla_{\boldsymbol{\theta}} \log \Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \boldsymbol{\theta}). \quad (2.32)$$

By replacing $\mathbf{g}(\cdot)$ by $\tilde{\mathbf{g}}(\cdot)$ from (2.32) in the MBP-PSML update from (2.20), we obtain

that the i th iteration step of the second-best PSML estimator is given by

$$\nabla_{\theta} \log f(\mathbf{x}; \hat{\theta}^{(i)}) = \tilde{\mathbf{g}}(\hat{\theta}^{(i-1)}), \quad \forall \mathbf{x} \in \tilde{\mathcal{A}}_{m, \tilde{m}}. \quad (2.33)$$

Similarly, the approximation from (2.23) can be replaced by its second-best PSML version:

$$\hat{\theta}^{(i)} = \hat{\theta}^{(\text{ML})} - \mathbf{J}_{\mathbf{x}}^{-1}(\hat{\theta}^{(i-1)}) \tilde{\mathbf{g}}(\hat{\theta}^{(i-1)}) \quad \forall \mathbf{x} \in \tilde{\mathcal{A}}_{m, \tilde{m}}. \quad (2.34)$$

In many cases, although the probability of selection, $\Pr(\Psi_{\mathbf{x}} = m; \theta)$, is intractable, the probability of the pairwise selection, $\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \theta)$, is tractable. By using the conditional probability properties the log-probability of selection can be decomposed as follows:

$$\log \Pr(\Psi_{\mathbf{x}} = m; \theta) = \log \Pr(\Psi_{\mathbf{x}} \in \{m, \tilde{m}\}; \theta) + \log \Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \theta). \quad (2.35)$$

Thus, the use of $\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \theta)$ in (2.30) instead of $\Pr(\Psi_{\mathbf{x}} = m; \theta)$ is equivalent to neglecting the term $\log \Pr(\Psi_{\mathbf{x}} \in \{m, \tilde{m}\}; \theta)$ in the PSLM maximization. Several works analyze scenarios and conditions where this probability is, indeed, negligible [117, 118]. However, usually this probability is non-negligible and the proposed second-best PSML is an ad-hoc method. An exception is for the trivial case when the number of parameters $M = 2$, the selection is pairwise, and, $\tilde{\mathbf{g}}(\theta)$ from (2.32) coincides with $\mathbf{g}(\theta)$ from (2.16). For this special case, the second-best PSML estimator coincides with the PSML estimator for this special case.

In some cases the pairwise selection probability, $\Pr(\tilde{\Psi}_{m, \tilde{m}} = m; \theta)$, is only a function of θ_m and $\theta_{\tilde{m}}$. Thus, (2.32) implies that for these cases $\tilde{g}_k(\theta) = 0, \forall k = 1, \dots, M, k \neq m, \tilde{m}$, and $\tilde{g}_k(\theta) = \tilde{g}_k(\theta_m, \theta_{\tilde{m}})$ for $k = m$ or $k = \tilde{m}$. By substituting these results in (2.34) it can be seen that at the i th iteration the estimator of θ is given by

$$\hat{\theta}^{(i)} = \hat{\theta}^{(\text{ML})} - \left[\mathbf{J}_{\mathbf{x}}^{-1}(\hat{\theta}^{(i-1)}) \right]_{:,m} \tilde{g}_m(\hat{\theta}_m^{(i-1)}, \hat{\theta}_{\tilde{m}}^{(i-1)}) - \left[\mathbf{J}_{\mathbf{x}}^{-1}(\hat{\theta}^{(i-1)}) \right]_{:,\tilde{m}} \tilde{g}_{\tilde{m}}(\hat{\theta}_m^{(i-1)}, \hat{\theta}_{\tilde{m}}^{(i-1)}). \quad (2.36)$$

According to (2.36), in the general case, even in this special case, all the parameters should be updated to obtain the second-best PSML. However, for the following scenarios we can update only the selected and the second best parameters, and set all the others to their ML estimators:

- A. For the independent populations model from Subsection 2.1.1, the FIM, $\mathbf{J}_{\mathbf{x}}(\theta)$, is a diagonal matrix. Thus, in this scenario, (2.36) implies that only the estimators of θ_m and $\theta_{\tilde{m}}$ should be updated at each iteration via (2.36) and the other estimators

of the parameters are equal to their ML value,

$$\hat{\theta}_k^{(2B-PSML)} = \hat{\theta}_k^{(ML)}, \quad \forall k \in \{1, \dots, M\}, \quad k \neq \{m, \tilde{m}\}. \quad (2.37)$$

This scenario is exemplified in the simulations in Subsection 2.7.3.

- B. For the case where the FIM, $\mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta})$, is a constant matrix, such as for location family of pdfs [7], the i th iteration step from (2.36) is reduced to

$$\hat{\boldsymbol{\theta}}^{(i)} = \hat{\boldsymbol{\theta}}^{(ML)} - \left[\mathbf{J}_{\mathbf{x}}^{-1} \right]_{:,m} \tilde{g}_m \left(\hat{\theta}_m^{(i-1)}, \hat{\theta}_{\tilde{m}}^{(i-1)}(\mathbf{x}) \right) - \left[\mathbf{J}_{\mathbf{x}}^{-1} \right]_{:,\tilde{m}} \tilde{g}_{\tilde{m}} \left(\hat{\theta}_m^{(i-1)}, \hat{\theta}_{\tilde{m}}^{(i-1)} \right). \quad (2.38)$$

In this scenario, the update of the estimators of $\hat{\theta}_m^{(i-1)}$ and $\hat{\theta}_{\tilde{m}}^{(i-1)}$ via (2.38) is not a function of the estimators of the other parameters. Since the PSMSE risk, as defined in (2.5), takes into account only the estimation errors of the selected parameter, there is no need to update the non-selected parameters at each iteration that do not affect the estimation of θ_m . Thus, without loss of performance, we can set these estimators to their associated ML estimators, as in (2.37), and only update $\hat{\theta}_m^{(i)}$ and $\hat{\theta}_{\tilde{m}}^{(i)}$ at each iteration. This scenario is exemplified in the simulations in Subsection 2.7.1.

The second-best PSML algorithm is summarized in Algorithm 2.

Algorithm 2 : Second-best PSML

Require: observation vector, \mathbf{x} , convergence parameter, δ .

- 1: set m according to $\Psi_{\mathbf{x}}$
- 2: set \tilde{m} : the second best selection
- 3: initialize: $i = 0$, $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}^{(ML)}$
- 4: **repeat**
- 5: set $i = i + 1$
- 6: solve (2.33) or its approximation in (2.34) to obtain the next iteration, $\hat{\boldsymbol{\theta}}^{(i)}$
- 7: **until** $\left\| \hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}}^{(i-1)} \right\| \leq \delta$

Ensure: second-best estimator, $\hat{\boldsymbol{\theta}}^{(2B-PSML)} = \hat{\boldsymbol{\theta}}^{(i)}$.

2.4.3 SA-PSML

In this subsection, we derive the SA method [119–121]. In particular, by using Monte Carlo averaging, we approximate the multi-dimensional integrals needed to calculate the gradient of the probability of selection from (2.16). Then, the approximated gradient is plugged into the MBP-PSML algorithm from Subsection 2.4.1. To this end, we draw samples directly from the distribution of the first-stage pdf, $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}_0)$, for a given $\boldsymbol{\theta}_0$,

as described, for example, in [122, Ch. 2]. When such generation is impossible, we use Markov chain Monte Carlo (MCMC) samplers [122, Ch. 6] to perform the data generation step.

The probability of selecting the m th parameter can be written as

$$\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) = \int_{\Omega_{\mathbf{x}}} \mathbb{1}_{\{\mathbf{x} \in \mathcal{A}_m\}} f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}. \quad (2.39)$$

By substituting (2.39) in (2.16) we obtain that

$$\mathbf{g}(\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})}{\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})} = \frac{\nabla_{\boldsymbol{\theta}} \int_{\Omega_{\mathbf{x}}} \mathbb{1}_{\{\mathbf{x} \in \mathcal{A}_m\}} f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}}{\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})}. \quad (2.40)$$

Under regularity Condition C.4, the operations of integration w.r.t. \mathbf{x} and differentiation w.r.t. $\boldsymbol{\theta}$ can be interchanged such that

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \int_{\Omega_{\mathbf{x}}} \mathbb{1}_{\{\mathbf{x} \in \mathcal{A}_m\}} f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} &= \int_{\Omega_{\mathbf{x}}} \nabla_{\boldsymbol{\theta}} \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) \mathbb{1}_{\{\mathbf{x} \in \mathcal{A}_m\}} f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \\ &= \mathbf{E}_{\boldsymbol{\theta}} \left[\nabla_{\boldsymbol{\theta}} \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) \mathbb{1}_{\{\mathbf{x} \in \mathcal{A}_m\}} \right]. \end{aligned} \quad (2.41)$$

By substituting (2.41) in (2.40) we obtain that

$$\mathbf{g}(\boldsymbol{\theta}) = \frac{\mathbf{E}_{\boldsymbol{\theta}} \left[\nabla_{\boldsymbol{\theta}} \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) \mathbb{1}_{\{\mathbf{x} \in \mathcal{A}_m\}} \right]}{\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})}. \quad (2.42)$$

The representation in (2.42) allows us to first calculate the gradient, $\nabla_{\boldsymbol{\theta}} \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})$, which is tractable, under our assumptions, and then use Monte Carlo evaluation of the expectation in (2.42). To this end, for any $\boldsymbol{\theta}_0 \in \mathbb{R}^M$, we draw i.i.d. samples, $\{\tilde{\mathbf{x}}^{(k)}\}_{k=1}^K$, from $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}_0)$, and use these samples to approximate:

$$\mathbf{g}(\boldsymbol{\theta}_0) \approx \hat{\mathbf{g}}(\boldsymbol{\theta}_0) \triangleq \frac{\sum_{k=1}^K \nabla_{\boldsymbol{\theta}} \log f_{\mathbf{x}}(\tilde{\mathbf{x}}^{(k)}; \boldsymbol{\theta}_0) \mathbb{1}_{\{\tilde{\mathbf{x}}^{(k)} \in \mathcal{A}_m\}}}{\sum_{k=1}^K \mathbb{1}_{\{\tilde{\mathbf{x}}^{(k)} \in \mathcal{A}_m\}}}. \quad (2.43)$$

In order to avoid numerical errors, if the denominator in (2.43) is smaller than a pre-determined threshold, we set $\hat{\mathbf{g}}(\boldsymbol{\theta}_0) = \mathbf{0}$. From the strong law of large numbers, the

approximation in (2.43) converges almost surely to $\mathbf{g}(\boldsymbol{\theta}_0)$, $\forall \boldsymbol{\theta}_0 \in \mathbb{R}^M$.

Each iteration step of the SA-PSML method, $\hat{\boldsymbol{\theta}}^{(i)}$, is obtained by replacing $\mathbf{g}(\cdot)$ in (2.20) with $\hat{\mathbf{g}}(\cdot)$ from (2.43), i.e. as the solution of

$$\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \hat{\boldsymbol{\theta}}^{(i)}) = \hat{\mathbf{g}}(\hat{\boldsymbol{\theta}}^{(i-1)}). \quad (2.44)$$

Similarly, the approximation in (2.23) can be used with the approximated gradient:

$$\hat{\boldsymbol{\theta}}^{(i)} = \hat{\boldsymbol{\theta}}^{(\text{ML})} - \mathbf{J}_{\mathbf{x}}^{-1}(\hat{\boldsymbol{\theta}}^{(i-1)}) \hat{\mathbf{g}}(\hat{\boldsymbol{\theta}}^{(i-1)}). \quad (2.45)$$

The SA-PSML method is summarized in Algorithm 3.

Algorithm 3 : SA-PSML

Require: observation vector, \mathbf{x} , convergence parameter, δ .

- 1: set m according to $\Psi_{\mathbf{x}}$
- 2: initialize: $i = 0$, $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}^{(\text{ML})}$
- 3: **repeat**
- 4: generate sample vectors $\{\tilde{\mathbf{x}}^{(k)}\}_{k=1}^K \sim f(\mathbf{x}; \hat{\boldsymbol{\theta}}^{(i-1)})$
- 5: evaluate $\hat{\mathbf{g}}(\hat{\boldsymbol{\theta}}^{(i-1)})$ from (2.43)
- 6: solve (2.44) or its approximation from (2.45) to obtain the next iteration: $\hat{\boldsymbol{\theta}}^{(i)}(\mathbf{x}, \mathbf{y})$
- 7: **until** $\|\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}}^{(i-1)}\| \leq \delta$

Ensure: SA-PSML estimator, $\hat{\boldsymbol{\theta}}^{(\text{SA-PSML})} = \hat{\boldsymbol{\theta}}^{(i)}$.

A major advantage of the SA-PSML method is that it does not require the knowledge of the selection rule, but only the ability to apply it given observations. That is, to compute the approximations in (2.43) we can generate the data sets, and insert the data sets into the “black box” to obtain the selection for each data set without the need of any other knowledge. This property is useful where the mechanism of the selection rule is not clear, or is complicated. This problem arises in experimental designs where we have access to a generative model, a black-box that can generate multiple realizations of the selection, while the decision rule is not clear, for example, if the selection rule is based on a chemical or biological reaction [44] whose mathematical model is not clear. Another common example is where the selection rule is based on machine learning classification algorithms [123–125] or a deep neural network classifier, where the analytic representation is usually unknown or very complicated. In Subsection 2.7.4, we demonstrate a scenario where the selection-rule is not specifically known.

2.5 Ψ -CRB

The CRB [1] provides a lower bound on the mean squared error of any mean-unbiased estimator and is used as a benchmark in non-Bayesian estimation. However, the conventional CRB does not take into account the selection process; thus, it is inappropriate for estimation after parameter selection [9, 10, 63]. The Ψ -CRB was developed in [9] as an alternative, and it provides a lower bound on the PSMSE of any Ψ -unbiased estimator. Similar to the PSML estimator, this bound may be intractable. Thus, we develop new low-complexity procedure in order to evaluate this bound.

Theorem 2.2. (*Ψ -CRB [9]*) *Let the regularity Conditions C.2–C.4 be satisfied, and $\hat{\boldsymbol{\theta}}$ be an Ψ -unbiased estimator of $\boldsymbol{\theta}$, with a finite second moment. Then, the PSMSE is bounded by the following Ψ -CRB:*

$$\mathbf{E}_{\boldsymbol{\theta}}[C^{(\Psi)}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] \geq \sum_{m=1}^M \Pr(\Psi(\mathbf{x}) = m; \boldsymbol{\theta}) \left[(\mathbf{J}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta}))^{-1} \right]_{m,m}, \quad (2.46)$$

where the post-selection FIM (PSFIM) is

$$\mathbf{J}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta}) \triangleq \mathbf{E}_{\boldsymbol{\theta}} \left[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}; \boldsymbol{\theta}) | \Psi_{\mathbf{x}} = m \right] - \mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}, \quad (2.47)$$

$\forall m = 1, \dots, M$, and $\mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}$ is defined in (2.26).

Proof. From the covariance-form of the Cauchy-Schwarz inequality

$$\mathbf{E}_{\boldsymbol{\theta}}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T | \Psi = m] \geq \boldsymbol{\Gamma}^{(m)} (\mathbf{J}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta}))^{-1} \boldsymbol{\Gamma}^{(m)}, \quad (2.48)$$

where

$$\boldsymbol{\Gamma}^{(m)} \triangleq \mathbf{E}_{\boldsymbol{\theta}}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x} | \Psi = m; \boldsymbol{\theta}) | \Psi = m]. \quad (2.49)$$

Since the matrices both sides of (2.48) are positive-semidefinite matrices the inequity in (2.48) implies that the diagonal of these matrices satisfy the inequality [116], i.e.

$$\begin{aligned} \mathbf{E}_{\boldsymbol{\theta}}[(\hat{\theta} - \theta)^2 | \Psi = m] &= \left[\mathbf{E}_{\boldsymbol{\theta}}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T | \Psi = m] \right]_{m,m} \\ &\geq \sum_{k=1}^M \sum_{l=1}^M [\boldsymbol{\Gamma}^{(m)}]_{m,k} [(\mathbf{J}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta}))^{-1}]_{k,l} [\boldsymbol{\Gamma}^{(m)}]_{l,m}. \end{aligned} \quad (2.50)$$

Notice that

$$\begin{aligned}
[\mathbf{\Gamma}^{(m)}]_{k,l} &= \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_k - \theta_k) \frac{\partial \log f(\mathbf{x}|\Psi = m; \boldsymbol{\theta})}{\partial \theta_l} \Big| \Psi = m \right] \\
&= \int_{\mathcal{A}_m} (\hat{\theta}_k - \theta_k) \frac{\partial \log f(\mathbf{x}|\Psi = m; \boldsymbol{\theta})}{\partial \theta_l} f(\mathbf{x}|\Psi = m; \boldsymbol{\theta}) d\mathbf{x} \\
&= \int_{\mathcal{A}_m} (\hat{\theta}_k - \theta_k) \frac{\partial f(\mathbf{x}|\Psi = m; \boldsymbol{\theta})}{\partial \theta_l} d\mathbf{x}
\end{aligned} \tag{2.51}$$

Under Condition C.4 and by using integration by parts, (D.6) can be rewritten as follows

$$\begin{aligned}
[\mathbf{\Gamma}^{(m)}]_{k,l} &= \\
&= \frac{\partial}{\partial \theta_l} \int_{\mathcal{A}_m} (\hat{\theta}_k - \theta_k) f(\mathbf{x}|\Psi = m; \boldsymbol{\theta}) d\mathbf{x} - \int_{\mathcal{A}_m} \frac{\partial (\hat{\theta}_k - \theta_k)}{\partial \theta_l} f(\mathbf{x}|\Psi = m; \boldsymbol{\theta}) d\mathbf{x} \\
&= \frac{\partial}{\partial \theta_l} \int_{\mathcal{A}_m} (\hat{\theta}_k - \theta_k) f(\mathbf{x}|\Psi = m; \boldsymbol{\theta}) d\mathbf{x} + \delta_{k,l} \\
&= \frac{\partial}{\partial \theta_l} \mathbb{E}_{\boldsymbol{\theta}} [(\hat{\theta}_k - \theta_k) | \Psi = m] + \delta_{k,l}.
\end{aligned} \tag{2.52}$$

Since $\hat{\boldsymbol{\theta}}$ is assume to be a Ψ -unbiased estimator, i.e. $\mathbb{E}_{\boldsymbol{\theta}} [(\hat{\theta}_m - \theta_m) | \Psi = m] = 0$, therefore,

$$[\mathbf{\Gamma}^{(m)}]_{m,k} = [\mathbf{\Gamma}^{(m)}]_{k,m} = \delta_{m,k}. \tag{2.53}$$

Substitution of (2.53) in (2.50) results in the following marginal bound

$$\mathbb{E}_{\boldsymbol{\theta}} [(\hat{\theta} - \theta)^2 | \Psi = m] \geq [(\mathbf{J}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta}))^{-1}]_{m,m}. \tag{2.54}$$

Since the probabilities of selection are non-negative, by substituting the marginal bounds from (2.54) in the PSMSE from (2.5), we obtain the Ψ -CRB from (2.46). \square

The calculation of the PSFIMs from (2.47) is often intractable due to the need for calculation of the probability of selection and the conditional expectation in (2.47). Similarly to the empirical FIM [126, 127] we propose a Monte Carlo approach to approximate the PSFIMs and the Ψ -CRB inspired by the SA-PSML methods. The proposed SA-PSFIM utilizes the PSL structure and, as a result, the structure of the PSFIM.

2.5.1 Empirical Ψ -CRB

By substituting (2.60), (2.16), and (2.26) in (2.47) we obtain that

$$\mathbf{J}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[\nabla_{\boldsymbol{\theta}} \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) | \Psi_{\mathbf{x}} = m \right] - \mathbf{g}(\boldsymbol{\theta}) \mathbf{g}^T(\boldsymbol{\theta}). \tag{2.55}$$

Under our assumptions, the second-stage FIM, $\mathbf{J}_y(\boldsymbol{\theta})$, can be analytically computed. However, the conditional expectation, as well as $\mathbf{g}(\boldsymbol{\theta})$ on the r.h.s. of (2.55), are intractable. Similarly to the derivation of (2.42), we can rewrite the conditional expectation by using an indicator function as follows

$$\mathbb{E}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) | \Psi_{\mathbf{x}} = m] = \frac{\mathbb{E}_{\boldsymbol{\theta}} \left[\nabla_{\boldsymbol{\theta}} \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) \mathbb{1}_{\{\mathbf{x} \in \mathcal{A}_m\}} \right]}{\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})}. \quad (2.56)$$

Thus, similarly to in the case of the SA-PSML from Subsection 2.4.3, we draw K i.i.d. samples, $\{\tilde{\mathbf{x}}^{(k)}\}_{k=1}^K$, from the true pdf, $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})$. We use these samples to approximate

$$\mathbf{J}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta}) \approx \hat{\mathbf{J}}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta}) \triangleq \frac{\sum_{k=1}^K \nabla_{\boldsymbol{\theta}} \log f_{\mathbf{x}}(\tilde{\mathbf{x}}^{(k)}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f_{\mathbf{x}}(\tilde{\mathbf{x}}^{(k)}; \boldsymbol{\theta}) \mathbb{1}_{\{\tilde{\mathbf{x}}^{(k)} \in \mathcal{A}_m\}}}{\sum_{k=1}^K \mathbb{1}_{\{\tilde{\mathbf{x}}^{(k)} \in \mathcal{A}_m\}}} - \hat{\mathbf{g}}(\boldsymbol{\theta}) \hat{\mathbf{g}}^T(\boldsymbol{\theta}). \quad (2.57)$$

It can be seen that the first term on the r.h.s. of (2.57) approximates the conditional expectation from (2.56) and $\hat{\mathbf{g}}(\boldsymbol{\theta})$ on the second term is approximated using (2.43).

The SA-PSFIM algorithm is summarized in Algorithm 4.

Algorithm 4 : SA-PSFIM

Require: parameter vector $\boldsymbol{\theta}$ and the selection m

- 1: generate sample vectors $\{\tilde{\mathbf{x}}^{(k)}\}_{k=1}^K \sim f(\mathbf{x}; \boldsymbol{\theta})$
- 2: evaluate $\hat{\mathbf{g}}(\boldsymbol{\theta})$ from (2.43)
- 3: evaluate $\hat{\mathbf{J}}_{\mathbf{x}}^{(m)}$ from (2.57)

Ensure: $\hat{\mathbf{J}}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta})$.

By substituting the empirical SA-PSFIM from (2.57) and the probability of selection approximated as $\frac{1}{K} \sum_{k=1}^K \mathbb{1}_{\{\tilde{\mathbf{x}}^{(k)} \in \mathcal{A}_m\}}$ in (2.46), we obtain an approximation of the Ψ -CRB:

$$\hat{B}(\boldsymbol{\theta}) \triangleq \sum_{m=1}^M \frac{1}{K} \sum_{k=1}^K \mathbb{1}_{\{\tilde{\mathbf{x}}^{(k)} \in \mathcal{A}_m\}} \left[\left(\hat{\mathbf{J}}_{\mathbf{x}}^{(m)}(\boldsymbol{\theta}) \right)^{-1} \right]_{m,m}. \quad (2.58)$$

By taking $K \gg M$ in (2.57), the SA-PSFIM can be assumed to be a non-singular matrix, and (2.58) is well-defined.

2.6 Generalization To a Two-Stage Observation Model

In this section, we present a new model for estimation after parameter selection in a two-stage data acquisition scheme. This model is a generalization of the classical two-stage model for independent populations [59]. We derive the two-stage versions of the Ψ -unbiasedness in the Lehmann sense [7] and the PSML estimator, and Ψ -CRB, that extend our single-stage results.

We consider \mathbf{x} presented in Fig. 2.1 as the first-stage of data acquisition. In addition, the data-based selection rule, $\Psi_{\mathbf{x}}$ remains the same. In the second stage of data acquisition, given that the selection is $\Psi_{\mathbf{x}} = m$, a second observation vector, \mathbf{y} , is observed from $\Omega_{\mathbf{y}}^{(m)}$ with a corresponding pdf $f_m(\mathbf{y}; \boldsymbol{\theta})$, $\forall m = 1, \dots, M$. That is, we assume that the conditional pdf of \mathbf{y} given $\Psi_{\mathbf{x}} = m$ is given by

$$f(\mathbf{y}|\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) = f_m(\mathbf{y}; \boldsymbol{\theta}) \quad \forall \mathbf{y} \in \Omega_{\mathbf{y}}^{(m)}. \quad (2.59)$$

It should be noted that the only influence of \mathbf{x} on the second-stage observations, \mathbf{y} , is by choosing the generating observation model, i.e. the specific pdf, $f_m(\cdot)$. This assumption describes a realistic scenario in which the sample-acquisition mechanism is adapted to the selection after the selection. However, the selection by the experimenter does not change the statistical behavior of the observations, which is governed by “nature”. For example, in channel estimation, one may adapt to the selection and acquire only samples from a channel associated with the selected parameter. Yet, the channel’s statistical behavior is the same for all the samples acquired before/after the selection. Therefore, by using (2.59) the joint pdf of the two-stage observation vectors is

$$f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) f_m(\mathbf{y}; \boldsymbol{\theta}), \quad \forall \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \Omega_{\mathbf{y}}, \quad (2.60)$$

where we denote $\Omega_{\mathbf{y}} \triangleq \bigcup_{m=1}^M \Omega_{\mathbf{y}}^{(m)}$ and $\Omega \triangleq \Omega_{\mathbf{x}} \times \Omega_{\mathbf{y}}$. By using these definitions, (2.60), and the rules of marginal probability, the pdf of the second-stage observation vector is given by

$$\begin{aligned} f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta}) &= \int_{\Omega_{\mathbf{x}}} \sum_{m=1}^M f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) f_m(\mathbf{y}; \boldsymbol{\theta}) \mathbb{1}_{\{\mathbf{x} \in \mathcal{A}\}} d\mathbf{x} \\ &= \sum_{m=1}^M f_m(\mathbf{y}; \boldsymbol{\theta}) \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}), \quad \forall \mathbf{y} \in \Omega_{\mathbf{y}}. \end{aligned} \quad (2.61)$$

Additionally, by using Bayes rule it can be verified that

$$f(\mathbf{x}, \mathbf{y}|\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) = \frac{f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})}{\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})}, \quad (2.62)$$

for $\mathbf{x} \in \mathcal{A}, \mathbf{y} \in \Omega_{\mathbf{y}}^{(m)}, \forall m = 1, \dots, M$, where $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ is defined in (2.60). In addition, we define $f(\mathbf{x}, \mathbf{y} | \Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) = 0$, for any $\mathbf{x} \notin \mathcal{A}_m$. Finally, we denote by $\hat{\boldsymbol{\theta}} : \Omega \rightarrow \mathbb{R}^M$ an estimator of $\boldsymbol{\theta}$ based on the two-stage observation vectors, \mathbf{x} and \mathbf{y} .

Notice that if $\Omega_{\mathbf{y}} = \emptyset$, i.e. there are no observations in the second stage, the two-stage model is reduced to our single-stage estimation after parameter selection model presented in [9, 10]. It should be noted that all the results in this chapter are also applicable for a single-stage model. In addition, in the case where (2.60) does not hold, one can merge \mathbf{x} and \mathbf{y} into new single-stage vector and formulate the problem as a single-stage estimation after parameter selection.

In the two-stage model, the estimator, $\hat{\boldsymbol{\theta}}$ is now a function of both \mathbf{x} and \mathbf{y} . Therefore, the conditional expectation on the r.h.s. of (2.5) is calculated by using the joint pdf of the two-stage observations, $f(\mathbf{x}, \mathbf{y} | \Psi_{\mathbf{x}}; \boldsymbol{\theta})$, defined in (2.62), while the selection probability is calculated by using $f(\mathbf{x} | \Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$, i.e. it is only a function of the pdf of the first-stage observations. The PSMSE in (2.5), which is the MSE over the selected parameter, is widely used in the mathematical statistics literature and in practical experiment design [50, 57, 58, 60, 61, 104, 105].

2.6.1 Two-stage Ψ -unbiasedness

In the following we propose the Ψ -unbiasedness for the two-stage model, i.e. the unbiasedness in the Lehmann sense w.r.t. the PSMSE for the two-stage observation model.

Proposition 2.2. (*Ψ -unbiasedness*) *An estimator $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \times \Omega_{\mathbf{y}} \rightarrow \mathbb{R}^M$ is an Ψ -unbiased estimator in the Lehmann sense w.r.t. the PSSE cost function if*

$$E_{\boldsymbol{\theta}} \left[\hat{\boldsymbol{\theta}}_m(\mathbf{x}, \mathbf{y}) - \boldsymbol{\theta}_m | \Psi_{\mathbf{x}} = m \right] = 0, \quad \forall m = 1, \dots, M \text{ such that } \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) \neq 0. \quad (2.63)$$

Proof. This proposition can be proved along the path of the proof of Proposition 2.1 by replacing the single-stage observations pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, with the two-stage pdf, $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ \square

It should be noted that in the Ψ -unbiasedness definition in (2.63), the estimator is a function of the two-stage data, but the conditional expectation is w.r.t. the selection event, which is only a function of the first-stage data. Thus, (2.63) highlights a main advantages of the two-step model in the context of “selection bias”: in various single-stage models, no Ψ -unbiased estimator exists. In contrast, the two-stage model has an Ψ -unbiased estimator [50, 59, 62]. In particular, we prove in Appendix A.i that, for any setting with an existing mean-unbiased estimator without the selection, we can find an

Ψ -unbiased estimator for the two-stage model with at least one sample at the second stage.

2.6.2 Two-stage PSML Estimator and MBP Algorithm

Similarly to (2.14) we define the PSML estimator for the two-stage observation model as

$$\hat{\boldsymbol{\theta}}^{(\text{PSML})}(\mathbf{x}, \mathbf{y}) \triangleq \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{x}, \mathbf{y} | \Psi_{\mathbf{x}} = m; \boldsymbol{\theta}), \quad (2.64)$$

$\forall \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \Omega_{\mathbf{y}}$. By substituting (2.62) in (2.13) we obtain that the PSML can be decomposed as follows:

$$\hat{\boldsymbol{\theta}}^{(\text{PSML})}(\mathbf{x}, \mathbf{y}) = \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) - \log \Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}). \quad (2.65)$$

The MBP algorithm for the PSML estimator under the two-stage model remains the same as presented in Algorithm 1 with the exception of replacing the pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, with the joint pdf $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$. The information dominance condition for convergence of the MBP-PSML under the two-stage model can be written as

$$\left\| \mathbf{J}_{\mathbf{x}, \mathbf{y}}^{-1}(\boldsymbol{\theta}) \mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}(\boldsymbol{\theta}) \right\| < 1, \quad (2.66)$$

where

$$\mathbf{J}_{\mathbf{x}, \mathbf{y}}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{\theta}} \left[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) \right] \quad (2.67)$$

is the conventional two-stage FIM. By using the extension of the data refinement inequality for Fisher information [115] implies that the single-stage information is less than or equal to the two-stage information, i.e.

$$\mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta}) \preceq \mathbf{J}_{\mathbf{x}, \mathbf{y}}(\boldsymbol{\theta}). \quad (2.68)$$

where $\mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta})$ is defined in (2.22). Therefore, since we assumed $\mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta})$ is a non-singular matrix, then $\mathbf{J}_{\mathbf{x}, \mathbf{y}}(\boldsymbol{\theta})$ is also a non-singular matrix. In addition, by using (2.27), (2.68) implies that

$$\mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)} \prec \mathbf{J}_{\mathbf{x}, \mathbf{y}}(\boldsymbol{\theta}). \quad (2.69)$$

Since $\mathbf{J}_{\mathbf{x}, \mathbf{y}}(\boldsymbol{\theta})$ is assumed to be a non-singular matrix it must be a positive definite matrix. In addition, $\mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}$ is a positive semidefinite matrix. Therefore, (2.69) implies that [116, Th. 7.7.3]

$$\mathbf{J}_{\mathbf{x}, \mathbf{y}}^{-1}(\boldsymbol{\theta}) \mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}(\boldsymbol{\theta}) \prec \mathbf{I}, \quad (2.70)$$

and that (2.25) is satisfied, which guarantees the convergence of the MBP-PSML algorithm to the PSML estimator.

2.6.3 Two-stage Ψ -CRB

Theorem 2.3. (*Two-stage Ψ -CRB*) *Let the regularity Conditions C.2–C.4 be satisfied, and $\hat{\boldsymbol{\theta}}$ be an Ψ -unbiased estimator of $\boldsymbol{\theta}$, with a finite second moment. Then, the PSMSE of $\hat{\boldsymbol{\theta}}$ is bounded by the following Ψ -CRB:*

$$\mathbb{E}_{\boldsymbol{\theta}} \left[C^{(\Psi)} \left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}) \right) \right] \geq \sum_{m=1}^M \Pr(\Psi(\mathbf{x}) = m; \boldsymbol{\theta}) \left[\left(\mathbf{J}_{\mathbf{x}, \mathbf{y}}^{(m)}(\boldsymbol{\theta}) \right)^{-1} \right]_{m,m}, \quad (2.71)$$

where the post-selection FIM (PSFIM) is

$$\mathbf{J}_{\mathbf{x}, \mathbf{y}}^{(m)}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{\theta}} \left[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) | \Psi_{\mathbf{x}} = m \right] - \mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}, \quad (2.72)$$

$\forall m = 1, \dots, M$, and $\mathbf{J}_{\Psi_{\mathbf{x}}}^{(m)}$ is defined in (2.26).

Proof. The proof of Theorem 2.3 can be proved along the path of the proof of the single-stage Ψ -CRB proof in Theorem 2.2, and can be obtained by replacing the single-stage observations pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, with the two-stage pdf, $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$. \square

2.7 Simulations

In this section, we evaluate the performance of the following methods:

1. The second-best PSML estimator, $\hat{\boldsymbol{\theta}}^{(2\text{B-PSML})}(\mathbf{x}, \mathbf{y})$, from Algorithm 2.
2. The SA-PSML estimator, $\hat{\boldsymbol{\theta}}^{(\text{SA-PSML})}(\mathbf{x}, \mathbf{y})$ from Algorithm 3.
3. The ML estimator, $\hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{x}, \mathbf{y})$, from (2.12).
4. The split-the-data estimator, which uses only the second-stage observations, \mathbf{y} , for the estimation. We use the following form of the ML estimator based only on \mathbf{y} :

$$\hat{\boldsymbol{\theta}}_{\mathbf{y}}^{(\text{ML})}(\mathbf{y}) \triangleq \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{y}; \boldsymbol{\theta}). \quad (2.73)$$

5. The first-stage ML estimator,

$$\hat{\boldsymbol{\theta}}_{\mathbf{x}}^{(\text{ML})}(\mathbf{x}) \triangleq \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{x}; \boldsymbol{\theta}). \quad (2.74)$$

The performance of these estimators is compared with the empirical Ψ -CRB from Algorithm 4. The conventional CRB is not presented in this section, since it does not provide a valid bound on the PSMSE and it is significantly higher than the estimators' performance.

The Ψ -bias and PSMSE of all estimators was calculated over 50000 Monte Carlo simulations. The maximal number of iterations of the SA-PSML methods is limited to 50. We set the threshold for the denominator in (2.43) to $10^{-7} \frac{N_{\mathbf{x}}^2}{M}$, while the number of generated samples is $K = 1000$.

2.7.1 Linear Gaussian model

The linear Gaussian model with dependent or independent populations is widely used in various applications. In clinical research, several treatments are compared, where each treatment has an unknown treatment effect, modeled as a Gaussian distributed variable [49, 50, 52, 59, 60]. We consider the following model with *correlated* Gaussian populations:

$$\begin{aligned}\mathbf{x}_n &= \mathbf{H}_{\mathbf{x}}\boldsymbol{\theta} + \mathbf{w}_n, & n = 1, \dots, N_{\mathbf{x}} \\ \mathbf{y}_n &= \mathbf{H}_{\mathbf{y}}^{(m)}\boldsymbol{\theta} + \mathbf{v}_n, & n = 1, \dots, N_{\mathbf{y}},\end{aligned}\tag{2.75}$$

where $N_{\mathbf{x}}$ and $N_{\mathbf{y}}$ are the number of samples in the first and second stages, respectively, $\mathbf{H}_{\mathbf{x}} \in \mathbb{R}^{K_{\mathbf{x}} \times M}$, $\mathbf{H}_{\mathbf{y}}^{(m)} \in \mathbb{R}^{K_{\mathbf{y}} \times M}$ are assumed to be known, full-rank matrices, where $\mathbf{H}_{\mathbf{y}}^{(m)}$ is determined according to the first-stage selection from the set of known matrices, $\mathbf{H}_{\mathbf{y}}^{(1)}, \dots, \mathbf{H}_{\mathbf{y}}^{(M)}$. That is, if $\Psi(\mathbf{x}) = m$, the second-stage data, \mathbf{y} , is observed with the matrix $\mathbf{H}_{\mathbf{y}}^{(m)}$. The noise vectors, $\{\mathbf{w}_n\}_{n=1}^{N_{\mathbf{x}}}$ and $\{\mathbf{v}_n\}_{n=1}^{N_{\mathbf{y}}}$ are statistically independent series of time-independent white Gaussian noise vectors with known covariance matrices, $\boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{v}}$, respectively. Therefore the first-stage observation vector is $\mathbf{x} = [\mathbf{x}_1^T, \dots, \mathbf{x}_{N_{\mathbf{x}}}^T]^T$ and the second-stage observation vector is $\mathbf{y} = [\mathbf{y}_1^T, \dots, \mathbf{y}_{N_{\mathbf{y}}}^T]^T$. A commonly-used selection rule is the following [49, 50, 59]:

$$\Psi_{\mathbf{x}} = \arg \max_{m=1, \dots, M} [\hat{\boldsymbol{\theta}}_{\mathbf{x}}^{(\text{ML})}(\mathbf{x})]_m,\tag{2.76}$$

where the single-stage ML estimator from (2.74) is given by

$$\hat{\boldsymbol{\theta}}_{\mathbf{x}}^{(\text{ML})}(\mathbf{x}) = \left(\mathbf{H}_{\mathbf{x}}^T \boldsymbol{\Sigma}_{\mathbf{w}}^{-1} \mathbf{H}_{\mathbf{x}} \right)^{-1} \mathbf{H}_{\mathbf{x}}^T \boldsymbol{\Sigma}_{\mathbf{w}}^{-1} \bar{\mathbf{x}},\tag{2.77}$$

in which $\bar{\mathbf{x}} \triangleq \frac{1}{N_{\mathbf{x}}} \sum_{n=1}^{N_{\mathbf{x}}} \mathbf{x}_n$. If $\mathbf{H}_{\mathbf{x}} = \mathbf{I}$, then, the selection rule from (2.76) is reduced to the commonly-used selection of the largest-mean population. The probability of selection, $\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$, for the rule in (2.76) is intractable for $M > 2$. Thus, the PSML from (2.14) cannot be directly implemented and low-complexity methods are required.

In this case, split-the-data estimator from (2.73) is given by

$$\hat{\boldsymbol{\theta}}_{\mathbf{y}}^{(\text{ML})}(\mathbf{y}) = \left((\mathbf{H}_{\mathbf{y}}^{(m)})^T \boldsymbol{\Sigma}_{\mathbf{v}}^{-1} \mathbf{H}_{\mathbf{y}}^{(m)} \right)^{-1} (\mathbf{H}_{\mathbf{y}}^{(m)})^T \boldsymbol{\Sigma}_{\mathbf{v}}^{-1} \bar{\mathbf{y}}, \quad \mathbf{x} \in \mathcal{A}, \quad (2.78)$$

where $\bar{\mathbf{y}} \triangleq \frac{1}{N_{\mathbf{y}}} \sum_{n=1}^{N_{\mathbf{y}}} \mathbf{y}_n$. According to the proof in the Appendix, the estimator in (2.78) is an Ψ -unbiased estimator of $\boldsymbol{\theta}$. The ML estimator based on both \mathbf{x} and \mathbf{y} from (2.12) is given by

$$\hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{x}, \mathbf{y}) = \mathbf{J}_{\mathbf{x}, \mathbf{y}}^{-1} \left(N_{\mathbf{x}} \mathbf{H}_{\mathbf{x}}^T \boldsymbol{\Sigma}_{\mathbf{w}}^{-1} \bar{\mathbf{x}} + N_{\mathbf{y}} (\mathbf{H}_{\mathbf{y}}^{(m)})^T \boldsymbol{\Sigma}_{\mathbf{v}}^{-1} \bar{\mathbf{y}} \right), \quad (2.79)$$

$\forall \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \Omega_{\mathbf{y}}$, where the conventional two-stage FIM from (2.67) is given by

$$\mathbf{J}_{\mathbf{x}, \mathbf{y}} = N_{\mathbf{x}} \mathbf{H}_{\mathbf{x}}^T \boldsymbol{\Sigma}_{\mathbf{w}}^{-1} \mathbf{H}_{\mathbf{x}} + N_{\mathbf{y}} (\mathbf{H}_{\mathbf{y}}^{(m)})^T \boldsymbol{\Sigma}_{\mathbf{v}}^{-1} \mathbf{H}_{\mathbf{y}}^{(m)}. \quad (2.80)$$

In order to derive the second-best PSML for this scenario, we examine the probability of pairwise selection from (2.29) for the selection rule (2.76),

$$\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \boldsymbol{\theta}) = \Pr(\hat{\boldsymbol{\theta}}_m^{(\text{ML})}(\mathbf{x}) \geq \hat{\boldsymbol{\theta}}_{\tilde{m}}^{(\text{ML})}(\mathbf{x})) = \Phi \left(\boldsymbol{\Delta}_{m, \tilde{m}}^T \boldsymbol{\theta} \right), \quad (2.81)$$

where \tilde{m} is the index of the second-best selection, $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard Gaussian pdf and cumulative distribution function (cdf), respectively, and

$$\boldsymbol{\Delta}_{m, \tilde{m}} \triangleq \frac{1}{\sqrt{(\mathbf{e}_m - \mathbf{e}_{\tilde{m}})^T \mathbf{J}_{\mathbf{x}}^{-1} (\mathbf{e}_m - \mathbf{e}_{\tilde{m}})}} (\mathbf{e}_m - \mathbf{e}_{\tilde{m}}), \quad (2.82)$$

where

$$\mathbf{J}_{\mathbf{x}} = N_{\mathbf{x}} \mathbf{H}_{\mathbf{x}}^T \boldsymbol{\Sigma}_{\mathbf{w}}^{-1} \mathbf{H}_{\mathbf{x}}, \quad (2.83)$$

and \mathbf{e}_m is the m th column vector of the identity matrix. By substituting (2.81) in (2.32), it can be verified that

$$\tilde{\mathbf{g}}(\boldsymbol{\theta}) = \frac{\phi(\boldsymbol{\Delta}_{m, \tilde{m}}^T \boldsymbol{\theta})}{\Phi(\boldsymbol{\Delta}_{m, \tilde{m}}^T \boldsymbol{\theta})} (\mathbf{e}_m - \mathbf{e}_{\tilde{m}}). \quad (2.84)$$

For this case, $\hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{x}, \mathbf{y})$ is an efficient estimator [1, Ch. 7]. Therefore, we can use (2.23) at the iteration step of the SA-PSML methods. By using the efficiency of this case and substituting (2.84) in (2.34), each iteration of the second-best PSML is given by

$$\hat{\boldsymbol{\theta}}^{(i)}(\mathbf{x}, \mathbf{y}) = \hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{x}, \mathbf{y}) - \mathbf{J}_{\mathbf{x}, \mathbf{y}}^{-1} \frac{\phi \left(\boldsymbol{\Delta}_{m, \tilde{m}}^T \hat{\boldsymbol{\theta}}^{(i-1)} \right)}{\Phi \left(\boldsymbol{\Delta}_{m, \tilde{m}}^T \hat{\boldsymbol{\theta}}^{(i-1)} \right)} (\mathbf{e}_m - \mathbf{e}_{\tilde{m}}), \quad (2.85)$$

$\forall \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \Omega_{\mathbf{y}}$. For this Gaussian model we also compare the results with the James-Stein shrinkage estimator [128, 129]:

$$\hat{\boldsymbol{\theta}}^{(\text{JS})}(\mathbf{x}, \mathbf{y}) = \left(1 - \frac{M-2}{(\hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{x}, \mathbf{y}))^T \mathbf{J}_{\mathbf{x}, \mathbf{y}} \hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{x}, \mathbf{y})} \right) \hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{x}, \mathbf{y}), \quad (2.86)$$

and the extension of the Cohen-Sackrowitz (CS) estimator [59] for correlated populations, as described in [60, Eq. (2.1)]. It was shown in [59, 60] that the CS estimator satisfies an unbiasedness condition that is stricter than our Ψ -unbiasedness definition from (2.6). Thus, the CS estimator is also an Ψ -unbiased estimator. However, it has strict requirements [60] and, thus, it has poor PSMSE performance, as shown in the following simulations. These estimators are specifically designed for the linear Gaussian model and there is no solution for the general case.

In Fig. 2.2 the Ψ -bias and PSMSE of the different estimators are presented versus the total number of observations, $N = N_{\mathbf{x}} + N_{\mathbf{y}}$, such that $N_{\mathbf{x}} = 0.8N$, $N_{\mathbf{y}} = 0.2N$, $M = 25$, $\mathbf{H}_{\mathbf{x}} = \mathbf{H}_{\mathbf{y}}^{(m)} = \mathbf{I}, \forall m = 1, \dots, M$, $\boldsymbol{\Sigma}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{v}} = \boldsymbol{\Sigma}$, such that $[\boldsymbol{\Sigma}]_{i,j} = (1 + |i - j|)^{-2}$, and $\theta_1 = 1.05, \theta_2 = 1.01, \theta_3 = 1.02, \theta_k = 1, k = 2, \dots, M-1, \theta_M = 0$. It can be seen that the proposed PSML methods have lower Ψ -bias and PSMSE than the ML estimator. The CS estimator, $\hat{\boldsymbol{\theta}}^{(\text{CS})}$, and the split-the-data estimator, $\hat{\boldsymbol{\theta}}_{\mathbf{y}}^{(\text{ML})}(\mathbf{y})$, are Ψ -unbiased estimators, but the unbiasedness comes at the expense of the PSMSE, which is higher even than the PSMSE of the ML estimator in this case. In addition, this figure demonstrates that the empirical Ψ -CRB is a lower bound on the PSMSE of the Ψ -unbiased estimators, $\hat{\boldsymbol{\theta}}^{(\text{CS})}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{y}}^{(\text{ML})}(\mathbf{y})$, and is achieved asymptotically by the PSML estimators. The James-Stein shrinkage estimator dominates the ML estimator, but it is dominated by the PSML methods. Similarly to the variance-bias trade off, the PSML methods are Ψ -biased, but achieve lower PSMSE than the unbiased methods and than the empirical Ψ -CRB.

In Fig. 2.3 we compared the probability of selection, $\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})$, and the pairwise probability of selection from (2.29), $\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \boldsymbol{\theta})$, for the setting of Figs. 2.2(a) and 2.2(b) for $m = 1$ and $\tilde{m} = 3$. The probability of selection was calculated numerically while for the pairwise probability of selection was calculated analytically according to (2.81). It can be seen that these two probabilities coincide only asymptotically, which explains the advantage of the SA-PSML over the second-best PSML outside the asymptotic region.

To demonstrate the complexity of the proposed methods for different problem dimensions, the average processing period, “runtime”, is evaluated by running the algorithms using Matlab 2017b on an Intel Xeon(TM) Processor E5-2660 v4. Fig. 2.4 shows the runtime of the PSML method versus the number of unknown parameters, M , for $N = 250$

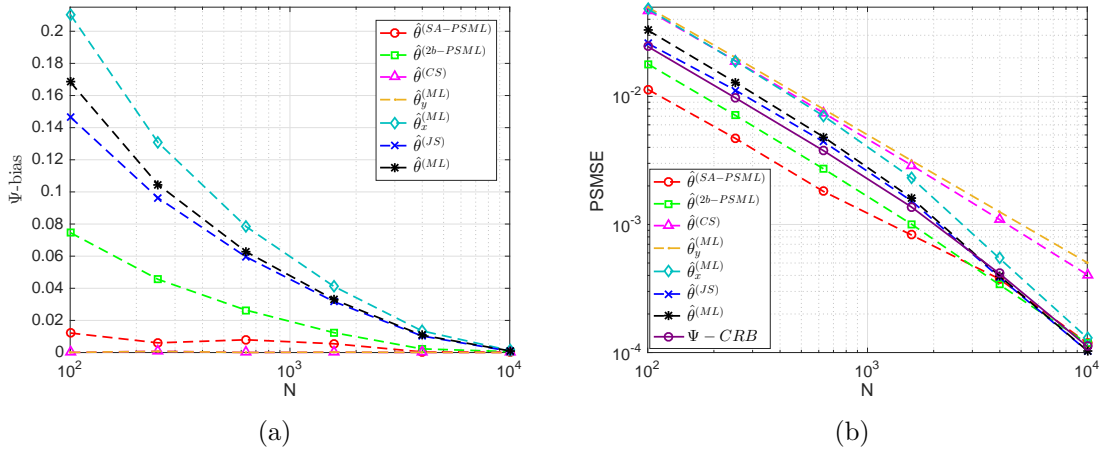


Figure 2.2: Linear Gaussian model: The Ψ -bias (a) and PSMSE (b) of the SA-PSML, the second-best PSML, CS, split-the-data, and the ML estimator versus the number of observations, N .

and $N = 10^4$, and $\theta_k = 1, \forall k = 1, \dots, M$. It can be seen that for the SA-PSML method the runtime increases with the problem dimensions. The second-best PSML has the lowest runtime, which is approximately constant with M and with N since it is based on the pairwise probability for every M . For larger observation number the SA-PSML requires on average fewer iterations to converge therefore the average runtime is smaller.

In Figs. 2.5(a)–2.5(c) the Ψ -bias and PSMSE and mean runtime of the SA-PSML versus the number of Monte-Carlo simulations, K , for various number of observations, N . Although Figs. 2.5(a) and 2.5(b) exemplify the fact that as K increases the SA-PSML is more accurate and the performance is better, Fig. 2.5(c) shows that the computational complexity increases correspondingly. It can be seen that the influence of N on the run-time is minor.

2.7.2 Bernoulli model

We consider a Bernoulli observation model. The observation of each population yields a binary value according to a Bernoulli distribution with unknown probability of success. At the first stage all M populations are observed to obtain N_x i.i.d. observations from every population. Based on these observations one population is selected according to a selection rule Ψ_x . Then, another N_y i.i.d observations are gathered from the selected population. The goal is to estimate the probability of success of the selected population. This model is useful in multi-armed bandit problems [130,131], where there are M arms, the m th arm yields a binary reward according to a Bernoulli distribution with unknown probability

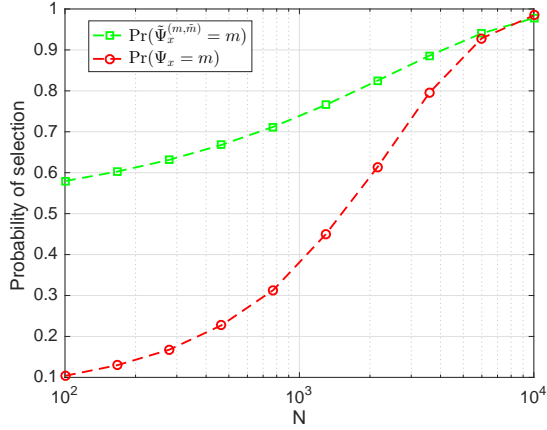


Figure 2.3: Linear Gaussian model: Comparison between the probability of selection and the pairwise probability of selection.

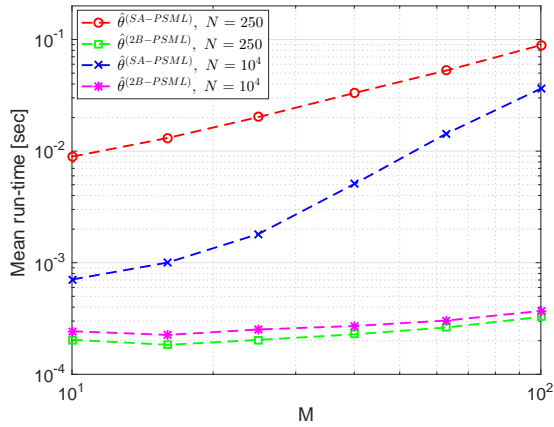


Figure 2.4: Linear Gaussian model: Run-time of the SA-PSML and the second-best PSML methods versus the number of parameters, M .

of success, θ_m . Another example arises in medical trials [52], where the response to the m th treatment is according to a Bernoulli distribution with unknown probability, θ_m , and we would select the treatment with the highest response rate. Therefore, for each n th sample $x_k[n] \sim \text{Ber}(\theta_k)$, $\forall k = 1, \dots, M$. We denote the first-stage observation vector as $\mathbf{x} = [x_1[1], \dots, x_M[1], x_1[2], \dots, x_M[2], \dots, x_M[N_{\mathbf{x}}]]^T$. In the following, we assume that the selection rule selects the arm with the highest averaged reward, which in this case is:

$$\Psi_{\mathbf{x}} = \arg \max_{k=1, \dots, M} \hat{\theta}_k^{(\text{ML})}(\mathbf{x}), \quad (2.87)$$

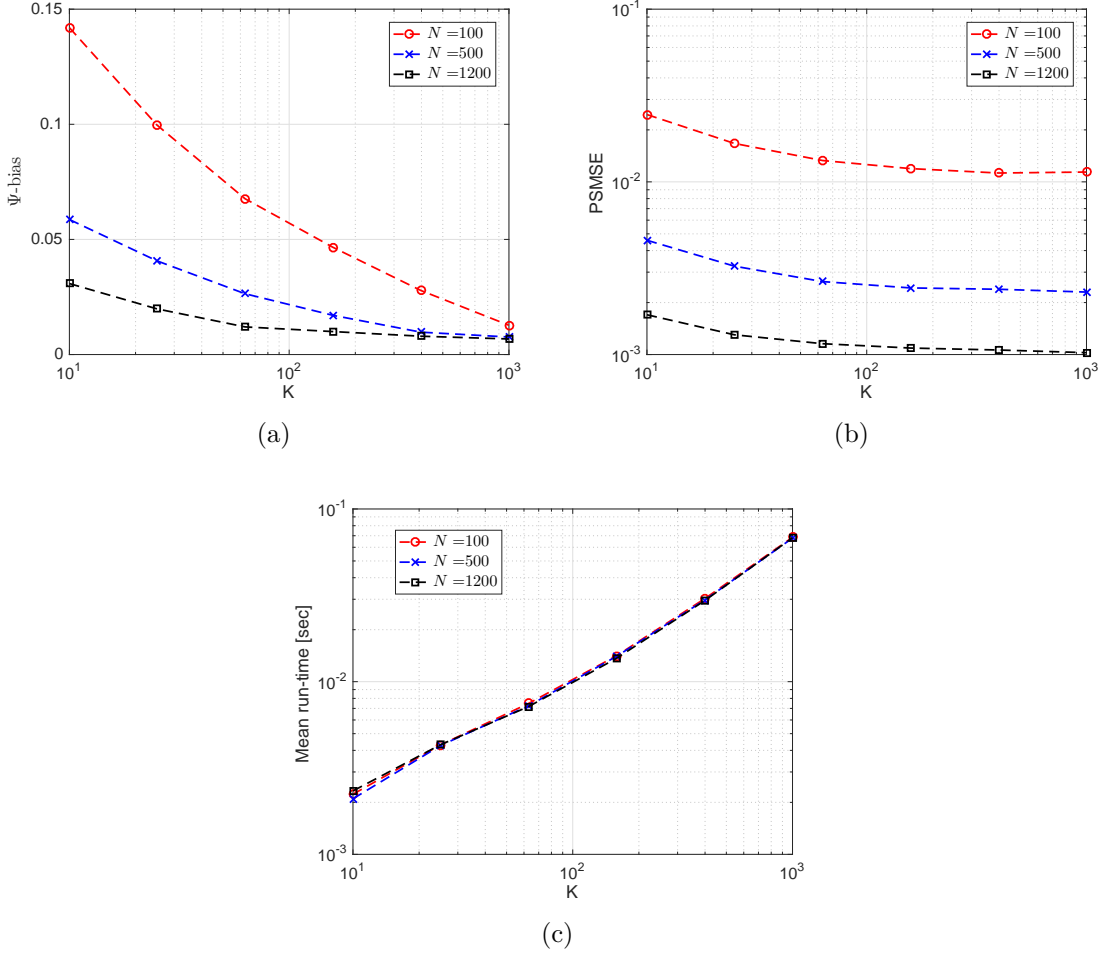


Figure 2.5: Linear Gaussian model: The Ψ -bias (a) and PSMSE (b) and mean run-time (c) of the SA-PSML estimator Vs. the number of Monte-Carlo simulations, K where the number of observation is $N = 100, 500, 1200$.

where the k th element of the single-stage ML estimator from (2.74) is given by

$$\hat{\theta}_k^{(\text{ML})}(\mathbf{x}) \triangleq \frac{1}{N_{\mathbf{x}}} \sum_{n=1}^{N_{\mathbf{x}}} x_k[n], \quad k = 1, \dots, M, \quad (2.88)$$

$k = 1, \dots, M$. In the second stage, only the selected population is sampled; therefore, the second-stage observation vector is $\mathbf{y} = [y_m[1], \dots, y_m[N_{\mathbf{y}}]]^T$, where $\mathbf{x} \in \mathcal{A}_m$, i.e. where the first-stage selection is m . Thus, the estimator from (2.73), which is the ML estimation of θ_m based on \mathbf{y} , is given by

$$\hat{\theta}_m^{(\text{ML})}(\mathbf{y}) \triangleq \frac{1}{N_{\mathbf{y}}} \sum_{n=1}^{N_{\mathbf{y}}} y_m[n], \quad (2.89)$$

and it is not defined for $k \neq m$. The k th element of the ML estimator based on both \mathbf{x} and \mathbf{y} is given by

$$\hat{\theta}_k^{(\text{ML})}(\mathbf{x}, \mathbf{y}) \triangleq \begin{cases} \frac{1}{N_{\mathbf{x}} + N_{\mathbf{y}}} (N_{\mathbf{x}} \hat{\theta}_m^{(\text{ML})}(\mathbf{x}) + N_{\mathbf{y}} \hat{\theta}_m^{(\text{ML})}(\mathbf{y})), & k = m \\ \hat{\theta}_k^{(\text{ML})}(\mathbf{x}), & k \neq m, \end{cases} \quad (2.90)$$

$k = 1, \dots, M$. In order to derive the second-best PSML, we examine the probability of pairwise selection from (2.29),

$$\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \boldsymbol{\theta}) = \Pr(\hat{\theta}_{\tilde{m}}^{(\text{ML})}(\mathbf{x}) \leq \hat{\theta}_m^{(\text{ML})}(\mathbf{x})). \quad (2.91)$$

The random variables $N_{\mathbf{x}} \hat{\theta}_m^{(\text{ML})}(\mathbf{x})$, $m = 1, \dots, M$ have a binomial distribution with $N_{\mathbf{x}}$ trials, and probability θ_m . We denote the binomial probability mass function with N trials and probability θ as:

$$F(n; N, \theta) \triangleq \binom{N}{n} \theta^n (1 - \theta)^{N-n}. \quad (2.92)$$

By using (2.91) and (2.92), the probability for pairwise selection is given by

$$\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \boldsymbol{\theta}) = \sum_{n=0}^{N_{\mathbf{x}}} \sum_{l=0}^n F(n; N_{\mathbf{x}}, \theta_m) F(l; N_{\mathbf{x}}, \theta_{\tilde{m}}). \quad (2.93)$$

One can notice that the derivative w.r.t θ of $F(n; \theta)$ is

$$\frac{\partial F(n; N, \theta)}{\partial \theta} = \xi(n, N, \theta) F(n; N, \theta), \quad (2.94)$$

where

$$\xi(n, N, \theta) \triangleq \frac{n - N\theta}{\theta(1 - \theta)}. \quad (2.95)$$

Therefore by substituting (2.91) in (2.32) we obtain that

$$\tilde{\mathbf{g}}(\boldsymbol{\theta}) = \sum_{n=0}^{N_{\mathbf{x}}} \sum_{l=0}^n \frac{F(n; N_{\mathbf{x}}, \theta_m) F(l; N_{\mathbf{x}}, \theta_k)}{\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, k)} = m; \boldsymbol{\theta})} (\xi(n, N_{\mathbf{x}}, \theta_m) \mathbf{e}_m + \xi(l, N_{\mathbf{x}}, \theta_k) \mathbf{e}_k). \quad (2.96)$$

The two-stage FIM from (2.67) for this scenario is a diagonal matrix, with the diagonal elements $[\mathbf{J}_{\mathbf{x}, \mathbf{y}}(\boldsymbol{\theta})]_{m, m} = \frac{N}{\theta_m(1 - \theta_m)} \forall m = 1, \dots, M$.

In Fig. 2.6 the Ψ -bias and PSMSE performance are presented versus the difference between θ_1 and the rest of parameters, Δ . In this case, $N = 150$, $N_{\mathbf{x}} = 0.75N$, $N_{\mathbf{y}} = 0.25N$, $M = 25$, and $\theta_1 = 0.5 + \Delta$, $\theta_k = 0.5$, $k = 2, \dots, M$. It can be seen that the proposed

PSML methods achieve better performance than the ML estimator in terms of both Ψ -bias and PSMSE. The split-the-data estimator, $\hat{\theta}^{(ML)}(\mathbf{y})$, is an Ψ -unbiased estimator, but its PSMSE performance is the highest since it uses only part of the observations.

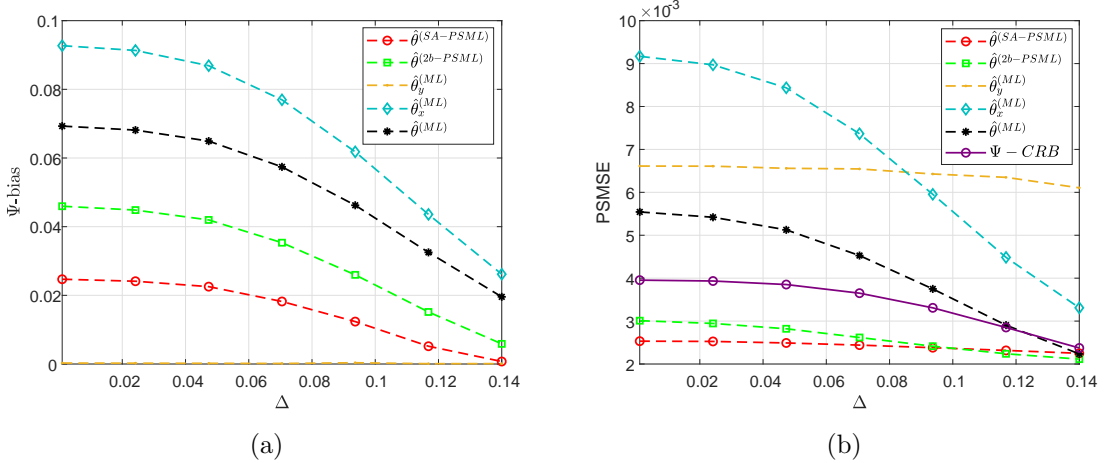


Figure 2.6: Bernoulli case: The Ψ -bias (a) and PSMSE (b) of the SA-PSML and second-best PSML versus Δ , the difference between θ_1 and the rest of the parameters, compared to split-the-data and the ML estimators.

2.7.3 Spectrum estimation after channel selection

In this subsection, we consider a problem of two-stage spectrum estimation after channel selection. We assume a multi-channel cognitive medium access control problem [42, 43], where a secondary user (SU) avoids channels that are occupied by a primary user (PU). The SU should not only detect a free channel, but also choose the optimal one [132]. Then, the second-stage observations set is acquired, and the goal is to estimate the parameter of the selected channel based on the two-stage observations. Unlike other studies on spectrum estimation, in which the main goal is optimal channel selection, here, we focus on the consequence estimation of the selected channel by taking into account the selection process.

We assume a frequency-flat and fast-fading channel; therefore, the discrete-time input-output relation of the k th channel $k = 1, \dots, M$ is given by

$$\begin{aligned} x_k[n] &= h_k s_k^{(1)}[n] + w_k^{(1)}[n], \quad n = 1, \dots, N_{\mathbf{x}} \\ y_k[n] &= h_k s_k^{(2)}[n] + w_k^{(2)}[n], \quad n = 1, \dots, N_{\mathbf{y}}, \end{aligned} \quad (2.97)$$

where $h_k \in \{0, 1\}$, $\forall k = 1, \dots, M$ are unknown deterministic parameters that represent the state of the channels. That is, $h_k = 1$ indicates that the k th channel is occupied

by a PU and $h_k = 0$ indicates that the k th channel is free for transmission. The state parameters, h_k , $k = 1, \dots, M$ are considered to be constant over the sensing period. The signals $s_k^{(i)}[n]$, $i = 1, 2$, and the additive noise, $w_k^{(i)}[n]$, $i = 1, 2$, are mutually independent i.i.d. Gaussian signals, with zero mean and unknown variances, $\sigma_{s_k}^2$ and $\sigma_{w_k}^2$, respectively. Therefore, $x_k[\cdot], y_k[\cdot] \sim \mathcal{N}(0, \sigma_k^2)$, $\forall k = 1, \dots, M$, where $\sigma_k^2 \triangleq h_m \sigma_{s_k}^2 + \sigma_{w_k}^2$ and $\boldsymbol{\theta} \triangleq [\sigma_1^2, \dots, \sigma_M^2]^T$ is the unknown parameter vector that characterizes the channels. We denote the first-stage observation vector as $\mathbf{x} = [x_1[1], x_2[1], \dots, x_M[1], \dots, x_M[N_{\mathbf{x}}]]^T$.

A widely applied spectrum sensing technique in CR is the minimum energy selection rule [43, 133, 134],

$$\Psi_{\mathbf{x}} = \arg \min_{k=1, \dots, M} \hat{\theta}_k^{(\text{ML})}(\mathbf{x}), \quad (2.98)$$

where k th element of the single-stage ML estimator from (2.74) is given by

$$\hat{\theta}_k^{(\text{ML})}(\mathbf{x}) \triangleq \frac{1}{N_{\mathbf{x}}} \sum_{n=1}^{N_{\mathbf{x}}} x_k^2[n], \quad k = 1, \dots, M, \quad (2.99)$$

$k = 1, \dots, M$. In this scenario we assume that in the second stage, only observations from the selected channel are taken; therefore, the second-stage observation vector is $\mathbf{y} = [y_m[1], \dots, y_m[N_{\mathbf{y}}]]^T$, where $\mathbf{x} \in \mathcal{A}_m$, i.e. where the first-stage selection is m . Thus, the estimator from (2.73), which is the ML estimation of θ_m based on \mathbf{y} , is given by

$$\hat{\theta}_m^{(\text{ML})}(\mathbf{y}) = \frac{1}{N_{\mathbf{y}}} \sum_{n=1}^{N_{\mathbf{y}}} y_m^2[n], \quad (2.100)$$

and k th element of the ML estimator based on both \mathbf{x} and \mathbf{y} from (2.12) is given by

$$\hat{\theta}_k^{(\text{ML})}(\mathbf{x}, \mathbf{y}) \triangleq \begin{cases} \frac{1}{N_{\mathbf{x}} + N_{\mathbf{y}}} (N_{\mathbf{x}} \hat{\theta}_m^{(\text{ML})}(\mathbf{x}) + N_{\mathbf{y}} \hat{\theta}_m^{(\text{ML})}(\mathbf{y})), & k = m \\ \hat{\theta}_k^{(\text{ML})}(\mathbf{x}), & k \neq m. \end{cases} \quad (2.101)$$

In order to derive the second-best PSML, we examine the probability of pairwise selection from (2.29),

$$\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m, \tilde{m})} = m; \boldsymbol{\theta}) = \Pr\left(\frac{\hat{\theta}_m^{(\text{ML})}(\mathbf{x})}{\hat{\theta}_{\tilde{m}}^{(\text{ML})}(\mathbf{x})} \leq 1\right), \quad (2.102)$$

where \tilde{m} is the index of the second-best parameter. The random variables $\frac{N_{\mathbf{x}}}{\sigma_m^2} \hat{\theta}_m^{(\text{ML})}(\mathbf{x})$, $m = 1, \dots, M$ have a central χ -square distribution with $N_{\mathbf{x}}$ degrees of freedom, and thus, $\frac{\sigma_k^2 \hat{\theta}_m^{(\text{ML})}(\mathbf{x})}{\sigma_{\tilde{m}}^2 \hat{\theta}_{\tilde{m}}^{(\text{ML})}(\mathbf{x})}$ have a F -central distribution [135, Ch. 2]. Therefore, by using (2.102), the

probability for pairwise selection of the first selection over the second is given by

$$\Pr(\tilde{\Psi}_{\mathbf{x}}^{(m,\tilde{m})} = m; \boldsymbol{\theta}) = F\left(\frac{\sigma_{\tilde{m}}^2}{\sigma_m^2}\right), \quad (2.103)$$

and the derivative of its log w.r.t. $\boldsymbol{\theta}$, from (2.32) is

$$\tilde{\mathbf{g}}(\boldsymbol{\theta}) = \frac{\varphi(\zeta)}{\theta_m F(\zeta)}(\mathbf{e}_{\tilde{m}} - \zeta \mathbf{e}_m), \quad (2.104)$$

where $F(\cdot)$ and $\varphi(\cdot)$ are the standard cdf and pdf of the F distribution, respectively, and $\zeta \triangleq \frac{\theta_{\tilde{m}}}{\theta_m}$. The two-stage FIM from (2.67) for this scenario is a diagonal matrix, where its diagonal elements are given by $[\mathbf{J}_{\mathbf{x},\mathbf{y}}(\boldsymbol{\theta})]_{m,m} = \frac{N}{2\theta_m^2} \forall m = 1, \dots, M$. Since $\hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{x}, \mathbf{y})$ is an efficient estimator, by substituting (2.104) in (2.34) the iteration of the second-best PSML using MBP-PSML is obtained by

$$\hat{\boldsymbol{\theta}}^{(i)}(\mathbf{x}, \mathbf{y}) = \hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{x}, \mathbf{y}) - \mathbf{J}_{\mathbf{x},\mathbf{y}}^{-1}(\hat{\boldsymbol{\theta}}^{(i-1)}(\mathbf{x}, \mathbf{y})) \frac{\varphi(\zeta^{(i-1)})}{\hat{\theta}_m^{(i-1)} F(\zeta^{(i-1)})}(\mathbf{e}_{\tilde{m}} - \zeta^{(i-1)} \mathbf{e}_m), \quad (2.105)$$

$\forall \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \Omega_{\mathbf{y}}$, where $\zeta^{(i)} \triangleq \frac{\hat{\theta}_{\tilde{m}}^{(i)}}{\hat{\theta}_m^{(i)}}$.

In Fig. 2.7 the Ψ -bias and PSMSE performance for the spectrum estimation after channel selection problem are presented versus the total number of observations N . In this case, $N_{\mathbf{x}} = 0.9N$, $N_{\mathbf{y}} = 0.1N$, $M = 30$, and $\theta_1 = 0.95$, $\theta_2 = 0.96$, $\theta_3 = 0.98$, $\theta_4 = \theta_5 = \theta_6 = 3$, $\theta_k = 1$, $k = 7, \dots, M$. It can be seen that the proposed PSML methods achieve better performance than the ML estimator in both terms, Ψ -bias and PSMSE. The split-the-data estimator, $\hat{\boldsymbol{\theta}}^{(\text{ML})}(\mathbf{y})$, is an Ψ -unbiased estimator, but its PSMSE performance is the highest since it uses only part of the observations.

2.7.4 Spectrum estimation with “black-box” selection rule

In this subsection, we demonstrate the robustness of the proposed SA-PSML method for a case where the selection rule is unknown to the estimator. We consider the CR spectrum estimation after channel selection from Subsection 2.7.3, where the selection is based on the k -Nearest Neighbors (kNN) algorithm [124, 125]. The Nearest Neighbor decision rule classifies a point as the classification of the nearest point in a set of classified points. The kNN decision rule is an extension, where the decision is based on the majority vote among the k nearest points. The kNN algorithm has been suggested in [136, 137] in the context of spectrum sensing in CR systems.

Let \mathcal{X} be a set of labeled points in \mathbb{R}^M , i.e. the “correct” selection for every point

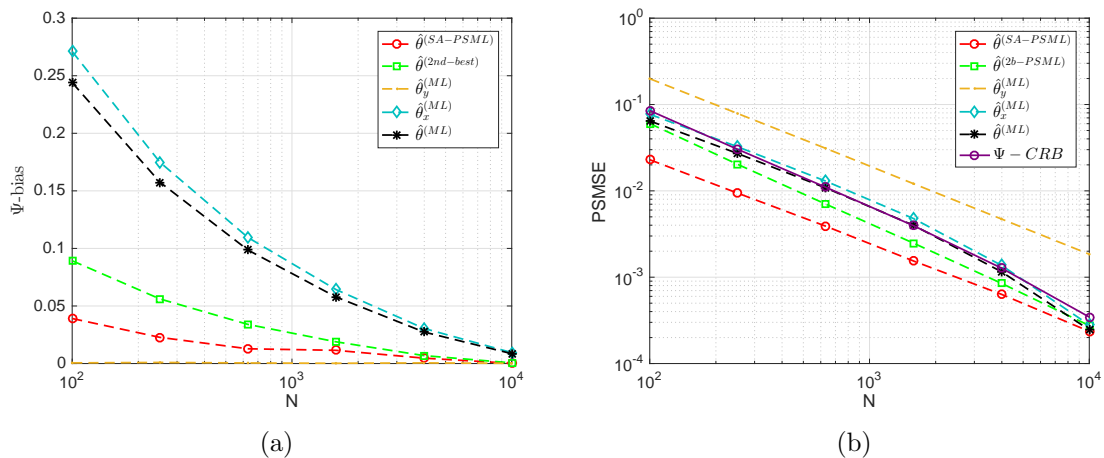


Figure 2.7: Spectrum estimation: The Ψ -bias of the SA-PSML and second-best PSML versus the number of observations N , compared to split-the-data and the ML estimators.

in \mathcal{X} is known. For the first stage observation set, \mathbf{x} , the kNN selection rule by the k nearest vectors in \mathcal{X} to $\hat{\theta}^{(ML)}(\mathbf{x})$, is defined in (2.99). In the following, we assume that the training data-set, \mathcal{X} , is inaccessible to the estimator directly; therefore, the kNN selection rule can be interpreted as a black-box procedure. However, we assume that we can generate multiple realizations from the observation model and determine the selection for each realization, to obtain the approximations from (2.43). In Fig. 2.8, the Ψ -bias and PSMSE of the proposed SA-PSML estimator and the ML estimator are shown versus the total number of observations, N , where $N_{\mathbf{x}} = 0.8N$, $N_{\mathbf{y}} = 0.2N$, $\theta_1 = 0.9, \theta_2 = \theta_3 = 0.95, \theta_k = 1, k = 4, \dots, M$ and $M = 25$. It can be seen that although the selection rule is unknown, the SA-PSML estimator have lower Ψ -bias and PSMSE than those of the ML estimator. The split-the-data estimator is Ψ -biased, but its PSMSE is the highest since it does not use all the observations.

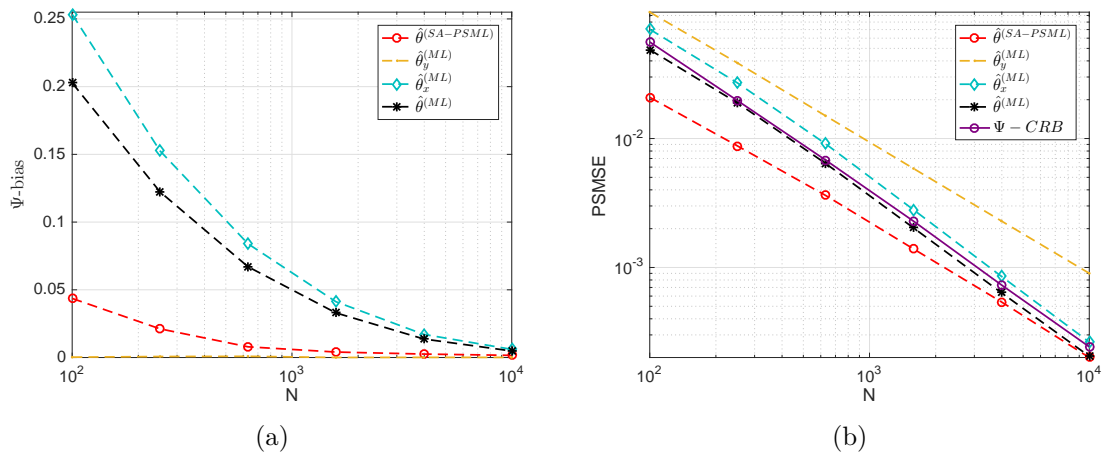


Figure 2.8: Spectrum estimation with a “black-box” kNN selection rule: The Ψ -bias (a) and PSMSE (b) of the SA-PSML, split-the-data, and ML estimators versus the number of observations, N .

Chapter 3

Post-Selection Estimation in Unidentifiable Models

In many modern applications the estimation of all the unknown parameters under a given model is time-consuming, not economical, or even impractical [138]. Therefore, a selection stage, that aims to identify the significant parameters subset, is usually conducted prior to the estimation. A common practice is to tackle these two tasks separately: First, a “parameters of interest” selection stage is conducted, and then, in the second stage, the selected unknown parameters are estimated by conventional estimation methods and the unselected parameters are usually set to zero. For example, in the context of estimating a sparse unknown parameter vector from noisy measurements, which has been extensively discussed in the literature [13, 67, 139–143]. In particular, in greedy compressive sensing algorithms, the support set is selected, and then the associated values on this support can be estimated by different estimation methods. However, these methods usually ignore the selection approach. The use of the same data for the selection step and for the estimation may introduce selection bias, changes the confidence intervals [55], and creates coupling between parameters that originally were decoupled [9]. Moreover, the subset of unknown parameters is itself stochastic, and this stochastic aspect is not accounted for by classical non-Bayesian estimation methods [144]. Therefore, new estimation methods that take into account the selection approach while maintaining reasonable complexity are required.

In this chapter we consider non-Bayesian estimation of deterministic parameters that have been selected according to a predetermined data-based selection rule. This work generalizes our results on identifiable single parameter estimation from Chapter 2 for the selection of a *subset* of parameters and for *unidentifiable* settings, where not all the parameters can be uniquely estimated. We discuss four post-selection estimators: the maximum-likelihood (ML) estimator, the coherent-ML estimator, which can be inter-

preted as the mismatched-ML estimator [84, 90, 145] under the considered model, the PSML estimator, and the coherent PSML estimator, and discuss their properties. Coherent estimators force the unselected parameters to zero and thus, can be implemented in multi-parameters, unidentifiable estimation problems with limited measurements. Then, we develop a practical algorithm, the stochastic approximation coherent PSML (SA-cPSML), for the implementation of the coherent PSML estimator, which accounts for the selection approach. Our simulations show that the SA-cPSML estimator outperforms the coherent-ML estimator for sparse vector recovery, where the selection is performed by the orthogonal matching pursuit (OMP) [141].

3.1 Model and Problem Formulation

Let $\mathbf{x} \in \Omega_{\mathbf{x}} \subseteq \mathbb{R}^N$ be an observation vector, where $\Omega_{\mathbf{x}}$, is the observation space with a probability density function $f(\mathbf{x}; \boldsymbol{\theta})$, which is parameterized by a deterministic parameter vector, $\boldsymbol{\theta} = [\theta_1, \dots, \theta_M]^T$. We assume that, prior to the estimation stage, a selection stage is conducted, in which the significant parameters that will be estimated are selected. In particular, we consider here the following common practical procedure. First, a support set of $\boldsymbol{\theta}$ is selected by a predetermined data-based selection-rule, where, $\hat{\Lambda} : \Omega_{\mathbf{x}} \rightarrow \{\Lambda_1, \dots, \Lambda_K\}$ is the selected support set, in which $\{\Lambda_k\}_{k=1}^K \in \mathcal{P}\{1, \dots, M\}$, are the candidate support sets. That is, if $m \in \hat{\Lambda}$, then θ_m is selected as a ‘‘parameter of interest’’, while if $m \notin \hat{\Lambda}$, then θ_m is considered a nuisance parameter. In the second stage, the parameters of the selected support set, $\boldsymbol{\theta}_{\hat{\Lambda}}$, are estimated based on the same observation vector, \mathbf{x} .

We assume here the common practice, which is to force the unselected parameters, $\boldsymbol{\theta}_{\hat{\Lambda}^c}$, to zero. In order to formalize this practice, we define ‘‘coherency’’ with the selection rule, similarly to the non-Bayesian and Bayesian coherency definitions for post model-selection estimation in [13, 81].

Definition 3.1. *An estimator $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \rightarrow \Omega_{\boldsymbol{\theta}}$ is coherent with a selection rule, $\hat{\Lambda}$, if*

$$\hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c} = \mathbf{0}. \quad (3.1)$$

The proposed post-selection architecture is presented schematically in Fig. 3.1. Additionally, by using Bayes rule it can be verified that

$$f(\mathbf{x} | \hat{\Lambda} = \Lambda_k; \boldsymbol{\theta}) = \frac{f(\mathbf{x}; \boldsymbol{\theta})}{\Pr(\hat{\Lambda} = \Lambda_k; \boldsymbol{\theta})}, \quad \forall \mathbf{x} \in \mathcal{A}_k, \quad (3.2)$$

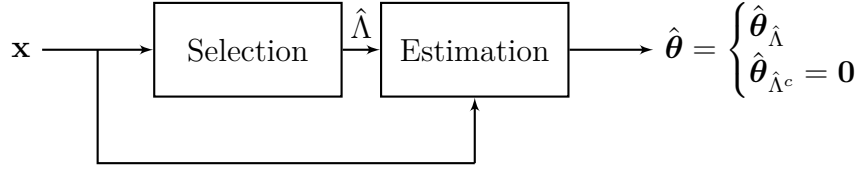


Figure 3.1: Post-selection estimation scheme: First, the set of significant parameters is selected based on the observation vector, \mathbf{x} , by a known predetermined selection rule, which results in a selected support set, $\hat{\Lambda}$. Second, the selected parameters, $\boldsymbol{\theta}_{\hat{\Lambda}}$, are estimated based on the same observation vector, while the unselected parameters, $\boldsymbol{\theta}_{\hat{\Lambda}^c}$, are estimated as zero.

where $\Pr(\hat{\Lambda} = \Lambda_k; \boldsymbol{\theta})$ is the probability that Λ_k is the selected support set $k = 1, \dots, K$. It is assumed that the deterministic sets $\mathcal{A}_k \triangleq \{\mathbf{x} \in \Omega_{\mathbf{x}} : \hat{\Lambda} = \Lambda_k\}$, $k = 1, \dots, K$ are a partition of $\Omega_{\mathbf{x}}$; thus, $\Pr(\hat{\Lambda} = \Lambda_k; \boldsymbol{\theta}) = \Pr(\mathbf{x} \in \mathcal{A}_k)$.

This model is a generalization of our previous model for estimation after parameter selection in [9, 65], as follows:

1. In [9, 65] the selection was of a *single* parameter of interest, i.e. it is assumed that $|\hat{\Lambda}| = 1$, while here the selection is of a subset of *multiple* parameters, i.e. $1 \leq |\hat{\Lambda}| \ll M$.
2. In [9, 65] the model is assumed to be identifiable, while here we consider also unidentifiable models [146, Def. 5.2]. In particular, in the considered model, $\boldsymbol{\theta}$ may be unidentifiable on the basis of \mathbf{x} , i.e. there may exist $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ such that $f(\mathbf{x}; \boldsymbol{\theta}_1) = f(\mathbf{x}; \boldsymbol{\theta}_2)$. As a result, the observation may not be sufficient for a consistent estimation of the full parameter vector, and the selection procedure is necessary. In particular, the considered model includes the special case of a linear model with more unknown parameters than observations, i.e. $M < N$, as described in Section 5.3. It should be noted, however, that we do assume that under the candidate support sets, $\{\Lambda_k\}_{k=1}^K$, the selected parameter vector, $\boldsymbol{\theta}_{\Lambda_k}$, is identifiable on the basis of \mathbf{x} under the assumption that the distribution is $f(\mathbf{x}; \boldsymbol{\theta}_{\Lambda_k})$.

It should be emphasized that in this work the support set selection process, $\hat{\Lambda}$, is taken for granted and we discuss the consequent estimation procedures. Accordingly, given a selection criterion, we ask, how can we improve the estimation performance? Therefore, we evaluate the estimation performance of the selected parameters, without including the estimation errors of the unselected parameters that are forced to zero by the selection stage. Hence, we use the PSSE cost function (2.4), which is given by

$$C^{(\hat{\Lambda})}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \triangleq \sum_{m=1}^M (\hat{\theta}_m - \theta_m)^2 \mathbb{1}_{\{m \in \hat{\Lambda}\}}. \quad (3.3)$$

Thus, for an estimator $\hat{\boldsymbol{\theta}}$ with a bounded second moment, the PSMSE (2.5), which is the expected cost function over (3.3) is:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} \left[C^{(\hat{\Lambda})}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \right] &= \sum_{m=1}^M \mathbb{E}_{\boldsymbol{\theta}} \left[(\hat{\theta}_m - \theta_m)^2 \mathbb{1}_{\{m \in \hat{\Lambda}\}} \right] \\ &= \sum_{k=1}^K \Pr(\hat{\Lambda} = \Lambda_k) \sum_{m \in \Lambda_k} \mathbb{E}_{\boldsymbol{\theta}} [(\hat{\theta}_m - \theta_m)^2 | \hat{\Lambda} = \Lambda_k], \end{aligned} \quad (3.4)$$

where the last equality is obtained by using the law of total expectation.

3.2 Post-parameter-selection Estimation Methods

In this section, we present different estimators of the parameter vector, $\boldsymbol{\theta}$, based on the observation vector, \mathbf{x} , for a given model-selection rule, $\hat{\Lambda}$, and discuss their properties.

3.2.1 ML Estimator

The commonly-used ML estimator [1] is defined as

$$\hat{\boldsymbol{\theta}}^{(\text{ML})} = \underset{\boldsymbol{\theta} \in \mathbb{R}^M}{\operatorname{argmax}} \log f(\mathbf{x}; \boldsymbol{\theta}). \quad (3.5)$$

The ML estimator is not a coherent estimator in the sense of Definition 3.1 since it does not take the selection procedure into consideration. Moreover, in a case of unidentifiable models, the ML estimator is not well defined. Therefore, it may not be a practical estimator for estimation after parameter selection.

3.2.2 Coherent ML Estimator

The coherent-ML estimator maximizes the likelihood function under the constraint of coherency, defined in (3.1). Thus, the coherent-ML estimator is the solution of

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{(\text{coherent-ML})} &= \underset{\boldsymbol{\theta} \in \mathbb{R}^M}{\operatorname{argmax}} \log f(\mathbf{x}; \boldsymbol{\theta}), \\ \text{s.t. } \boldsymbol{\theta}_{\Lambda_k^c} &= \mathbf{0}, \quad \forall \mathbf{x} \in \mathcal{A}_k. \end{aligned} \quad (3.6)$$

Since we assume that for any candidate support set, Λ_k , the associated parameter vector, $\boldsymbol{\theta}_{\Lambda_k}$, is identifiable on the basis of \mathbf{x} , then, the coherent ML estimator in (3.6) has a unique solution for any $k = 1, \dots, K$. In addition, the coherent ML estimator satisfies Definition 3.1, since from (3.6) it can be seen that $\hat{\boldsymbol{\theta}}_{\Lambda_k^c}^{(\text{coherent-ML})} = \mathbf{0}$. However, this estimator

does not exploit knowledge of the *selection mechanism*. The coherent ML estimator can be interpreted the mismatched-ML estimator [84, 90], which is derived under possibly misspecified model.

3.2.3 PSML Estimator

The PSML estimator, [9], maximizes the conditional likelihood function conditioned on the selection:

$$\hat{\boldsymbol{\theta}}^{(\text{PSML})} \triangleq \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{x} | \hat{\Lambda} = \Lambda_k; \boldsymbol{\theta}), \quad \forall \mathbf{x} \in \mathcal{A}_k. \quad (3.7)$$

The PSML estimator has displayed better performance than the ML estimator, in terms of post-selection-bias and PSMSE, in various scenarios when a *single* parameter is selected and with identifiable parameters [9, 65]. Moreover, it was shown in [9] that if an efficient estimator w.r.t. the PSMSE exists, then it is the PSML estimator. Nevertheless, similarly to the ML estimator, it is not a coherent estimator in the sense of Definition 3.1 and thus, it is not well-defined in a case of unidentifiable model. Therefore, it is also not a practical nor consistent estimator for estimation after the selection of parameter subset with possible unidentifiability.

3.2.4 Coherent PSML Estimator

Motivated by the theory behind the PSML estimator, in this chapter we propose the following coherent PSML estimator:

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{(\text{coherent-PSML})} &= \operatorname{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^M} \log f(\mathbf{x} | \hat{\Lambda} = \Lambda_k; \boldsymbol{\theta}), \\ \text{s.t. } \boldsymbol{\theta}_{\Lambda_k^c} &= \mathbf{0}, \quad , \forall \mathbf{x} \in \mathcal{A}_k. \end{aligned} \quad (3.8)$$

The coherent PSML estimator satisfies Definition 3.1, since from (3.8) it can be seen that $\hat{\boldsymbol{\theta}}_{\Lambda_k^c}^{(\text{coherent-PSML})} = \mathbf{0}$ and thus, it is a coherent estimator that can be implemented in practice. Substituting (3.2) in the log-likelihood function (3.8), results in

$$\log f(\mathbf{x} | \hat{\Lambda} = \Lambda_k; \boldsymbol{\theta}) = \log f(\mathbf{x}; \boldsymbol{\theta}) - \log \Pr(\hat{\Lambda} = \Lambda_k; \boldsymbol{\theta}). \quad (3.9)$$

Therefore, the PSML estimator can be interpreted as a penalized ML estimator, where the log-probability is the penalty function determined by the selection procedure. Under some regularity conditions, by substituting the constraint $\boldsymbol{\theta}_{\Lambda_k^c} = \mathbf{0}$ into the conditional log-likelihood function, the coherent PSML estimator from (3.8) can be obtained as the

solution to the following score equation:

$$\nabla_{\theta_{\Lambda_k}} \log f(\mathbf{x} | \hat{\Lambda} = \Lambda_k; \boldsymbol{\theta}) \Big|_{\theta_{\Lambda_k^c} = \mathbf{0}} = \mathbf{0}, \quad \forall \mathbf{x} \in \mathcal{A}_k. \quad (3.10)$$

By substituting (3.9) in (3.10), the coherent PSML estimator is obtained by the solution of

$$\nabla_{\theta_{\Lambda_k}} \log f(\mathbf{x}; \boldsymbol{\theta}) \Big|_{\theta_{\Lambda_k^c} = \mathbf{0}} = \mathbf{g}_k(\boldsymbol{\theta}) \Big|_{\theta_{\Lambda_k^c} = \mathbf{0}}, \quad (3.11)$$

where

$$\mathbf{g}_k(\boldsymbol{\theta}) \triangleq \nabla_{\theta_{\Lambda_k}} \log \Pr(\hat{\Lambda} = \Lambda_k; \boldsymbol{\theta}). \quad (3.12)$$

In general, an analytical solution of (3.11) is intractable, furthermore, the evaluation of $\mathbf{g}_k(\cdot)$ is impractical even at a given point. Therefore, numerical solutions should be considered in order to implement the proposed coherent PSML estimator.

3.3 Practical Implementation of the coherent PSML estimator

In this section, we present the implementation of the coherent PSML estimator from (3.8), by an iterative method. In the proposed SA-cPSML method, first, we set an initial coherent estimator, $\hat{\boldsymbol{\theta}}_{\Lambda_k}^{(0)}$. This initial estimator can be, for example, the coherent ML estimator from (3.6). Then, we iteratively refine the estimator of the selected parameters while keeping the unselected parameters as zeros. That is, the selection is not changed during the implementation of the SA-cPSML algorithm. Then, at each step we solve:

$$\nabla_{\theta_{\Lambda_k}} \log f(\mathbf{x}; \hat{\boldsymbol{\theta}}^{(i)}) = \mathbf{g}_k(\hat{\boldsymbol{\theta}}^{(i-1)}), \quad (3.13)$$

$\forall \mathbf{x} \in \mathcal{A}_k$. In practice, a closed form of (3.13) is not always available; hence, we suggest an iterative method inspired by the Fisher scoring method [1, Ch. 7.7]:

$$\hat{\boldsymbol{\theta}}_{\Lambda}^{(i)} = \hat{\boldsymbol{\theta}}_{\Lambda}^{(i-1)} - \mathbf{J}_{\Lambda}^{-1}(\hat{\boldsymbol{\theta}}^{(i-1)}) \mathbf{g}(\hat{\boldsymbol{\theta}}^{(i-1)}), \quad (3.14)$$

where the Fisher information matrix (FIM) is defined as

$$\mathbf{J}(\boldsymbol{\theta}) \triangleq E_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}; \boldsymbol{\theta})]. \quad (3.15)$$

It should be noted that since the true parameter may be unidentifiable, the FIM of the full parameter vector may be a singular matrix [147]. However, since we assume that $\boldsymbol{\theta}_{\Lambda_k}$ is identifiable, the submatrix of the FIM, $\mathbf{J}(\boldsymbol{\theta})$, under the k th model, $\mathbf{J}_{\Lambda_k}(\boldsymbol{\theta})$, is a non-singular matrix for any $k = 1, \dots, K$. Therefore, the iteration in (3.14) is well defined.

A major obstacle in the PSML methods is that the probability of selection and the gradient from (3.12) are intractable. To confront this problem, we suggest a stochastic-approximation (SA) approach [65], which numerically evaluates the gradient of the log-probability from (3.12) at each iteration. To this end, for any $\boldsymbol{\theta} \in \mathbb{R}^M$ we drew L i.i.d. vector samples, $\{\tilde{\mathbf{x}}^{(l)}\}_{l=1}^L$, from $f(\mathbf{x}; \boldsymbol{\theta})$. By using these sample, the stochastic approximation of $\mathbf{g}_k(\boldsymbol{\theta})$ from (3.12) is given by:

$$\hat{\mathbf{g}}_k(\boldsymbol{\theta}) \triangleq \frac{\sum_{l=1}^L \nabla_{\boldsymbol{\theta}_{\Lambda_k}} \log f(\tilde{\mathbf{x}}^{(l)}; \boldsymbol{\theta}) \mathbb{1}_{\{\hat{\Lambda}(\tilde{\mathbf{x}}^{(l)})=\Lambda_k\}}}{\sum_{l=1}^L \mathbb{1}_{\{\hat{\Lambda}(\tilde{\mathbf{x}}^{(l)})=\Lambda_k\}}}, \quad (3.16)$$

for any $k = 1, \dots, K$. By substituting (3.16) in (3.14) we obtain that the SA-cPSML iteration step is given by:

$$\hat{\boldsymbol{\theta}}_{\hat{\Lambda}}^{(i)} = \hat{\boldsymbol{\theta}}_{\hat{\Lambda}}^{(i-1)} - \mathbf{J}_{\hat{\Lambda}}^{-1}(\hat{\boldsymbol{\theta}}^{(i-1)}) \hat{\mathbf{g}}(\hat{\boldsymbol{\theta}}^{(i-1)}). \quad (3.17)$$

The SA-cPSML method is summarized in Algorithm 5.

Algorithm 5 : SA-cPSML

Require: observation vector, \mathbf{x} , selection approach, $\hat{\Lambda}$, convergence parameter, δ .

- 1: initialize: $i = 0$, $\hat{\boldsymbol{\theta}}^{(0)} = \hat{\boldsymbol{\theta}}^{(\text{coherent-ML})}$
- 2: **repeat**
- 3: generate sample vectors $\{\tilde{\mathbf{x}}^{(l)}\}_{l=1}^L \sim f(\mathbf{x}; \hat{\boldsymbol{\theta}}^{(i-1)})$
- 4: evaluate $\hat{\mathbf{g}}(\hat{\boldsymbol{\theta}}^{(i-1)})$ by (3.16)
- 5: update the next iteration, $\hat{\boldsymbol{\theta}}^{(i)}$, according to (3.17)
- 6: **until** $\|\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}}^{(i-1)}\| \leq \delta$

Ensure: SA-cPSML estimator, $\hat{\boldsymbol{\theta}}^{(\text{coherent-PSML})} = \hat{\boldsymbol{\theta}}^{(i)}$.

3.4 Simulations

We consider the following Gaussian linear model:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}, \quad (3.18)$$

where $\mathbf{x} \in \mathbb{R}^N$ is the observation vector, $\boldsymbol{\theta} \in \mathbb{R}^M$ is an unknown deterministic parameter vector, and $\mathbf{H} \in \mathbb{R}^{N \times M}$ is a known matrix. The noise vector, $\mathbf{w} \in \mathbb{R}^N$, is a white Gaussian

vector with a known covariance matrix, $\sigma^2\mathbf{I}$. We consider only candidate support sets that the submatrices of \mathbf{H} corresponding to these sets are of full (column) rank. Thus, Assumption 2 holds, even for the case of more unknown parameters than measurements, $N < M$.

Given the selected support set, $\hat{\Lambda}$, it can be shown that the coherent-ML estimator from (3.6) under the consider model is

$$\hat{\boldsymbol{\theta}}_{\hat{\Lambda}}^{(\text{coherent-ML})} \triangleq (\mathbf{H}_{\hat{\Lambda}}^T \mathbf{H}_{\hat{\Lambda}})^{-1} \mathbf{H}_{\hat{\Lambda}}^T \mathbf{x}, \quad (3.19)$$

where $\mathbf{H}_{\hat{\Lambda}}^T \mathbf{H}_{\hat{\Lambda}}$ is a non-singular matrix since we assume that Assumption 2 holds. Under the considered observations model from (3.18), the coherent PSML score equation from (3.11) is equivalent to

$$\frac{1}{\sigma^2} \mathbf{H}_{\Lambda_k}^T (\mathbf{x} - \mathbf{H}_{\Lambda_k} \boldsymbol{\theta}_{\Lambda_k}) = \mathbf{g}_k(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}_{\hat{\Lambda}} = \mathbf{0}}, \quad \forall \mathbf{x} \in \mathcal{A}_k. \quad (3.20)$$

However, evaluation of $\mathbf{g}_k(\boldsymbol{\theta})$ is usually intractable. Hence, we implement the proposed SA-cPSML from Algorithm 3, which is initialized by the coherent ML estimator from (3.19).

We consider the OMP algorithm [141] as the support-selection criterion, where the stopping rule is set to

$$\left\| \left(\mathbf{I} - \mathbf{H}_{\hat{\Lambda}^{(j)}} (\mathbf{H}_{\hat{\Lambda}^{(j)}}^T \mathbf{H}_{\hat{\Lambda}^{(j)}})^{-1} \mathbf{H}_{\hat{\Lambda}^{(j)}}^T \right) \mathbf{x} \right\| < \sigma \sqrt{N - j}, \quad (3.21)$$

in which j is the iteration index, $\hat{\Lambda}^{(j)}$ is the selected support set in the j th iteration, and $|\hat{\Lambda}^{(j)}| = j$. In addition, it is assumed that the maximum number of iterations is $\max_{k=1, \dots, K} |\Lambda_k|$, according to the candidate models.

In the following simulations, we set the matrix $\mathbf{H} = [\tilde{\mathbf{H}}, \mathbf{I}] \in \mathbb{R}^{32 \times 64}$, where $\tilde{\mathbf{H}}$ is a normalized Hadamard matrix, and then, we normalize the columns of the matrix \mathbf{H} . The deterministic vector $\boldsymbol{\theta} \in \mathbb{R}^{64}$ has been generated once by a zero-mean Gaussian distribution with a variance of 0.1, and then 3 elements are multiplied by 2. In Fig. 3.2 the PSMSE of the proposed SA-cPSML estimator and of the coherent-ML estimator, which is obtained as an output of the OMP algorithm, is presented versus $\frac{1}{\sigma^2}$. It can be seen that the proposed SA-cPSML estimator achieves lower PSMSE than the coherent-ML estimator for any value of σ^2 . It is worth mentioning that the PSMSE is not a monotonic function of $\frac{1}{\sigma^2}$, since for different noise levels the average number of parameters of interest varies, which affects the PSMSE in (3.4).

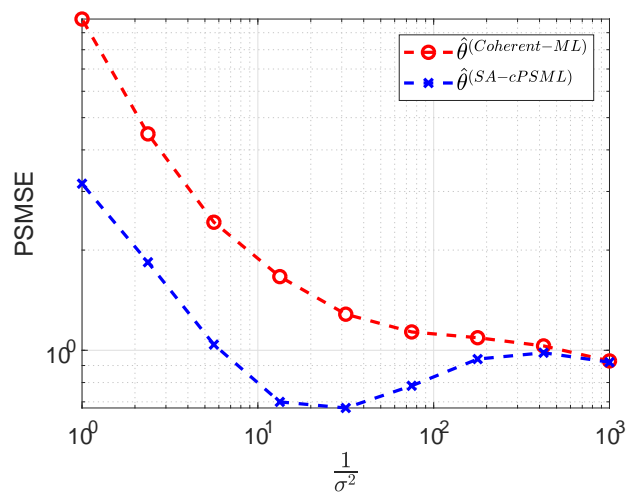


Figure 3.2: The PSMSE of the SA-cPSML compared to the coherent-ML

Chapter 4

Non-Bayesian Post-Model-Selection Estimation as Estimation Under Model Misspecification

Estimation under misspecified models has been discussed in the literature and garnered renewed attention recently. In particular, [82–84] discuss the asymptotic properties of the ML estimator under a misspecified model, also known as the misspecified ML (MML) estimator. For Bayesian parameter estimation, [85, 86] discuss the properties of Bayesian estimators under misspecified models. A Cramér-Rao-type lower bound that accounts for model misspecification, the misspecified CRB (MCRB), was developed in [87, 88] and was discussed in and derived for several scenarios in [89–92]. Modifications for the MCRB, such as a constrained MCRB for problems involving equality constraints [93] and a generalized MCRB for estimation problems in which the Hessian matrix is singular [94] were introduced. In [95], a cyclic MCRB, which is a lower bound on the mean cyclic error for periodic estimation problems under model misspecification, was derived. In [96, 97], the MCRB was used to design a model-selection procedure. In [98], a bilateral bound on the MSE under model mismatch that is applicable for Bayesian and non-Bayesian approaches was developed. Despite the elegant and useful theory presented in these works, none of the existing works deals with post-model-selection estimation the procedure that led to the selection of the misspecified model. In the conventional estimation under model misspecification, there is a clear definition of the assumed probability density function (pdf). However, in post-model-selection estimation, there are several candidate models. Therefore, the precise interpretation of the assumed model in this context should be meticulously considered. As a result, the existing misspecified bounds and estimators cannot be directly applied for the considered post-model-selection estimation.

4.1 Background: Estimation Under Model Misspecification

Estimation under model misspecification is a framework that relates to situations where the considered model, named the *assumed* model, may be different from the *true* model. Model misspecification can be caused by faults or model relaxations that aim to reduce the estimation complexity. The analysis of estimators under misspecified models should consider the statistics of both the true and the assumed models.

Let $(\Omega_{\mathbf{x}}, \mathcal{F}, P)$ denote a probability space, where Ω is the observation space, \mathcal{F} is the σ -algebra, and P is a probability measure on \mathcal{F} . We assume that the pdf w.r.t. P exists and is denoted by $p(\cdot)$. Let $\mathbf{x} \in \Omega_{\mathbf{x}}$ be a random observation vector, which is distributed according to the pdf, $p(\mathbf{x})$, which represents the *true* observation model. Under the misspecified model, it is assumed that \mathbf{x} is distributed according to the pdf $f(\mathbf{x}; \boldsymbol{\theta})$ that is parameterized by a deterministic parameter vector $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$. In the following, $f(\mathbf{x}; \boldsymbol{\theta})$ is considered to be the *assumed* pdf. For the sake of simplicity, in the following, it is assumed that $\Omega_{\boldsymbol{\theta}} \succ \mathbb{R}^M$, i.e. $\boldsymbol{\theta}$ is an M length real-valued vector. It is considered that the assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, is strictly positive and twice continuously differentiable w.r.t. $\boldsymbol{\theta}$ for any measurable $\mathbf{x} \in \Omega_{\mathbf{x}}$ and $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$. We denote a misspecified estimator based on the assumed model as $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \rightarrow \Omega_{\boldsymbol{\theta}}$. The MML estimator, also known as the quasi-ML estimator, is defined as the ML estimator of the unknown parameters under the assumed model [84, 88–90] as follows:

$$\hat{\boldsymbol{\theta}}_{\text{MML}} \triangleq \arg \max_{\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}} f(\mathbf{x}; \boldsymbol{\theta}). \quad (4.1)$$

In conventional non-Bayesian estimation problems, the definition of the estimation error is straightforward: the difference between the estimator and the value of the true parameter. However, the estimation error definition is not trivial in estimation under misspecification since the assumed pdf parameters, $\boldsymbol{\theta}$, do not necessarily appear in the true pdf, $p(\mathbf{x})$. Therefore, the *pseudo-true* parameter vector is defined (see e.g. [84, 87]).

Definition 4.1. (*Pseudo-true parameters*) For a true pdf, $p(\mathbf{x})$, and an assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, with an assumed parameter vector, $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$, the pseudo-true parameter vector is defined as

$$\boldsymbol{\vartheta} \triangleq \arg \min_{\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}} \mathcal{D}_{KL}(p(\mathbf{x}) || f(\mathbf{x}; \boldsymbol{\theta})), \quad (4.2)$$

where $\mathcal{D}_{KL}(\cdot || \cdot)$ denotes the Kullback-Leibler divergence (KLD) [148], which is given for

general pdfs $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ by

$$\mathcal{D}_{KL}(g_1(\mathbf{x})||g_2(\mathbf{x})) \triangleq E_{g_1} \left[\log \left(\frac{g_1(\mathbf{x})}{g_2(\mathbf{x})} \right) \right]. \quad (4.3)$$

Since $E_p[\log p(\mathbf{x})]$ is not a function of $\boldsymbol{\theta}$, (4.2) can be written as

$$\boldsymbol{\vartheta} = \arg \max_{\boldsymbol{\theta} \in \Omega} E_p[\log f(\mathbf{x}; \boldsymbol{\theta})]. \quad (4.4)$$

Under mild regularity conditions, the MML estimator is a consistent estimator of the pseudo-true parameter, $\boldsymbol{\vartheta}$ [82, 84]. By using Definition 4.1, we can define the following misspecified squared error (MSSE) as a cost function for evaluation of estimators under model misspecification:

$$\mathcal{C}_{\text{MSSE}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) \triangleq (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})^T. \quad (4.5)$$

The corresponding misspecified MSE (MSMSE) can be defined as follows:

$$\text{MSMSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) \triangleq E_p[\mathcal{C}_{\text{MSSE}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\vartheta})] = E_p[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})^T]. \quad (4.6)$$

In the following, we define the misspecified-unbiasedness (MS-unbiasedness), i.e. unbiasedness under the misspecified setting [87, 90], as follows:

Definition 4.2. (*MS-unbiasedness*) *Considering a misspecified estimation model, where $p(\mathbf{x})$ is the true pdf and $f(\mathbf{x}; \boldsymbol{\theta})$ is the assumed pdf, an estimator $\hat{\boldsymbol{\theta}}$ is MS-unbiased if*

$$E_p[\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}] = \mathbf{0}, \quad (4.7)$$

where $\boldsymbol{\vartheta}$ is the pseudo-true parameter vector defined in (4.2).

Proposition 4.1. *The MS-unbiasedness from Definition 4.2 coincides with the Lehmann unbiasedness from Definition 1.1 w.r.t. the MSSE from (4.5).*

The proof of Proposition 4.1 is presented in Appendix B.i.

4.1.0.1 Misspecified CRB

In this context, we consider the following regularity conditions:

RC.1. The maximum of $E_p[\log f(\mathbf{x}; \boldsymbol{\theta})]$ w.r.t. $\boldsymbol{\theta}$ is a unique interior point.

RC.2. The log-likelihood function, $\log f(\mathbf{x}; \boldsymbol{\theta})$, is a twice differentiable function, and the functions $\left| \frac{\partial \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i} \right|$ and $\left| \frac{\partial^2 \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i^2} \right|$, $i = 1, \dots, |\Omega_\theta|$, are dominated by a function $m(\mathbf{x})$, which is a square-integrable function w.r.t. $p(\mathbf{x})$.

RC.3. There is a neighborhood of the pseudo-true parameter vector, $\boldsymbol{\vartheta}$, such that the function $\frac{1}{f(\mathbf{x};\boldsymbol{\vartheta})} |\nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta})|$ evaluated at any $\boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\vartheta}$ is bounded (element-wise) by $m(\mathbf{x})$, which is a square-integrable function w.r.t. $p(\mathbf{x})$.

Theorem 4.1. *Let $f(\mathbf{x};\boldsymbol{\theta})$ be an assumed model that satisfies regularity conditions RC.1–RC.3 where the true model is $p(\mathbf{x})$. Let $\hat{\boldsymbol{\theta}}$ be an MS-unbiased estimator with a finite variance under the misspecified assumed model $f(\mathbf{x};\boldsymbol{\theta})$. The following MCRB is a lower bound on the MSMSE of $\hat{\boldsymbol{\theta}}$:*

$$\text{MSMSE}(\hat{\boldsymbol{\theta}}, \boldsymbol{\vartheta}) \succeq \mathbf{A}^{-1}(\boldsymbol{\vartheta})\mathbf{B}(\boldsymbol{\vartheta})\mathbf{A}^{-1}(\boldsymbol{\vartheta}), \quad (4.8)$$

where the MSMSE is defined in (4.6), the Matrix

$$\mathbf{A}(\boldsymbol{\theta}) \triangleq \mathbb{E}_p \left[\nabla_{\boldsymbol{\theta}}^2 \log f(\mathbf{x};\boldsymbol{\theta}) \right], \quad (4.9)$$

is assumed to be a non-singular matrix, and where

$$\mathbf{B}(\boldsymbol{\theta}) \triangleq \mathbb{E}_p \left[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x};\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x};\boldsymbol{\theta}) \right]. \quad (4.10)$$

The matrices $\mathbf{A}(\boldsymbol{\theta})$ and $\mathbf{B}(\boldsymbol{\theta})$ can be interpreted as the Hessian form and the outer-product form FIM, respectively.

The proof of Theorem 4.1 appears in [87, Th. 4.1]. Nevertheless, for the sake of clarity, we derive in Appendix B.ii a new version of this proof.

4.2 Model: Estimation After Model Selection

Let $\mathbf{x} \in \Omega_{\mathbf{x}}$ be an observation vector, where $\Omega_{\mathbf{x}}$ is the observation space, with the true pdf, $p(\mathbf{x})$, which can be a parametric model or a non-parametric model.

In practice, the exact true model is unknown, and we consider a set of candidate pdfs, $\{f_k(\mathbf{x};\boldsymbol{\theta}^{(k)})\}_{k=1}^K$. The k th pdf is parameterized by an unknown deterministic parameter vector, $\boldsymbol{\theta}^{(k)} \in \Omega_k$, where Ω_k denotes the parameter space of the k th pdf. The K different pdfs represent the K different models. We assume that any candidate pdf, $f_k(\mathbf{x};\boldsymbol{\theta}^{(k)})$, $k \in \{1, \dots, K\}$, is strictly positive for any $\mathbf{x} \in \Omega_{\mathbf{x}}$ and $\boldsymbol{\theta}^{(k)} \in \Omega_k$. However, the true pdf does not necessarily belong to the set of candidate pdfs. Finally, the vector $\boldsymbol{\theta} \triangleq [(\boldsymbol{\theta}^{(1)})^T, \dots, (\boldsymbol{\theta}^{(K)})^T]^T \in \Theta$ is the augmented vector that contains the parameters of all candidate models, where $\Theta \triangleq \Omega_1 \times \dots \times \Omega_K$.

The data is assumed to have complex values in general, i.e. $x_i \in \mathbb{C}$, as often encountered in signal and array processing [149], and the set of parameters $\boldsymbol{\theta}$ consists of members that may be complex, real, or a mixture of both.

Post-model-selection estimation arises in many signal processing problems and can be described as a two-stage approach: in the first stage, the model is selected from the candidate models based on the observations. In the second stage, the parameters of the selected model are estimated based on the same observations. The selection stage is conducted according to a predetermined data-based selection rule, $\Psi : \Omega_{\mathbf{x}} \rightarrow \{1, \dots, K\}$.

We denote the deterministic sets that are associated with the selection of each model by

$$\mathcal{A}_k \triangleq \{\mathbf{x} \in \Omega_{\mathbf{x}} : \Psi = k\}, \quad k = 1, \dots, K, \quad (4.11)$$

and assume that $\{\mathcal{A}_k\}_{k=1}^K$ creates a disjoint partition of $\Omega_{\mathbf{x}}$, i.e. $\mathcal{A}_k \cap \mathcal{A}_m = \emptyset$, $m \neq k$, and $\cup_{k=1}^K \mathcal{A}_k = \Omega_{\mathbf{x}}$.

The probability of selecting the k th model is denoted by

$$p_k \triangleq \int_{\mathcal{A}_k} p(\mathbf{x}) d\mathbf{x}, \quad k = 1, \dots, K. \quad (4.12)$$

In addition, we define

$$\pi_k(\boldsymbol{\theta}^{(k)}) \triangleq \int_{\mathcal{A}_k} f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) d\mathbf{x}, \quad k = 1, \dots, K. \quad (4.13)$$

It should be noted that since $p(\mathbf{x})$ is unknown, the probabilities p_k , $k = 1, \dots, K$ are unknown. It should be noted that the probabilities in (4.12), $\{p_k\}_{k=1}^K$, are computed using the same probability measure, and thus,

$$\sum_{k=1}^K p_k = \sum_{k=1}^K \int_{\mathcal{A}_k} p(\mathbf{x}) d\mathbf{x} = \int_{\Omega_{\mathbf{x}}} p(\mathbf{x}) d\mathbf{x} = 1. \quad (4.14)$$

In contrast, the probabilities $\{\pi_k(\boldsymbol{\theta}^{(k)})\}_{k=1}^K$ are computed by integration w.r.t. a different probability measure for each k . Thus, in the general case, the sum of the probabilities

$$\sum_{k=1}^K \pi_k(\boldsymbol{\theta}^{(k)}) = \sum_{k=1}^K \int_{\mathcal{A}_k} f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) d\mathbf{x} \neq 1. \quad (4.15)$$

In this chapter, it is assumed that the model-selection rule, Ψ , is predetermined, and the goal is to analyze the consequent estimation. The parameters that are estimated may be different under the different selected models. Thus, $\hat{\boldsymbol{\theta}}^{(k)}$ denotes an estimator of the parameter vector of the k th model, $\boldsymbol{\theta}^{(k)}$, $\forall k \in \{1, \dots, K\}$.

4.3 Post-Selection Estimation as Estimation Under Model Misspecification

The problem of post-model-selection estimation described in Section 4.2 has been widely discussed in the literature on selective inference (see e.g. in [13, 63, 70–75, 78, 80, 81]). In this chapter, we take the approach of treating it as estimation under model misspecification, which is described in Section 4.1. Treating and analyzing the post-model-selection estimation within the framework of estimation under model misspecification is challenging since there is no clear definition of the *assumed model* in this case. This ambiguity arises due to the existence of several candidate models, and since the selected model is data-dependent, i.e. it differs for different observation vectors \mathbf{x} . In Subsections 4.3.1–4.3.3, we describe three interpretations of post-model-selection estimation as an estimation under model misspecification by describing their associated *assumed pdf* under the misspecified model, denoted by $f_I(\mathbf{x}; \boldsymbol{\theta})$, $f_{II}(\mathbf{x}; \boldsymbol{\theta})$, and $f_{III}(\mathbf{x}; \boldsymbol{\theta})$.

While the first interpretation aligns with common practice in selective inference (see e.g. [71, 78, 139, 150–153]), the second and third interpretations, to the best of our knowledge, are introduced here for the first time. Our perspective emphasizes that the third interpretation is particularly suitable for the considered setting, demonstrating coherency with the selection rule and exhibiting superior practical performance in the tested simulations.

4.3.1 Naive Interpretation

A natural approach is to treat the selected model as the assumed one and disregard the selection procedure. In this approach, if the k th model has been selected in the first stage, the pdf associated with the selected model, $f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)})$, is considered to be the assumed pdf. Mathematically, this implies that the assumed pdf is

$$f_I(\mathbf{x}; \boldsymbol{\theta}) = f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}), \quad \forall \mathbf{x} \in \mathcal{A}_k, \quad k = 1, \dots, K. \quad (4.16)$$

By using the indicator function, (4.16) can be rewritten as

$$f_I(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^K f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k}, \quad \forall \mathbf{x} \in \Omega_{\mathbf{x}}. \quad (4.17)$$

However, the function $f_I(\mathbf{x}; \boldsymbol{\theta})$ is not a valid pdf in the general case, as it does not integrate to unity. This can be shown by integrating $f_I(\mathbf{x}; \boldsymbol{\theta})$ w.r.t. \mathbf{x} over the observation space, $\Omega_{\mathbf{x}}$:

$$\begin{aligned} \int_{\Omega_{\mathbf{x}}} f_I(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} &= \int_{\Omega_{\mathbf{x}}} \sum_{k=1}^K f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k} d\mathbf{x} \\ &= \sum_{k=1}^K \int_{\mathcal{A}_k} f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) d\mathbf{x} = \sum_{k=1}^K \pi_k(\boldsymbol{\theta}^{(k)}), \end{aligned} \quad (4.18)$$

where the first and second equalities are obtained by substituting (4.17) and by changing the order of summing and integration, respectively, and the last equality is obtained by substituting (4.13). As shown in (4.15), the r.h.s. of (4.18) is not necessarily 1.

Although $f_I(\mathbf{x}; \boldsymbol{\theta})$ is not a valid pdf, in practice, it can be used for the estimation approach, e.g. by using the ML estimator under the selected model [71, 78, 139, 150–153] (see more in Subsection 4.4.1). However, in terms of analysis, referring to $f_I(\mathbf{x}; \boldsymbol{\theta})$ as the assumed pdf that defines the misspecified model is inappropriate, and there is no guarantee that results from Section 4.1 hold in this case. In particular, the MCRB may not be a valid bound under this setting.

4.3.2 Normalized Interpretation

Since the naive interpretation in Subsection 4.3.1 led to a non-valid pdf, a natural remedy is to take the assumed pdf to be a normalized version of $f_I(\mathbf{x}; \boldsymbol{\theta})$ from (4.17) as follows:

$$f_{II}(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\alpha(\boldsymbol{\theta})} f_I(\mathbf{x}; \boldsymbol{\theta}). \quad (4.19)$$

The normalization factor in (4.19) is

$$\alpha(\boldsymbol{\theta}) \triangleq \sum_{k=1}^K \pi_k(\boldsymbol{\theta}^{(k)}), \quad (4.20)$$

where $\pi_k(\boldsymbol{\theta}^{(k)})$ is defined in (4.13). By substituting (4.17) in (4.19) we obtain that

$$f_{II}(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\alpha(\boldsymbol{\theta})} \sum_{k=1}^K f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k}, \quad \forall \mathbf{x} \in \Omega_{\mathbf{x}}. \quad (4.21)$$

By using (4.19) and the fact that $\alpha(\boldsymbol{\theta})$ is not a function of \mathbf{x} , it can be verified that

$$\int_{\Omega_{\mathbf{x}}} f_{II}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} = \frac{1}{\alpha(\boldsymbol{\theta})} \int_{\Omega_{\mathbf{x}}} f_I(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} = \frac{1}{\alpha(\boldsymbol{\theta})} \sum_{k=1}^K \pi_k(\boldsymbol{\theta}^{(k)}) = 1, \quad (4.22)$$

where the second equality is obtained by substituting (4.18), and the last equality is obtained by substituting the definition of $\alpha(\boldsymbol{\theta})$ from (4.20). Thus, $f_{II}(\mathbf{x}; \boldsymbol{\theta})$, which has non-negative values and is integrated to unity, is indeed a valid pdf.

It is important to note that the normalization factor, $\alpha(\boldsymbol{\theta})$, is a function of the parameters of all candidate models, $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(K)}$. As a result, for any realization of \mathbf{x} , $f_{II}(\mathbf{x}; \boldsymbol{\theta})$ from (4.19) is a function of all the unknown parameters of both the selected and the unselected models. This is as opposed to $f_I(\mathbf{x}; \boldsymbol{\theta})$ from (4.16), in which if $\mathbf{x} \in \mathcal{A}_k$, then $f_I(\mathbf{x}; \boldsymbol{\theta})$ only depends on the k th parameter vector, $\boldsymbol{\theta}^{(k)}$, associated with the selected model. Thus, the second interpretation creates a coupling between the unknown parameters, a theoretical issue that contradicts the intuitive coherence expected during the selection stage. Intuitively, if the k th model has been selected, we are usually only interested in estimating the parameters under the selected model, and the parameters of the unselected models are irrelevant. This intuition aligns with the concept of *coherent estimators* that ignores the parameters of the unselected models, or treats them as zero (see e.g. in our previous works in [13, 81]). Furthermore, from a practical point of view, $f_{II}(\mathbf{x}; \boldsymbol{\theta})$ complicates the estimation process since, in this case, we need to estimate all $\boldsymbol{\theta}$ for any \mathbf{x} . Thus, this interpretation may seem cumbersome for the considered setting of post-model-selection estimation.

4.3.3 Selective Inference Interpretation

In order to balance competing objectives (i.e. using a valid pdf that is also tractable and reasonable), an alternative approach is proposed here. Instead of normalizing $f_I(\mathbf{x}; \boldsymbol{\theta})$ from (4.16) using a single normalization factor, such as $\alpha(\boldsymbol{\theta})$ from (4.20), this approach uses a separated normalization for each selected model (each marginal pdf $f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)})$ on the r.h.s. of (4.17)). By doing so, the normalization process does not induce coupling between the parameters of the different models. Specifically, in this framework, the assumed pdf is defined as follows:

$$f_{III}(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^K \frac{c_k}{\pi_k(\boldsymbol{\theta}^{(k)})} f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k}, \quad (4.23)$$

where $\{c_k\}_{k=1}^K$ are any set of constants that satisfy:

- C.1. $c_k \geq 0$, $k = 1, \dots, K$;
- C.2. c_k is not a function of $\boldsymbol{\theta}$ and/or \mathbf{x} , $k = 1, \dots, K$;
- C.3. $\sum_{k=1}^K c_k = 1$.

By using the Bayes rule and (4.13), one can notice that

$$\frac{f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)})}{\pi_k(\boldsymbol{\theta}^{(k)})} = f_k(\mathbf{x}|\Psi = k; \boldsymbol{\theta}^{(k)}), \quad \forall \mathbf{x} \in \mathcal{A}_k, \quad (4.24)$$

which is the conditional assumed pdf according to the k th model, conditioned by the event of selecting the k th model, $\Psi = k$. The denominator in the l.h.s. on (4.24) is computed by (4.13), since $\pi_k(\boldsymbol{\theta}^{(k)})$ is the probability of $\Psi = k$ according to the k th assumed model. By substituting (4.24) in (4.23), we obtain that

$$f_{III}(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^K c_k f_k(\mathbf{x}|\Psi = k; \boldsymbol{\theta}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k}. \quad (4.25)$$

One can verify that $f_{III}(\mathbf{x}; \boldsymbol{\theta})$ is a valid pdf, since

$$\begin{aligned} \int_{\Omega_{\mathbf{x}}} f_{III}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} &= \int_{\Omega_{\mathbf{x}}} \sum_{k=1}^K c_k f_k(\mathbf{x}|\Psi = k; \boldsymbol{\theta}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k} d\mathbf{x} \\ &= \sum_{k=1}^K \int_{\mathcal{A}_k} c_k f_k(\mathbf{x}|\Psi = k; \boldsymbol{\theta}^{(k)}) d\mathbf{x} = \sum_{k=1}^K c_k = 1, \end{aligned} \quad (4.26)$$

where the first equality is obtained by substituting (4.25), the second equality is obtained by changing the order of summing and integration, and the third equality is obtained by using the conditional pdf property $\int_{\mathcal{A}_k} f_k(\mathbf{x}|\Psi = k; \boldsymbol{\theta}^{(k)}) d\mathbf{x} = 1$, and the fact that c_k are not functions of \mathbf{x} (Condition C.2)). The last equality is obtained by substituting Condition C.3. Therefore, (4.26) implies that $f_{III}(\mathbf{x}; \boldsymbol{\theta})$ is a valid pdf as long as it is a nonnegative function, which is obtained from Condition C.1.

In general, one can choose any coefficients $\{c_k\}_{k=1}^K$ that satisfy Conditions C.1–C.3. In particular, we suggest to choose

$$c_k = p_k, \quad \forall k \in \{1, \dots, K\}, \quad (4.27)$$

where p_k is the true probability of selection from (4.12). It can be verified that this choice satisfies Conditions C.1–C.3. By substituting (4.27) in (4.23), we obtain

$$f_{III}(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^K \frac{p_k}{\pi_k(\boldsymbol{\theta}^{(k)})} f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k}. \quad (4.28)$$

Thus, for this choice the pdf $f_{III}(\mathbf{x}; \boldsymbol{\theta})$ is a weighted average of the candidate pdfs, where the weight of the marginal pdf of the k th model, $f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)})$, is the likelihood ratio of the

correct and the candidate models, $\frac{p_k}{\pi_k(\boldsymbol{\theta}^{(k)})}$. While the probabilities $\{p_k\}_{k=1}^K$ are unknown, $f_{III}(\mathbf{x}; \boldsymbol{\theta})$ from (4.28) lead to a valid estimator and bound, as shown in Subsections 4.4.3 and 4.8.3.

It can be seen that for a given $\mathbf{x} \in \mathcal{A}_k$, similar to $f_I(\mathbf{x}; \boldsymbol{\theta})$ in (4.17), the assumed pdf in (4.23), $f_{III}(\mathbf{x}; \boldsymbol{\theta})$, is only a function of the parameter vector, $\boldsymbol{\theta}^{(k)}$, associated with the selected model. Thus, the assumed pdfs $f_I(\mathbf{x}; \boldsymbol{\theta})$ and $f_{III}(\mathbf{x}; \boldsymbol{\theta})$ are *coherent* with the selection approach. Coherency in estimation after model selection refers to the property that the estimation approach accounts for the model selection (see more in [13, 66, 81]). In our case, this coherency means that if it is assumed that the observations obey the k th model, then the assumed pdf is only determined by $\boldsymbol{\theta}^{(k)}$. This is as opposed to $f_{II}(\mathbf{x}; \boldsymbol{\theta})$ in (4.19), which is a function of the augmented vector, $\boldsymbol{\theta}$. In addition, in previous works on selective inference it was shown that the conditional pdf of the observation conditioned on the selection is the appropriate pdf for post-selection estimation and analysis (see e.g. [9, 74–77]). Thus, $f_{III}(\mathbf{x}; \boldsymbol{\theta})$ corresponds with the existing theory.

Remark 4.1. *To conclude this section, we note that for the special case where there is a single candidate model, i.e. $K = 1$ and $\mathcal{A}_1 = \Omega_{\mathbf{x}}$, all the interpretations above coincide: $f_I(\mathbf{x}; \boldsymbol{\theta}) = f_{II}(\mathbf{x}; \boldsymbol{\theta}) = f_{III}(\mathbf{x}; \boldsymbol{\theta})$, since (4.13) and (4.20) imply that in this case $\alpha(\boldsymbol{\theta}) = \pi_1(\boldsymbol{\theta}) = 1$. Moreover, in this case, the considered scheme is reduced to conventional parameter estimation under model misspecification [84, 87–89], where the single candidate model is represented by the pdf $f_1(\mathbf{x}; \boldsymbol{\theta}^{(1)})$. If, in addition, this only candidate is the true model, i.e. if $f_1(\mathbf{x}; \boldsymbol{\theta}^{(1)}) = p(\mathbf{x})$, the considered scheme is reduced to the conventional non-Bayesian parameter estimation problem.*

4.4 Post-Model-Selection Estimators

In this section, we introduce post-model-selection estimators. We start by presenting the oracle ML estimator as a benchmark. Then, in Subsections 4.4.1–4.4.3, we present the MML estimators as post-model-selection estimators according to the interpretations presented above in Section 4.3.

4.4.1 Maximum Selected Likelihood (MSL) Estimator

An intuitive post-model-selection estimator is the MSL estimator, which is obtained by setting the estimator to be the ML estimator of the selected model. Therefore, the MSL estimator is the MML estimator under the naive interpretation presented in Sub-

section 4.3.1, i.e. it is obtained by substituting (4.16) in (4.1), as follows:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{MSL}}^{(k)} &\triangleq \arg \max_{\boldsymbol{\theta}^{(k)} \in \Omega_k} f_I(\mathbf{x}; \boldsymbol{\theta}) \\ &= \arg \max_{\boldsymbol{\theta}^{(k)} \in \Omega_k} f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}), \quad \forall \mathbf{x} \in \mathcal{A}_k.\end{aligned}\quad (4.29)$$

Although $f_I(\mathbf{x}; \boldsymbol{\theta})$ is not a valid likelihood function (as explained in Subsection 4.3.1), the MSL estimator is well defined and widely used in practice [71, 78, 139, 150–153].

4.4.2 Maximum Selected Normalized Likelihood (MSNL) Estimator

The MSNL estimator is the MML estimator under the normalized interpretation and is obtained by substituting the assumed likelihood from (4.21) in (4.1), as follows:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{MSNL}} &\triangleq \arg \max_{\boldsymbol{\theta} \in \Theta} f_{II}(\mathbf{x}; \boldsymbol{\theta}) \\ &= \arg \max_{\boldsymbol{\theta} \in \Theta} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) - \log \alpha(\boldsymbol{\theta}), \quad \forall \mathbf{x} \in \mathcal{A}_k.\end{aligned}\quad (4.30)$$

The last equality is obtained by using the fact that the log function is a monotonically increasing function, the properties of the indicator function, and the fact that $\{\mathcal{A}_k\}$ creates a partition of $\Omega_{\mathbf{x}}$. The maximization in (4.30) is w.r.t. $\boldsymbol{\theta}$, which also includes the parameters of the unselected models, $\boldsymbol{\theta}^{(l)}$, $l \neq k$, $\forall \mathbf{x} \in \mathcal{A}_k$, via $\alpha(\boldsymbol{\theta}) = \sum_{l=1}^K \pi_l(\boldsymbol{\theta}^{(l)})$ from (4.20). Since $\alpha(\boldsymbol{\theta})$ is not a function of the observation vector, \mathbf{x} , and is determined by the selection rule, Ψ , by using the fact that the log function is a monotonically increasing function, the properties of the indicator function, and the fact that $\{\mathcal{A}_k\}_{k=1}^K$ creates a partition of $\Omega_{\mathbf{x}}$, the estimator of the parameters of the selected model, $\boldsymbol{\theta}^{(k)}$, can be written as

$$\hat{\boldsymbol{\theta}}_{\text{MSNL}}^{(k)} = \arg \max_{\boldsymbol{\theta}^{(k)} \in \Omega_k} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) - \log \left(\pi_k(\boldsymbol{\theta}^{(k)}) + \sum_{l \neq k} \pi_l \right), \quad \forall \mathbf{x} \in \mathcal{A}_k, \quad (4.31)$$

where

$$\underline{\pi}_k \triangleq \min_{\boldsymbol{\theta}^{(k)} \in \Omega_k} \pi_k(\boldsymbol{\theta}^{(k)}), \quad k \in \{1, \dots, K\}. \quad (4.32)$$

The estimator in (4.31) can be interpreted as a penalized ML estimator, where the penalty is determined by the selection rule and the probabilities under the other candidate models. In practice, the minimization in (4.32) can be performed in advance (offline) as it is not a function of the observation vector, \mathbf{x} . Then, only the MNSL estimator of the selected

parameters in (4.31) is estimated based on the data in \mathbf{x} . The minimization in (4.32) may become impractical for a large number of candidate models, especially when some have a small probability of being selected. In addition, this may adversely impact performance by introducing additional uncertainty into the estimation approach.

4.4.3 Post-Selection ML (PSML) Estimator

The PSML estimator is the MML estimator under the selective inference interpretation from Subsection 4.3.3, and is given by

$$\hat{\boldsymbol{\theta}}_{\text{PSML}}^{(k)} \triangleq \arg \max_{\boldsymbol{\theta}^{(k)} \in \Omega_k} f_{III}(\mathbf{x}; \boldsymbol{\theta}^{(k)}), \quad \forall \mathbf{x} \in \mathcal{A}_k. \quad (4.33)$$

By substituting (4.25) in (4.33), and since $\{c_k\}_{k=1}^K$ are positive constants (see Condition C.1), independent of $\boldsymbol{\theta}$ (see Condition C.3), the maximization in (4.33) is equivalent in this case to

$$\hat{\boldsymbol{\theta}}_{\text{PSML}}^{(k)} = \arg \max_{\boldsymbol{\theta}^{(k)} \in \Omega_k} f_k(\mathbf{x} | \Psi = k; \boldsymbol{\theta}^{(k)}), \quad \forall \mathbf{x} \in \mathcal{A}_k. \quad (4.34)$$

Since the log function is a monotonically increasing function, by using the Bayes rule in (4.24), (4.34) can be written as

$$\hat{\boldsymbol{\theta}}_{\text{PSML}}^{(k)} = \arg \max_{\boldsymbol{\theta}^{(k)} \in \Omega_k} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) - \log \pi_k(\boldsymbol{\theta}^{(k)}), \quad \forall \mathbf{x} \in \mathcal{A}_k. \quad (4.35)$$

It can be seen that the PSML estimator is independent of the constants $\{c_k\}_{k=1}^K$. Hence, similarly to the MSL estimator, for any $\mathbf{x} \in \mathcal{A}_k$, $f_{III}(\mathbf{x}; \boldsymbol{\theta})$ is only a function of the parameters of the k th model, $\boldsymbol{\theta}^{(k)}$, in contrast with the MSNL estimator. The PSML estimator can be interpreted as a penalized ML estimator [12, 106], where the penalty term is $-\log \pi_k(\boldsymbol{\theta}^{(k)})$. Since $\pi_k(\boldsymbol{\theta}^{(k)})$ represents the probability of selection under the k th model, this penalty term can be interpreted as a measure of the uncertainty of the selection approach w.r.t. the selected model. Since the penalty term is not a pdf w.r.t. $\boldsymbol{\theta}^{(k)}$, the PSML estimator does not have a Bayesian interpretation.

Estimators that are based on maximizing conditional likelihood functions were shown to be more appropriate for post-selection estimation problems [9, 77]. Thus, the PSML estimator, which is the MML estimator under the selective inference interpretation, is expected to yield good MSE performance, as demonstrated numerically in Section 5.3.

The following remarks describe some relations between the estimators and special cases where these estimators coincide.

Remark 4.2. *In Remark 4.1 it is shown that for the special case of a single candidate*

model, $K = 1$, the three interpretations coincide, i.e. $f_I(\mathbf{x}; \boldsymbol{\theta}) = f_{II}(\mathbf{x}; \boldsymbol{\theta}) = f_{III}(\mathbf{x}; \boldsymbol{\theta})$. Thus, in this case, the MSL, MSNL, and PSML estimators from (4.29), (4.30) and (4.33), respectively, are all reduced to the MML estimator in (4.1), associated with conventional estimation under misspecification.

Remark 4.3. In the case where $\pi_k(\boldsymbol{\theta}^{(k)})$ is not a function of $\boldsymbol{\theta}^{(k)}$ for a given k , it can be seen that the maximizations in (4.31) and (4.35) are equivalent to the maximization in (4.29) for $\mathbf{x} \in \mathcal{A}_k$. Therefore, in this case, the MSL, MSNL, and PSML estimators coincide for the selection of the k th model.

Remark 4.4. From the definition in (4.13), the probabilities satisfy $\pi_k(\boldsymbol{\theta}^{(k)}) \in [0, 1]$. Therefore, in the case where for some k , $\pi_l(\boldsymbol{\theta}^{(l)})$ achieves its minimum at zero $\forall l \neq k$, i.e. $\underline{\pi}_l = 0 \forall l \neq k$, then the MSNL estimator from (4.31) coincides with the PSML estimator in (4.35) under the selection of the k th model. This can be verified by substitution of (4.32) with $\underline{\pi}_l = 0 \forall l \neq k$ in the MSNL estimator in (4.31), which results in $\hat{\boldsymbol{\theta}}_{MSNL}^{(k)} = \hat{\boldsymbol{\theta}}_{PSML}^{(k)}$.

4.5 Post-Model-Selection Performance Analysis

In the context of post-model-selection estimation as presented in Section 4.2, if the k th model has been selected, we focus on estimating the parameter vector $\boldsymbol{\theta}^{(k)}$ of the selected k th model, where the parameters from the unselected models can be interpreted as nuisance parameters. Hence, in order to analyze post-model-selection estimators that may estimate different quantities for different observation vectors, we introduce the k th-MSE, the associated post-model-selection (PS)-pseudo-true parameter, and the post-model-selection unbiasedness in Subsection 4.5.1. Subsequently, we derive the PS-pseudo-true parameter vectors for the different interpretations in Section 4.6.

4.5.1 Post-Model-Selection MSE

The k th-MSE for estimating $\boldsymbol{\theta}^{(k)}$ under the selection of the k th model is the following $|\Omega_k| \times |\Omega_k|$ matrix:

$$\mathbf{MSE}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\theta}^{(k)}) \triangleq \mathbb{E}_p[(\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)})(\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)})^H | \Psi = k]. \quad (4.36)$$

Thus, for each k , the dimensions of the k th-MSE may be different.

The MCRB in Theorem 4.1 is evaluated at the pseudo-true parameter. Thus, in order to discuss the post-model-selection performance, it is crucial to define the relevant pseudo-

true parameter vectors. The following definition generalizes the pseudo-true parameter vector definition in Definition 4.1 to the misspecified case.

Definition 4.3. (*PS-pseudo-true parameter vector*) Let $f(\mathbf{x}; \boldsymbol{\theta})$ be a general assumed pdf, which is strictly positive for any $\mathbf{x} \in \Omega_{\mathbf{x}}$, in the post-model-selection estimation setting. The PS-pseudo-true parameter vector w.r.t. $f(\mathbf{x}; \boldsymbol{\theta})$ given that the k th model is selected, $\boldsymbol{\vartheta} \triangleq [(\boldsymbol{\vartheta}^{(1)})^T, \dots, (\boldsymbol{\vartheta}^{(K)})^T]^T$, is defined as follows:

$$\boldsymbol{\vartheta} \triangleq \arg \min_{\boldsymbol{\theta} \in \Theta} \mathcal{D}_{KL}(p(\mathbf{x}|\Psi = k)||f(\mathbf{x}; \boldsymbol{\theta})), \quad (4.37)$$

where, according to the Bayes rule and similar to (4.24), the conditional true pdf is given by

$$p(\mathbf{x}|\Psi = k) = \frac{p(\mathbf{x})}{p_k}, \quad \forall \mathbf{x} \in \mathcal{A}_k. \quad (4.38)$$

In other words, the k th PS-pseudo-true parameter vector is obtained by minimization of the KLD between the true pdf conditioned by the selection $\Psi = k$, $p(\mathbf{x}|\Psi = k)$, and the assumed pdf. According to this definition, $\boldsymbol{\vartheta}^{(k)}$ is the point that is *the closest* to the ground truth, given the k th selection. It should be noted that the support of $p(\mathbf{x}|\Psi = k)$ is \mathcal{A}_k , which is included in the support of $f(\mathbf{x}; \boldsymbol{\theta})$, $\Omega_{\mathbf{x}}$, and thus, the KLD on the r.h.s. of (4.37) is well defined. In addition,

$$\mathcal{D}_{KL}(p(\mathbf{x}|\Psi = k)||f(\mathbf{x}; \boldsymbol{\theta})) = \mathbb{E}_p[\log p(\mathbf{x}|\Psi = k)|\Psi = k] - \mathbb{E}_p[\log f(\mathbf{x}; \boldsymbol{\theta})|\Psi = k]. \quad (4.39)$$

This definition is equivalent to Definition 4.1, but with conditional true and assumed pdfs, which fits the conditional MSE in (4.36), i.e. conditioned by the event of $\Psi = k$. As in Definition 4.1, since $p(\mathbf{x}|\Psi = k)$ is not a function of $\boldsymbol{\theta}$, the minimization in (4.37) is reduced to

$$\boldsymbol{\vartheta} = \arg \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_p[\log f(\mathbf{x}; \boldsymbol{\theta})|\Psi = k]. \quad (4.40)$$

In a similar manner to the conventional estimation under model misspecification, by using Definition 4.3 we can define the k th-MSE of an estimator under the assumed pdf $f(\mathbf{x}; \boldsymbol{\theta})$ as the k th-MSE from (4.36) evaluated at the PS-pseudo-true parameter vector, $\boldsymbol{\vartheta}^{(k)}$, from (4.40) as follows:

$$\mathbf{MSE}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\vartheta}^{(k)}) \triangleq \mathbb{E}_p \left[(\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\vartheta}^{(k)})(\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\vartheta}^{(k)})^H | \Psi = k \right]. \quad (4.41)$$

Based on the conditional MSE in (4.41), we can define the associated post-model-selection MS-unbiasedness (PSMS-unbiasedness) definition, which generalizes the MS-

unbiasedness from Definition 4.2.

Definition 4.4. (*PSMS-unbiasedness*) An estimator of the k th parameter vector, $\hat{\boldsymbol{\theta}}^{(k)}$, is a PSMS-unbiased estimator w.r.t. a generic assumed pdf $f(\mathbf{x}; \boldsymbol{\theta})$ if

$$\mathbb{E}_p[\hat{\boldsymbol{\theta}}^{(k)} | \Psi = k] = \boldsymbol{\vartheta}^{(k)}, \quad (4.42)$$

where $\boldsymbol{\vartheta}^{(k)}$ is defined in (4.37).

This unbiasedness definition is similar to the definition of selective unbiasedness (Ψ -unbiasedness) from [9, 13, 65]. However, here we also incorporate the model misspecification by taking the PS-pseudo-true parameter vector w.r.t. the assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, into account. It can be seen that for the special case of a single candidate model ($K = 1$), the PSMS-unbiasedness (4.42) coincides with the MS-unbiasedness under misspecification in Definition 4.2. Finally, the PSMS-unbiasedness definition from Definition 4.4 can be interpreted as the unbiasedness in the Lehmann sense [146] w.r.t. the k th-MSE defined in (4.41). This can be seen by setting the k th-MSE from (4.41) in the definition of Lehmann unbiasedness.

4.6 Post-Selection Pseudo-True Parameter Vectors

In Definition 4.3, the PS-pseudo-true parameter vector is defined for a generic assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$. In this subsection, we derive the pseudo-true parameter vectors of the three interpretations of post-model-selection estimation.

4.6.1 Naive Interpretation

By substituting $f(\mathbf{x}; \boldsymbol{\theta}) = f_I(\mathbf{x}; \boldsymbol{\theta})$ from (4.17) in (4.40), we obtain that the k th PS-pseudo-true parameter vector under the naive interpretation is

$$\boldsymbol{\vartheta}_I^{(k)} = \arg \max_{\boldsymbol{\theta}^{(k)} \in \Omega_k} \mathbb{E}_p [\log f_k(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k], \quad (4.43)$$

$k = 1, \dots, K$. As explained in Subsection 4.3.1, $f_I(\mathbf{x}; \boldsymbol{\theta})$ is not a valid pdf. Thus, using Definition 4.1 will not result in a KLD measure in the conventional sense. On the other hand, Definition 4.3 implies that $\boldsymbol{\vartheta}_I^{(k)}$ is the minimizer of the valid KLD between $f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)})$ and the *conditional* true pdf, $p(\mathbf{x}; \boldsymbol{\varphi} | \Psi = k)$, in the support \mathcal{A}_k . The relationship between the two definitions of pseudo-true parameter vectors for the naive interpretation is described in the following claim.

Claim 4.1. *The PS-pseudo-true parameter vector under the naive interpretation, $\boldsymbol{\vartheta}_I \triangleq [(\boldsymbol{\vartheta}_I^{(1)})^T, \dots, (\boldsymbol{\vartheta}_I^{(K)})^T]^T$, where $\boldsymbol{\vartheta}_I^{(k)}$ is defined in (4.43), i.e. according to Definition 4.3, coincides with the pseudo-true parameter vector according to Definition 4.1 under the naive interpretation.*

Proof. By substituting $f(\mathbf{x}; \boldsymbol{\theta}) = f_I(\mathbf{x}; \boldsymbol{\theta})$ in (4.4), we obtain that, according to Definition 4.1, the pseudo-true parameter under the first, naive interpretation from Subsection 4.3.1 is

$$\tilde{\boldsymbol{\vartheta}}_I \triangleq \arg \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_p [\log f_I(\mathbf{x}; \boldsymbol{\theta})]. \quad (4.44)$$

By substituting (4.17) in the r.h.s of (4.44), we obtain that

$$\begin{aligned} \mathbb{E}_p [\log f_I(\mathbf{x}; \boldsymbol{\theta})] &= \int_{\Omega_{\mathbf{x}}} p(\mathbf{x}; \boldsymbol{\varphi}) \log \left(\sum_{k=1}^K f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k} \right) d\mathbf{x} \\ &= \sum_{k=1}^K \int_{\mathcal{A}_k} p(\mathbf{x}; \boldsymbol{\varphi}) \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) d\mathbf{x}, \end{aligned} \quad (4.45)$$

where the last equality is obtained by changing the order of summing and integration, and using $\mathcal{A}_k \cap \mathcal{A}_m = \emptyset$, $m \neq k$, and the indicator function properties. By substituting (4.38) in (4.45) and using the conditional expectation definition, one obtains

$$\mathbb{E}_p [\log f_I(\mathbf{x}; \boldsymbol{\theta})] = \sum_{k=1}^K p_k \mathbb{E}_p [\log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) | \Psi = k]. \quad (4.46)$$

By substituting (4.46) in (4.44), we obtain that the maximization in (4.44) is separable w.r.t. the parameters of the different models. Thus, it can be implemented by solving the K maximization problems in (4.43), i.e. $\tilde{\boldsymbol{\vartheta}}_I = \boldsymbol{\vartheta}_I$, which completes the proof. \square

4.6.2 Normalized Interpretation

By substituting $f(\mathbf{x}; \boldsymbol{\theta}) = f_{II}(\mathbf{x}; \boldsymbol{\theta})$ from (4.21) in (4.40), we obtain that the PS-pseudo-true parameter vector under the normalized interpretation is

$$\boldsymbol{\vartheta}_{II} = \arg \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_p [\log f_k(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k] - \log \alpha(\boldsymbol{\theta}). \quad (4.47)$$

In a similar manner to the MSNL estimator in (4.30), $\log \alpha(\boldsymbol{\theta})$ is a function of all the parameters $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(K)}$. Nevertheless, since $\boldsymbol{\theta}^{(l)}$, $l \neq k$, appears only in $\alpha(\boldsymbol{\theta})$, and since the logarithmic function is a monotonically increasing function, the maximization is obtained by substituting the minimal values of π_l from (4.32). Therefore, (4.47) is

equivalent to

$$\boldsymbol{\vartheta}_{II}^{(k)} = \arg \max_{\boldsymbol{\theta}^{(k)} \in \Omega_k} \mathbb{E}_p [\log f_k(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k] - \log \left(\pi_k(\boldsymbol{\theta}^{(k)}) + \sum_{l \neq k} \pi_l \right). \quad (4.48)$$

The maximization in (4.48) is similar to (4.43) with an additional penalty term in the last row.

4.6.3 Selective Inference Interpretation

By substituting $f(\mathbf{x}; \boldsymbol{\theta}) = f_{III}(\mathbf{x}; \boldsymbol{\theta})$ from (4.25) in (4.40), we obtain that the k th PS-pseudo-true parameter vector under the selective inference interpretation is given by

$$\boldsymbol{\vartheta}_{III}^{(k)} = \arg \max_{\boldsymbol{\theta}^{(k)} \in \Omega_k} \mathbb{E}_p [\log f_k(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k] - \log \pi_k(\boldsymbol{\theta}^{(k)}). \quad (4.49)$$

One can notice that the maximization in (4.49) is similar to (4.43) with a penalty term manifested by $\log \pi_k(\boldsymbol{\theta}^{(k)})$. In other words, under the naive interpretation, the PS-pseudo-true parameter vector is obtained by minimization of the KLD between the true model conditioned by the selection $\Psi = k$, and the k th model conditioned by the selection $\Psi = k$. The following claim describes the relationship between this definition and the pseudo-true parameter vector in Definition 4.1.

Claim 4.2. *The PS-pseudo-true parameter vector under the selective inference interpretation, $\boldsymbol{\vartheta}_{III} \triangleq [(\boldsymbol{\vartheta}_{III}^{(1)})^T, \dots, (\boldsymbol{\vartheta}_{III}^{(K)})^T]^T$, where $\boldsymbol{\vartheta}_{III}^{(k)}$ is defined in (4.49), i.e. according to Definition 4.3, coincides with the pseudo-true parameter vector according to Definition 4.1 under the selective inference interpretation.*

Proof. By substituting $f(\mathbf{x}; \boldsymbol{\theta}) = f_{III}(\mathbf{x}; \boldsymbol{\theta})$ in (4.4), we obtain that, according to Definition 4.1, the pseudo-true parameter under the selective inference interpretation is

$$\tilde{\boldsymbol{\vartheta}}_{III} \triangleq \arg \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_p [\log f_{III}(\mathbf{x}; \boldsymbol{\theta})]. \quad (4.50)$$

By substituting (4.25) in (4.50) we obtain that

$$\begin{aligned} \mathbb{E}_p [\log f_{III}(\mathbf{x}; \boldsymbol{\theta})] &= \int_{\Omega_{\mathbf{x}}} p(\mathbf{x}; \boldsymbol{\varphi}) \log \left(\sum_{k=1}^K c_k f_k(\mathbf{x} | \Psi = k; \boldsymbol{\theta}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k} \right) d\mathbf{x} \\ &= \sum_{k=1}^K \int_{\mathcal{A}_k} p(\mathbf{x}; \boldsymbol{\varphi}) \log f_k(\mathbf{x} | \Psi = k; \boldsymbol{\theta}^{(k)}) d\mathbf{x} + p_k \log c_k, \end{aligned} \quad (4.51)$$

where the second equality is obtained by changing the order of summing and using the

property of summing over non-overlapping events, as in (4.45). By using (4.38) and the definition of conditional expectation, it can be verified that for any measurable function, $g(\mathbf{x})$,

$$\int_{\mathcal{A}_k} p(\mathbf{x}; \boldsymbol{\varphi}) g(\mathbf{x}) d\mathbf{x} = p_k \mathbb{E}_p[g(\mathbf{x}) | \Psi = k]. \quad (4.52)$$

Therefore, (4.51) can be written as

$$\mathbb{E}_p[\log f_{III}(\mathbf{x}; \boldsymbol{\theta})] = \sum_{k=1}^K p_k \left(\mathbb{E}_p[\log f_k(\mathbf{x} | \Psi = k; \boldsymbol{\theta}^{(k)}) | \Psi = k] + \log c_k \right). \quad (4.53)$$

By substituting (4.53) in (4.50), we obtain that the maximization in (4.50) is separable w.r.t. the parameters of the different models, and thus it can be implemented by solving the K maximization problems in (4.49), i.e. $\tilde{\boldsymbol{\vartheta}}_{III} = \boldsymbol{\vartheta}_{III}$, which completes the proof. \square

4.7 Post-Model-Selection Misspecified Cramér-Rao-Type Lower Bounds

The conventional MCRB presented in Section 4.1 is a Cramér-Rao-type bound on the MSMSE that takes into account model misspecification. In this section, we derive the PS-MCRB for performance analysis and system design in the framework of post-model-selection estimation. First, we define the regularity conditions in Subsection 4.7.1. Then, we derive the MCRBs that incorporate the model-selection procedure as the form of model misspecification. In particular, in Subsection 4.7.2, we develop the marginal k th PS-MCRB on the k th-MSE for a given selection, $k \in \{1, \dots, K\}$. Then, in Subsection 4.7.3, we introduce a global MSE and present a global PS-MCRB across all selections. In Section 4.8, we derive the PS-MCRB according to the three interpretations presented in Section 4.3. Finally, in Subsection 4.8.4, we discuss the properties of the PS-MCRB and present some special cases.

4.7.1 Post Selection Regularity Conditions

Based on the regularity conditions of the MCRB for the general case (see e.g. [87, 90]), we define the following regularity conditions for the post-model-selection scheme with a general assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, which can be e.g. one of the assumed pdfs, $f_i(\mathbf{x}; \boldsymbol{\theta})$, $i = I, II, III$.

RC.1. The maximum of $\mathbb{E}_p[\log f(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k]$ w.r.t. $\boldsymbol{\theta}^{(k)}$ from (4.40) is unique interior

point, $\forall k \in \{1, \dots, K\}$.

RC.2. The log-likelihood function, $\log f(\mathbf{x}; \boldsymbol{\theta})$, is a twice differentiable function, and the functions $\left| \frac{\partial \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i} \right|$ and $\left| \frac{\partial^2 \log f(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i^2} \right|$, $i = 1, \dots, |\Omega_\Theta|$, are dominated by a function $m(\mathbf{x})$, which is a square-integrable function w.r.t. the conditional true pdf, $p(\mathbf{x}|\Psi = k)$ from (4.38), $\forall k \in \{1, \dots, K\}$.

RC.3. There is a neighborhood of a general PS-pseudo-true parameter vector, $\boldsymbol{\vartheta}^{(k)}$, such that for every $\boldsymbol{\theta}^{(k)}$ in this neighborhood

$$\left(\frac{1}{f(\mathbf{x}; \boldsymbol{\vartheta}^{(k)})} \left| \frac{\partial f(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i^{(k)}} \right| \right) \leq m(\mathbf{x}), \quad (4.54)$$

where $m(\mathbf{x})$ is a square-integrable function w.r.t. $p(\mathbf{x}|\Psi = k)$, defined in (4.38).

RC.4. The k th post-model-selection Hessian form information matrix, evaluated at the k th PS-pseudo-true parameter vector

$$\mathbf{A}^{(k)}(\boldsymbol{\vartheta}^{(k)}) \triangleq \mathbb{E}_p \left[\nabla_{\boldsymbol{\theta}^{(k)}}^2 \log f(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k \right] \Big|_{\boldsymbol{\theta}^{(k)} = \boldsymbol{\vartheta}^{(k)}}, \quad (4.55)$$

is a $|\Omega_k| \times |\Omega_k|$ non-singular matrix, where the conditional expectation in (4.55) is obtained by integration w.r.t. the conditional true pdf $p(\mathbf{x}|\Psi = k)$ from (4.38), $\forall k \in \{1, \dots, K\}$.

regularity condition RC.1 ensures the uniqueness of the pseudo-true parameter vector, $\boldsymbol{\vartheta}^{(k)}$. regularity conditions RC.2 and RC.3 enable differentiation under the integral sign of the conditional expectation of any finite-variance function of \mathbf{x} . regularity condition RC.4 ensures that the inverse of $\mathbf{A}^{(k)}(\boldsymbol{\vartheta}^{(k)})$ from (4.55) is well defined. All regularity conditions are w.r.t. the conditional true pdf in (4.38), which accounts for the selection approach. One can notice that these regularity conditions are the similar to regularity conditions regularity conditions RC.1–RC.3 presented for the conventional case in Subsection 4.1.0.1. In fact, The only difference is that here we replaced the true pdf with the conditional true pdf, $p(\mathbf{x}|\Psi = k)$.

4.7.2 Marginal PS-MCRB

In the following, we derive the marginal PS-MCRBs on the k th-MSE under a generally assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, and under the regularity conditions from Subsection 4.7.1.

Theorem 4.2. (*k*th PS-MCRB) Let $f(\mathbf{x}; \boldsymbol{\theta})$ be a general assumed pdf for a post-model-selection scheme with a selection rule Ψ that satisfy regularity conditions RC.1–RC.4. The *k*th-MSE of any finite variance, PSMS-unbiased estimator, $\hat{\boldsymbol{\theta}}^{(k)}$, satisfies

$$\mathbf{MSE}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\vartheta}^{(k)}) \succeq \mathbf{MCRB}^{(k)}(\boldsymbol{\vartheta}^{(k)}), \quad (4.56)$$

where the *k*th-PS-MCRB is given by

$$\mathbf{MCRB}^{(k)}(\boldsymbol{\vartheta}^{(k)}) \triangleq (\mathbf{A}^{(k)}(\boldsymbol{\vartheta}^{(k)}))^{-1} \mathbf{B}^{(k)}(\boldsymbol{\vartheta}^{(k)}) (\mathbf{A}^{(k)}(\boldsymbol{\vartheta}^{(k)}))^{-1}, \quad (4.57)$$

the pseudo-true parameter vector $\boldsymbol{\vartheta}^{(k)}$ and the post-model-selection Hessian form information matrix, $\mathbf{A}^{(k)}(\boldsymbol{\vartheta}^{(k)})$, are defined in (4.37) and (4.55), respectively, and the outer-product form of the *k*th post-model-selection information matrix is

$$\mathbf{B}^{(k)}(\boldsymbol{\vartheta}^{(k)}) \triangleq \mathbb{E}_p \left[\nabla_{\boldsymbol{\theta}^{(k)}} \log f(\mathbf{x}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}^{(k)}}^H \log f(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k \right]. \quad (4.58)$$

Proof. The proof of Theorem 4.2 can be obtained along the lines of the proof of the MCRB for the conventional case from Theorem 4.1, by replacing the true pdf $p(\mathbf{x})$ with the conditional true pdf, $p(\mathbf{x} | \Psi = k)$, and replacing the estimator $\hat{\boldsymbol{\theta}}$ with an estimator of the *k*th parameter vector, $\hat{\boldsymbol{\theta}}^{(k)}$. As a result, all the expectations involved in the terms $\mathbf{MSE}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\vartheta}^{(k)})$, $\mathbf{A}^{(k)}(\boldsymbol{\vartheta}^{(k)})$, and $\mathbf{B}^{(k)}(\boldsymbol{\vartheta}^{(k)})$ are conditional expectations that are computed using integration w.r.t. the conditional true pdf, $p(\mathbf{x} | \Psi = k)$. \square

4.7.3 Global PS-MCRB

The bound in Theorem 4.2 provides a marginal lower bound on each *k*th-MSE from (4.41). In general, this *k*th-MSE matrix may have different dimensions for every *k*, $|\Omega_k| \times |\Omega_k|$, where $|\Omega_k|$ is the dimensions of the *k*th parameter vector, $\boldsymbol{\theta}^{(k)}$. Thus, to derive a global bound, a general estimation task should be considered across all models. In this subsection, we consider the true pdf to be a parametric model, $p(\mathbf{x}; \boldsymbol{\varphi})$, where $\boldsymbol{\varphi} \in \Omega_\varphi$ is a deterministic parameter vector and Ω_φ denotes the true parameter space.

In this subsection, we use the additional assumption that for any candidate model there is a deterministic, continuously differentiable mapping from the *k*th model parameter space to Ω_φ , represented by $\boldsymbol{\varphi}_k : \Omega_k \rightarrow \Omega_\varphi$, such that $\boldsymbol{\varphi}_k(\boldsymbol{\theta}^{(k)}) = \boldsymbol{\varphi}$. Therefore, any practical post-model-selection estimator of $\boldsymbol{\varphi}$, $\hat{\boldsymbol{\varphi}} : \Omega_{\mathbf{x}} \rightarrow \Omega_\varphi$, can be written in the following form:

$$\hat{\boldsymbol{\varphi}} = \sum_{k=1}^K \boldsymbol{\varphi}_k(\hat{\boldsymbol{\theta}}^{(k)}) \mathbb{1}_{\mathbf{x} \in \mathcal{A}_k}. \quad (4.59)$$

The MSE of an estimator $\hat{\varphi}$ from (4.59) is defined as

$$\mathbf{MSE}(\hat{\varphi}, \varphi) \triangleq \mathbb{E}_p \left[(\hat{\varphi} - \varphi)(\hat{\varphi} - \varphi)^H \right]. \quad (4.60)$$

To incorporate the model-selection stage and analyze the estimators under each selected model separately, we decompose the MSE from (5.3) by substituting a general estimator $\hat{\varphi}$ from (4.59) and using the law of total expectation:

$$\mathbf{MSE}(\hat{\varphi}, \varphi) = \sum_{k=1}^K p_k \mathbb{E}_p \left[(\varphi_k(\hat{\boldsymbol{\theta}}^{(k)}) - \varphi)(\varphi_k(\hat{\boldsymbol{\theta}}^{(k)}) - \varphi)^H | \Psi = k \right], \quad (4.61)$$

where p_k is defined in (4.12), and is a function of φ . The conditional expectations in (4.61) are calculated w.r.t. the conditional true pdfs in (4.38). The following Theorem uses Theorem 4.2 to obtain the PS-MCRB on the global MSE from (5.3).

Theorem 4.3. (*PS-MCRB*) *Let us assume a post-model-selection model with the assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, and a selection rule Ψ that satisfy regularity conditions RC.1–RC.4, $\forall k \in \{1, \dots, K\}$. The MSE of any post-model-selection estimator $\hat{\varphi}$ in (4.59), which consists of the PSMS-unbiased for all $k \in \{1, \dots, K\}$, satisfies*

$$\mathbf{MSE}(\hat{\varphi}, \varphi) \succeq \sum_{k=1}^K p_k \left(\dot{\varphi}_k^H(\boldsymbol{\vartheta}^{(k)}) \mathbf{MCRB}^{(k)} \dot{\varphi}_k(\boldsymbol{\vartheta}^{(k)}) + (\varphi_k(\boldsymbol{\vartheta}^{(k)}) - \varphi)(\varphi_k(\boldsymbol{\vartheta}^{(k)}) - \varphi)^H \right), \quad (4.62)$$

where

$$\dot{\varphi}_k(\boldsymbol{\vartheta}^{(k)}) \triangleq \nabla_{\boldsymbol{\theta}^{(k)}} \varphi_k^H(\boldsymbol{\theta}^{(k)}) \Big|_{\boldsymbol{\theta}^{(k)} = \boldsymbol{\vartheta}^{(k)}} \quad (4.63)$$

is the $|\Omega_k| \times |\Omega_\varphi|$ Jacobian matrix of the k th mapping $\varphi_k(\cdot)$ evaluated at the relevant pseudo-true parameter vector, $\boldsymbol{\vartheta}^{(k)}$.

Proof. Since we assumed that for any k , $\varphi_k(\cdot)$ is a continuously differentiable mapping, similar to the conventional CRB on a functional transformation of the unknown parameter vector [87], the marginal bound in (4.56) can be generalized to the estimation of

$$\begin{aligned} & \mathbb{E}_p \left[(\varphi_k(\hat{\boldsymbol{\theta}}^{(k)}) - \varphi)(\varphi_k(\hat{\boldsymbol{\theta}}^{(k)}) - \varphi)^H | \Psi = k \right] \\ & \succeq \dot{\varphi}_k^H(\boldsymbol{\vartheta}^{(k)}) \mathbf{MCRB}^{(k)} \dot{\varphi}_k(\boldsymbol{\vartheta}^{(k)}) + (\varphi_k(\boldsymbol{\vartheta}^{(k)}) - \varphi)(\varphi_k(\boldsymbol{\vartheta}^{(k)}) - \varphi)^H, \end{aligned} \quad (4.64)$$

where the last term in (4.64) is since the transformation of the pseudo-true parameter vector, $\varphi_k(\boldsymbol{\vartheta}^{(k)})$, is not necessarily equal to φ . By plugging the marginal bounds from (4.64) for $k = 1, \dots, K$ in (4.61), and since the probabilities of selection, $\{p_k\}_{k=1}^K$, are non-negative, we obtain (4.62). \square

4.8 Interpretations of the PS-MCRB

In Theorems 4.2 and 4.3, we derived the k th-PS-MCRB and PS-MCRB for a general assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$. In the following, we implement the associated post-model-selection information matrices $\mathbf{A}^{(k)}(\cdot)$ and $\mathbf{B}^{(k)}(\cdot)$ from (4.55) and (4.58) for the three interpretations from Section 4.3. The different post-model-selection information matrices lead to different MCRBs.

4.8.1 Naive Interpretation

We note that (4.16) implies that

$$\nabla_{\boldsymbol{\theta}^{(k)}} \log f_I(\mathbf{x}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}), \quad \forall \mathbf{x} \in \mathcal{A}_k, \quad (4.65)$$

$k \in \{1, \dots, K\}$. Thus, by substituting (4.65) in (4.55) and (4.58) we obtain that the k th post-model-selection Hessian form and outer-product form information matrices, in this case, are

$$\mathbf{A}_I^{(k)}(\boldsymbol{\theta}^{(k)}) = E_p \left[\nabla_{\boldsymbol{\theta}^{(k)}}^2 \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) | \Psi = k \right] \quad (4.66)$$

and

$$\mathbf{B}_I^{(k)}(\boldsymbol{\theta}^{(k)}) = E_p \left[\nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) \nabla_{\boldsymbol{\theta}^{(k)}}^H \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) | \Psi = k \right], \quad (4.67)$$

respectively.

Although $f_I(\mathbf{x}; \boldsymbol{\theta})$ is not a valid pdf (see Subsection 4.3.1), the marginal PS-MCRB under the naive interpretation is a valid bound on the k th MSE of any PSMS-unbiased estimator w.r.t to the naive interpretation. This can be explained as follows. By considering $f(\mathbf{x}; \boldsymbol{\theta}) = f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)})$ as the assumed pdf for any $\mathbf{x} \in \Omega_{\mathbf{x}}$, the k th PS pseudo-true vector, $\boldsymbol{\vartheta}^{(k)}$ is identical by definition to the k th PS pseudo-true vector under the naive interpretation, $\boldsymbol{\vartheta}_I^{(k)}$. This implies that for $\mathbf{x} \in \mathcal{A}_k$, the PSMS-unbiasedness definition is the same as for the naive interpretation. Thus, if $f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)})$ satisfies regularity conditions RC.1–RC.4, then Theorem 4.2 applies for this case. In addition, (4.65) implies that the matrices $\mathbf{A}^{(k)}(\boldsymbol{\theta}^{(k)})$, $\mathbf{B}^{(k)}(\boldsymbol{\theta}^{(k)})$ coincide with with the matrices under the naive interpretation from (4.66) and (4.67), respectively. Therefore, the marginal PS-MCRB under this consideration coincides with the marginal PS-MCRB under the naive interpretation.

4.8.2 Normalized Interpretation

To obtain the k th post-model-selection information matrices under the normalized interpretation, we use the derivative of $\log f_{II}(\mathbf{x}; \boldsymbol{\theta})$ from (4.19):

$$\nabla_{\boldsymbol{\theta}^{(k)}} \log f_{II}(\mathbf{x}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) - \nabla_{\boldsymbol{\theta}^{(k)}} \log \alpha(\boldsymbol{\theta}), \quad \forall \mathbf{x} \in \mathcal{A}_k, \quad (4.68)$$

$k = 1, \dots, K$. By substituting (4.68) in (4.55) we obtain that

$$\mathbf{A}_{II}^{(k)}(\boldsymbol{\theta}) = \mathbf{A}_I^{(k)}(\boldsymbol{\theta}^{(k)}) - \nabla_{\boldsymbol{\theta}^{(k)}}^2 \log \alpha(\boldsymbol{\theta}), \quad (4.69)$$

where $\mathbf{A}_I^{(k)}(\cdot)$ is defined in (4.66). Similarly, by substituting (4.68) in (4.58) and using the fact that $\nabla_{\boldsymbol{\theta}^{(k)}} \log \alpha(\boldsymbol{\theta})$ is deterministic, we obtain that

$$\begin{aligned} \mathbf{B}_{II}^{(k)}(\boldsymbol{\theta}^{(k)}) &= \mathbf{B}_I^{(k)}(\boldsymbol{\theta}^{(k)}) \\ &\quad - \mathbb{E}_p[\nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k] \nabla_{\boldsymbol{\theta}^{(k)}}^H \log \alpha(\boldsymbol{\theta}) \\ &\quad - \nabla_{\boldsymbol{\theta}^{(k)}} \log \alpha(\boldsymbol{\theta}) \mathbb{E}_p[\nabla_{\boldsymbol{\theta}^{(k)}}^H \log f_k(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k] \\ &\quad + \nabla_{\boldsymbol{\theta}^{(k)}} \log \alpha(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}^{(k)}}^H \log \alpha(\boldsymbol{\theta}), \end{aligned} \quad (4.70)$$

where $\mathbf{B}_I^{(k)}(\cdot)$ is defined in (4.67). Since the pseudo-true parameter vector under this interpretation, $\boldsymbol{\vartheta}_{II}$, maximizes the r.h.s. of (4.47), then under regularity condition RC.2 for this case (i.e. twice differentiability of $f_{II}(\mathbf{x}; \boldsymbol{\theta})$), $\boldsymbol{\vartheta}_{II}^{(k)}$ is a stationary point that satisfies $\nabla_{\boldsymbol{\theta}^{(k)}} \log f_{II}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{0}$, which by (4.68) implies that

$$\nabla_{\boldsymbol{\theta}^{(k)}} \log \alpha(\boldsymbol{\vartheta}_{II}) = \mathbb{E}_p[\nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k] |_{\boldsymbol{\theta}^{(k)} = \boldsymbol{\vartheta}_{II}^{(k)}}, \quad (4.71)$$

$\forall k \in \{1, \dots, K\}$. By substituting (4.71) in (4.70) we obtain that

$$\mathbf{B}_{II}^{(k)}(\boldsymbol{\vartheta}_{II}^{(k)}) = \mathbf{B}_I^{(k)}(\boldsymbol{\vartheta}_{II}^{(k)}) - \nabla_{\boldsymbol{\theta}^{(k)}} \log \alpha(\boldsymbol{\vartheta}_{II}) \nabla_{\boldsymbol{\theta}^{(k)}}^H \log \alpha(\boldsymbol{\vartheta}_{II}). \quad (4.72)$$

It can be seen that $\mathbf{A}_{II}^{(k)}(\cdot)$ and $\mathbf{B}_{II}^{(k)}(\cdot)$ from (4.69) and (4.72), respectively, are both composed of the sum of the k th post-model-selection matrices of the naive interpretation, $\mathbf{A}_I^{(k)}(\cdot)$ and $\mathbf{B}_I^{(k)}(\cdot)$, and a second term which stems from the factor $\alpha(\boldsymbol{\theta})$ from (4.20) and is determined by the selection approach. However, the PS-MCRB under this normalized interpretation is based on evaluating $\mathbf{A}_{II}^{(k)}(\cdot)$ and $\mathbf{B}_{II}^{(k)}(\cdot)$ at the pseudo-true parameter vector under this interpretation, $\boldsymbol{\vartheta}_{II}$, and not at $\boldsymbol{\vartheta}_I$.

4.8.3 Selective Inference Interpretation

To obtain the k th post-model-selection information matrices under the selective inference interpretation, we use the derivative of $\log f_{III}(\mathbf{x}; \boldsymbol{\theta})$ from (4.25) (using the fact that c_k is not a function of \mathbf{x}):

$$\nabla_{\boldsymbol{\theta}^{(k)}} \log f_{III}(\mathbf{x}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) - \nabla_{\boldsymbol{\theta}^{(k)}} \log \pi_k(\boldsymbol{\theta}_{III}^{(k)}), \quad (4.73)$$

for any $\mathbf{x} \in \mathcal{A}_k$. By substituting (4.73) in (4.55) we obtain that

$$\mathbf{A}_{III}^{(k)}(\boldsymbol{\vartheta}_{III}^{(k)}) = \mathbf{A}_I^{(k)}(\boldsymbol{\vartheta}_{III}^{(k)}) - \nabla_{\boldsymbol{\theta}^{(k)}}^2 \log \pi_k(\boldsymbol{\vartheta}_{III}^{(k)}), \quad (4.74)$$

where $\mathbf{A}_I^{(k)}(\cdot)$ is defined in (4.66). Similarly, by substituting the gradient from (4.73) in (4.58) and using the fact that $\nabla_{\boldsymbol{\theta}^{(k)}} \log \pi_k(\boldsymbol{\vartheta}_{III}^{(k)})$ is deterministic, we obtain that

$$\begin{aligned} \mathbf{B}_{III}^{(k)}(\boldsymbol{\theta}^{(k)}) &= \mathbf{B}_I^{(k)}(\boldsymbol{\theta}^{(k)}) \\ &\quad - \mathbb{E}_p \left[\nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) | \Psi = k \right] \nabla_{\boldsymbol{\theta}^{(k)}}^H \log \pi_k(\boldsymbol{\theta}^{(k)}) \\ &\quad - \nabla_{\boldsymbol{\theta}^{(k)}} \log \pi_k(\boldsymbol{\theta}^{(k)}) \mathbb{E}_p \left[\nabla_{\boldsymbol{\theta}^{(k)}}^H \log f_k(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k \right] \\ &\quad + \nabla_{\boldsymbol{\theta}^{(k)}} \log \pi_k(\boldsymbol{\theta}^{(k)}) \nabla_{\boldsymbol{\theta}^{(k)}}^H \log \pi_k(\boldsymbol{\theta}^{(k)}), \end{aligned} \quad (4.75)$$

where $\mathbf{B}_I^{(k)}(\cdot)$ is defined in (4.67). Since the pseudo-true parameter vector, $\boldsymbol{\vartheta}_{III}^{(k)}$, maximizes the r.h.s. of (4.49), then under regularity condition RC.2 (i.e. twice differentiability of $f_{III}(\mathbf{x}; \boldsymbol{\theta})$), $\boldsymbol{\vartheta}_{III}^{(k)}$ is a stationary point that satisfy $\nabla_{\boldsymbol{\theta}^{(k)}} \log f_{III}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{0}$, which by (4.73) implies that

$$\nabla_{\boldsymbol{\theta}^{(k)}} \log \pi_k(\boldsymbol{\vartheta}_{III}^{(k)}) = \mathbb{E}_p \left[\nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}) | \Psi = k \right] \Big|_{\boldsymbol{\theta}^{(k)} = \boldsymbol{\vartheta}_{III}^{(k)}}, \quad (4.76)$$

$\forall k \in \{1, \dots, K\}$. By substituting (4.76) in (4.75) we obtain that

$$\mathbf{B}_{III}^{(k)}(\boldsymbol{\vartheta}_{III}^{(k)}) = \mathbf{B}_I^{(k)}(\boldsymbol{\vartheta}_{III}^{(k)}) - \nabla_{\boldsymbol{\theta}^{(k)}} \log \pi_k(\boldsymbol{\vartheta}_{III}^{(k)}) \nabla_{\boldsymbol{\theta}^{(k)}}^H \log \pi_k(\boldsymbol{\vartheta}_{III}^{(k)}). \quad (4.77)$$

It can be seen that $\mathbf{A}_{III}^{(k)}(\cdot)$ and $\mathbf{B}_{III}^{(k)}(\cdot)$ from (4.74) and (4.77), respectively, are composed of the sum of the k th post-model-selection information matrices of the naive interpretation, $\mathbf{A}_I^{(k)}(\cdot)$ and $\mathbf{B}_I^{(k)}(\cdot)$ from (4.66) and (4.67), and a second term which is determined by the model selection, based on derivatives of $\pi_k(\cdot)$. All these terms are evaluated at the associated pseudo-true parameter vector, $\boldsymbol{\vartheta}_{III}^{(k)}$.

4.8.4 Remarks and discussion

Remark 4.5. For the special case of a single candidate model from Remark 4.1, i.e. if $K = 1$, the post-model-selection pseudo-true parameter vector from Definition 4.3 coincides with the conventional pseudo-true parameter vector from Definition 4.1. regularity conditions RC.1–RC.4 coincides with the conventional regularity conditions in [87]. Moreover, the PS-MCRB coincides with the conventional MCRB described in Theorem 4.1.

Remark 4.6. The k th-MSE in (4.36) is defined w.r.t. the PS-pseudo-true parameter vector. Since each interpretation has different pseudo-true parameter vectors, the k th-MSE is defined differently for each interpretation. Thus, the k th-PS-MCRB under each interpretation is a lower bound on a different risk. Moreover, according to Definition 4.4, each interpretation induces a different PSMS-unbiasedness condition. Hence, each interpretation results in a bound for a different class of estimators.

Remark 4.7. Under mild regularity conditions [1] and if $\hat{\varphi}$ is a mean-unbiased estimator of φ , its MSE is bounded by the following oracle CRB:

$$\text{MSE}(\hat{\varphi}, \varphi) \succeq \text{E}_p[\nabla_{\varphi} \log p(\mathbf{x}; \varphi) \nabla_{\varphi}^H \log p(\mathbf{x}; \varphi)]. \quad (4.78)$$

The oracle CRB in (4.78) is commonly used in post-model-selection estimation analysis [13, 154, 155]. However, it is not a valid bound since it relies on knowledge of the true model, disregards the selection stage, and does not account for model misspecification, which impacts the estimation performance. Thus, it is only used here as a theoretical benchmark.

4.9 Example: Estimation after channel selection

Let $\mathbf{x} \in \mathbb{R}^N$ be an observation vector such that

$$\mathbf{x} = \varphi + \mathbf{w}, \quad (4.79)$$

where $\varphi \in \mathbb{R}^N$ is an unknown deterministic parameter vector to be estimated, and \mathbf{w} is white Gaussian noise with zero mean and a known covariance matrix, $\sigma^2 \mathbf{I}$. We consider two hypotheses regarding φ :

$$\begin{cases} \mathcal{H}_1 : \varphi = \mathbf{H}\boldsymbol{\theta}^{(1)} \\ \mathcal{H}_2 : \varphi = \boldsymbol{\theta}^{(2)}, \end{cases} \quad (4.80)$$

i.e. under \mathcal{H}_1 , the unknown parameter vector $\boldsymbol{\varphi} \in \mathbb{R}^N$ belongs to the column space of $\mathbf{H} \in \mathbb{R}^{N \times M}$, which is a known rank M matrix, where $M < N$. Under \mathcal{H}_2 , $\boldsymbol{\varphi} \in \mathbb{R}^N$ does not have a specific structure. Thus, the unknown parameter vector has a different dimension under each hypothesis (aka model): $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^M$ and $\boldsymbol{\theta}^{(2)} \in \mathbb{R}^N$. For example, in communication systems, \mathcal{H}_1 can describe a scenario where the signal is received from a known channel, \mathbf{H} , while hypothesis \mathcal{H}_2 describes the model of signal received from an unknown channel. In the considered setting, the candidate pdfs are both Gaussian with means $\mathbf{H}\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$, respectively.

Based on the observation vector \mathbf{x} one of the hypotheses/models is selected. We consider here the generalized information criterion (GIC) [153, 156] as the selection rule. It can be shown that for the considered model the GIC is given by

$$\Psi = \arg \max_{k \in \{1,2\}} -2 \log f_k(\mathbf{x}; \hat{\boldsymbol{\theta}}_{\text{MSL}}^{(k)}) + \gamma, \quad (4.81)$$

where $\hat{\boldsymbol{\theta}}_{\text{MSL}}^{(k)}$ is the MSL estimator from (4.29), which is also the marginal ML estimator under each model,

and γ is a penalty term that is determined according to the specific criterion. For example, for $\gamma = 2(N - M)$, we obtain the well-known AIC, and for $\gamma = \log(N)(N - M)$, we obtain the BIC, which coincides in this case with the MDL criterion [156]. Since $\{f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)})\}_{k=1}^2$ are Gaussian pdfs with the means from (4.79), (4.80), the MSL estimator is given by

$$\hat{\boldsymbol{\theta}}_{\text{MSL}}^{(k)} = \begin{cases} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}, & k = 1 \\ \mathbf{x}, & k = 2. \end{cases} \quad (4.82)$$

For the considered settings, by substituting the Gaussian pdfs and (4.82) in (4.81), we obtain that the following energy detector manifests the GIC

$$\Psi = \begin{cases} 1, & \frac{1}{\sigma^2} \mathbf{x}^T \mathbf{P}_{\mathbf{H}}^\perp \mathbf{x} \leq \gamma \\ 2, & \text{otherwise,} \end{cases} \quad (4.83)$$

where $\mathbf{P}_{\mathbf{H}}^\perp \triangleq \mathbf{I} - \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$.

In this case, any practical post-model-selection estimator of the parameter vector of interest $\boldsymbol{\varphi}$ has the following form:

$$\begin{aligned} \hat{\boldsymbol{\varphi}} &= \hat{\boldsymbol{\varphi}}^{(1)} \mathbb{1}_{\mathbf{x} \in \mathcal{A}_1} + \hat{\boldsymbol{\varphi}}^{(2)} \mathbb{1}_{\mathbf{x} \in \mathcal{A}_2} \\ &= \mathbf{H} \hat{\boldsymbol{\theta}}^{(1)} \mathbb{1}_{\mathbf{x} \in \mathcal{A}_1} + \hat{\boldsymbol{\theta}}^{(2)} \mathbb{1}_{\mathbf{x} \in \mathcal{A}_2}, \end{aligned} \quad (4.84)$$

where $\hat{\boldsymbol{\theta}}^{(k)}$, $k = 1, 2$, are the estimators of the unknown parameter vector under hypothesis $k = 1, 2$, and the events $\mathbf{x} \in \mathcal{A}_1$ and $\mathbf{x} \in \mathcal{A}_2$ are determined according to (4.83).

Under the considered settings, $\frac{1}{\sigma^2} \mathbf{x}^T \mathbf{P}_{\mathbf{H}}^\perp \mathbf{x}$ has a χ^2 distribution with $r \triangleq \text{Rank}(\mathbf{P}_{\mathbf{H}}^\perp) = N - M$ degrees of freedom, and non-centrality parameter [135, Ch. 2.3]

$$\lambda \triangleq \frac{\boldsymbol{\varphi}^T \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\varphi}}{\sigma^2}. \quad (4.85)$$

Thus, the true probabilities of selection from (4.12) in this case are

$$p_k = \begin{cases} F_r(\gamma; \lambda), & k = 1 \\ 1 - F_r(\gamma; \lambda), & k = 2, \end{cases} \quad (4.86)$$

where $F_r(\cdot; \lambda)$ is the χ^2 cdf with r degrees of freedom and non-centrality parameter λ . Since $\mathbf{P}_{\mathbf{H}}^\perp \mathbf{H} = \mathbf{0}$, under \mathcal{H}_1 the non-centrality in (4.85) vanishes and the test has a central χ^2 distribution $\forall \boldsymbol{\theta}^{(1)} \in \mathbb{R}^M$. Under \mathcal{H}_2 the non-centrality is given by

$$\lambda^{(2)} = \frac{(\boldsymbol{\theta}^{(2)})^T \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\theta}^{(2)}}{\sigma^2}. \quad (4.87)$$

Therefore, the assumed probabilities of selection from (4.13) are given by

$$\pi_k(\boldsymbol{\theta}^{(k)}) = \begin{cases} F_r(\gamma; 0), & k = 1 \\ 1 - F_r(\gamma; \lambda^{(2)}), & k = 2. \end{cases} \quad (4.88)$$

In the following, we present the estimators and bounds under the three interpretations for this scenario. Detailed derivations appear in Appendix C.

4.9.1 Estimators

In this subsection, we present the estimators from Section 4.4, according to each interpretation, for this example.

1. **Naive Interpretation:** The MSL estimator from (4.29) for these settings is (4.82) (see more in Appendix C.i.1).
2. **Normalized Interpretation:** Since π_1 is not a function of $\boldsymbol{\theta}^{(1)}$, $\alpha(\boldsymbol{\theta})$ is not a function of $\boldsymbol{\theta}^{(1)}$. Thus, the maximization in (4.30) for $k = 1$ w.r.t. $\boldsymbol{\theta}^{(1)}$ is equivalent in this case to the maximization in (4.29). Therefore, if $\mathbf{x} \in \mathcal{A}_1$, the MSNL estimator

is

$$\hat{\boldsymbol{\theta}}_{\text{MSNL}}^{(1)} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}. \quad (4.89)$$

If $\mathbf{x} \in \mathcal{A}_2$, the MSNL estimator is obtained by the maximization in (4.30) w.r.t. $\boldsymbol{\theta}^{(2)}$, which is obtained by setting the gradient of the r.h.s. of (4.30) to zero. In Appendix C.i.1, we show that this results in the following score equation:

$$\mathbf{x} - \boldsymbol{\theta}^{(2)} - \frac{F_r(\gamma; \lambda^{(2)}) - F_{r+2}(\gamma; \lambda^{(2)})}{F_r(\gamma; 0) + 1 - F_r(\gamma; \lambda^{(2)})} \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\theta}^{(2)} = \mathbf{0}. \quad (4.90)$$

Then, we set the MSNL estimator, $\hat{\boldsymbol{\theta}}_{\text{MSNL}}^{(2)}$, to be the solution of (4.90), which can be found numerically.

3. **Selective Inference Interpretation:** The PSML from Subsection 4.4.3 is given by the maximization in (4.35). Since in this case $\pi_1(\boldsymbol{\theta}^{(1)})$ is not a function of $\boldsymbol{\theta}^{(1)}$, the maximization in (4.35) w.r.t. $\boldsymbol{\theta}^{(1)}$ is equivalent to the maximization in (4.29), as before. Thus, if \mathcal{H}_1 is selected, the PSML estimator of $\boldsymbol{\theta}^{(1)}$ is

$$\hat{\boldsymbol{\theta}}_{\text{PSML}}^{(1)} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}, \quad (4.91)$$

which coincides with the MSL and MSNL estimators of $\boldsymbol{\theta}^{(1)}$ from (4.82) and (4.89), respectively. If \mathcal{H}_2 is selected, the PSML estimator of $\boldsymbol{\theta}^{(2)}$ is obtained by the maximization of (4.35) w.r.t. $\boldsymbol{\theta}^{(2)}$. This maximization is obtained by setting the gradient of the r.h.s. of (4.35) (with $\pi_k(\boldsymbol{\theta}^{(k)})$ from (4.88) and the Gaussian pdf $f_2(\mathbf{x}; \boldsymbol{\theta}^{(2)})$) to zero, which results in Appendix C.i.1

$$\mathbf{x} - \boldsymbol{\theta}^{(2)} - \frac{(F_r(\gamma; \lambda^{(2)}) - F_{r+2}(\gamma; \lambda^{(2)}))}{1 - F_r(\gamma; \lambda^{(2)})} \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\theta}^{(2)} = \mathbf{0}. \quad (4.92)$$

Then, we set the PSML estimator, $\hat{\boldsymbol{\theta}}_{\text{PSML}}^{(2)}$, to be the solution of (4.92), which can be found by numerical method.

4.9.2 Pseudo-True Parameter Vectors

In this subsection, we present the pseudo-true parameter vectors from Section 4.6, according to each interpretation. The full derivation appears in Appendix C.i.2.

1. **Naive Interpretation:** By substituting the considered setting in (4.43), we show in Appendix C.i.2 that the pseudo-true parameter vector according to the naive

interpretation is

$$\boldsymbol{\vartheta}_I^{(k)} = \begin{cases} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \boldsymbol{\mu}^{(1)}, & k = 1 \\ \boldsymbol{\mu}^{(2)}, & k = 2, \end{cases} \quad (4.93)$$

where the conditional expectation of \mathbf{x} given $\Psi = k$ is

$$\boldsymbol{\mu}^{(k)} \triangleq \mathbb{E}_p[\mathbf{x} | \Psi = k] = \boldsymbol{\varphi} + (-1)^k \frac{1}{p_k} (F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)) \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\varphi}, \quad k \in 1, 2. \quad (4.94)$$

The derivation of the closed-form expression for this conditional expectation can be found in Appendix C.ii.

2. **Normalized Interpretation:** Since in this case $\pi_1(\boldsymbol{\theta}^{(1)})$ is not a function of $\boldsymbol{\theta}^{(1)}$, the gradient of (4.48) w.r.t. $\boldsymbol{\theta}^{(1)}$ is as in the naive interpretation. Therefore, the pseudo-true parameter vector according to the normalized interpretation under \mathcal{H}_1 is

$$\boldsymbol{\vartheta}_{II}^{(1)} = \boldsymbol{\vartheta}_I^{(1)} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \boldsymbol{\mu}^{(1)}. \quad (4.95)$$

The pseudo-true parameter vector under \mathcal{H}_2 is obtained by setting the gradient of (4.48) w.r.t. $\boldsymbol{\theta}^{(2)}$ to zero, which results in

$$\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)} - \frac{(F_r(\gamma; \lambda^{(2)}) - F_{r+2}(\gamma; \lambda^{(2)}))}{\alpha(\boldsymbol{\theta})} \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\theta}^{(2)} = \mathbf{0}, \quad (4.96)$$

which can be solved numerically.

3. **Selective Inference Interpretation:** By substituting the considered setting in (4.49), we have

$$\boldsymbol{\vartheta}_{III}^{(1)} = \boldsymbol{\vartheta}_I^{(1)} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \boldsymbol{\mu}^{(1)}. \quad (4.97)$$

By setting the gradient of (4.49) w.r.t. $\boldsymbol{\theta}^{(2)}$ to zero, we obtain that the pseudo-true parameter vector under \mathcal{H}_2 ($\Psi = 2$) is

$$\boldsymbol{\vartheta}_{III}^{(2)} = \boldsymbol{\varphi}. \quad (4.98)$$

4.9.3 PS-MCRBs

In this subsection, we present the PS-MCRBs from Section 4.7, according to each interpretation. The Hessian matrix of the log-likelihood under each of the hypotheses is given

by

$$\nabla_{\boldsymbol{\theta}^{(k)}}^2 \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) = \begin{cases} -\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H}, & k = 1 \\ -\frac{1}{\sigma^2} \mathbf{I}, & k = 2, \end{cases} \quad (4.99)$$

which is a deterministic matrix independent of \mathbf{x} . Thus, by substituting (4.99) in (4.66), (4.69), and (4.74), we obtain the post-model-selection Hessian form information matrices:

$$\mathbf{A}_I^{(k)}(\boldsymbol{\vartheta}_I^{(k)}) = \begin{cases} -\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H}, & k = 1 \\ -\frac{1}{\sigma^2} \mathbf{I}, & k = 2, \end{cases} \quad (4.100)$$

$$\mathbf{A}_{II}^{(k)}(\boldsymbol{\vartheta}_{II}^{(k)}) = \begin{cases} -\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H}, & k = 1 \\ -\frac{1}{\sigma^2} \mathbf{I} - \nabla_{\boldsymbol{\theta}^{(2)}}^2 \log \alpha(\boldsymbol{\vartheta}_{II}), & k = 2, \end{cases} \quad (4.101)$$

and

$$\mathbf{A}_{III}^{(k)}(\boldsymbol{\vartheta}_{III}^{(k)}) = \begin{cases} -\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H}, & k = 1 \\ -\frac{1}{\sigma^2} \mathbf{I} - \nabla_{\boldsymbol{\theta}^{(2)}}^2 \log \pi_2(\boldsymbol{\vartheta}_{III}^{(2)}), & k = 2, \end{cases} \quad (4.102)$$

respectively. Closed-form expressions of $\nabla_{\boldsymbol{\theta}^{(2)}}^2 \log \pi_2(\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}^{(2)}}^2 \log \alpha(\boldsymbol{\theta})$ appear in Appendix C.i.3.

In addition, in Appendix C.i.3 it is shown that by substituting the considered settings and the associated pseudo-true parameter for each interpretation from Subsection C.i.2 in (4.67), (4.72) and (4.77), the outer-product form of the k th post-model-selection information matrices are identical:

$$\mathbf{B}_i^{(k)}(\boldsymbol{\vartheta}_i^{(k)}) = \begin{cases} \frac{1}{\sigma^4} \mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{x}}^{(1)} \mathbf{H}, & k = 1 \\ \frac{1}{\sigma^4} \boldsymbol{\Sigma}_{\mathbf{x}}^{(2)}, & k = 2, \end{cases} \quad (4.103)$$

where $i = I, II, III$ and

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{x}}^{(k)} &\triangleq \mathbb{E}_p[(\mathbf{x} - \boldsymbol{\mu}^{(k)})(\mathbf{x} - \boldsymbol{\mu}^{(k)})^T | \Psi = k] \\ &= \sigma^2 \mathbf{I} + (-1)^{k-1} \frac{1}{p_k} (F_r(\gamma; \lambda) - 2F_{r+2}(\gamma; \lambda) \\ &\quad + F_{r+4}(\gamma; \lambda)) \mathbf{P}_{\mathbf{H}}^{\perp} \boldsymbol{\varphi} \boldsymbol{\varphi}^T \mathbf{P}_{\mathbf{H}}^{\perp} \\ &\quad + (-1)^k \frac{1}{p_k} (F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)) \mathbf{P}_{\mathbf{H}}^{\perp} \sigma^2 \\ &\quad - \left(\frac{1}{p_k} (F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)) \right)^2 \mathbf{P}_{\mathbf{H}}^{\perp} \boldsymbol{\varphi} \boldsymbol{\varphi}^T \mathbf{P}_{\mathbf{H}}^{\perp}, \quad k \in 1, 2. \end{aligned} \quad (4.104)$$

A detailed derivation of this conditional covariance matrix is presented in Appendix C.ii.

4.9.4 Simulation Results

In Figs. 4.1 and 4.2, the performance of the MSL, MSNL, and PSML, estimators are presented and compared to their corresponding PS-MCRBs and to the oracle CRB, versus the threshold, γ , where the true hypothesis is \mathcal{H}_1 and \mathcal{H}_2 , i.e. where $\boldsymbol{\varphi} = \mathbf{H}\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\varphi} = \boldsymbol{\theta}^{(2)}$ in Figs. 4.1 and 4.2, respectively. In addition, in Fig. 4.1, we present the conventional MCRB that always assumes the wrong model (“anti-oracle”), $MCRB^{(2)}$. Similarly, in Fig. 4.2(a), we present the bias of the anti-oracle MML estimator, $\hat{\boldsymbol{\theta}}_{\text{MML}}^{(1)}$, and in Fig. 4.2(b), we present the conventional MCRB, $MCRB^{(1)}$. We use $N = 4, M = 2$ and the elements of $\mathbf{H} \in \mathbb{R}^{4 \times 2}$, $\boldsymbol{\theta}^{(1)}$, and $\boldsymbol{\theta}^{(2)}$ were generated once according to a standard Gaussian distribution. The performance is evaluated via 10^6 Monte-Carlo simulations for $\sigma^2 = 1$. The estimators and bounds appear in dashed and continuous lines, respectively. From the oracle point of view, i.e. if the true model is known, this is an estimation of the mean in a linear Gaussian model. Therefore, the oracle ML estimator achieves the oracle CRB in this case [1].

For this scenario of $N = 4, M = 2$, the threshold γ corresponding to the AIC is $\gamma = 2(N - M) = 4$, and similarly the BIC/MDL $\gamma = (N - M) \log(N) = 2 \log(4) \approx 2.77$. In Figs. 4.1 and 4.2(b) we show that in terms of the trace of the MSE matrix from (5.3), the PSML estimator outperforms the MSNL estimator, and both outperform the commonly-used MSL estimator. In addition, it can be seen that the proposed PS-MCRB for each interpretation is a valid bound that is more informative than the oracle CRB. In Fig. 4.2(a), the ℓ_1 norm of the bias, $\|\mathbb{E}_p[\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}]\|_1 = \sum_{n=1}^N |\mathbb{E}_p[\hat{\varphi}_n - \varphi_n]|$, is presented for the case where the true hypothesis is \mathcal{H}_2 . In this case, the estimators are biased in the conventional sense, since there is a probability of wrong selection that implies bias for any practical estimator. The oracle CRB is not a valid bound for these estimators, as can be seen in Fig. 4.2(b). In the case where the true hypothesis is \mathcal{H}_1 , the bias of all the estimators is negligible, and therefore is not shown here.

In both cases, for the smallest value of γ , $p_1 \approx 1$, and for the largest value $p_2 \approx 1$. At these extreme points, in practice, only one candidate model is selected. Thus, in Fig. 4.1, for the smallest γ , \mathcal{H}_2 (in this case, the wrong model) is selected a.s., and thus, all the estimators and bounds coincide with the conventional MCRB with $f_2(\mathbf{x}; \boldsymbol{\theta}^{(2)})$ as the assumed pdf. For the largest value of γ , \mathcal{H}_1 (the true model) is selected a.s., and thus, all the estimators and bounds coincide with the oracle ML and oracle CRB. Similarly, in Fig. 4.2, for the smallest γ , \mathcal{H}_2 (the true model) is selected a.s. and all estimators coincide with the oracle ML estimator, and the MSE of all the estimators and bounds coincide

with the oracle CRB. For the largest γ , \mathcal{H}_1 is (wrongly) selected a.s., and the biases of all the estimators coincide with the bias of the conventional MML that takes $f_1(\mathbf{x}; \boldsymbol{\theta}^{(1)})$ as the assumed pdf. Thus, all the estimators and bounds coincide with the conventional MCRB.

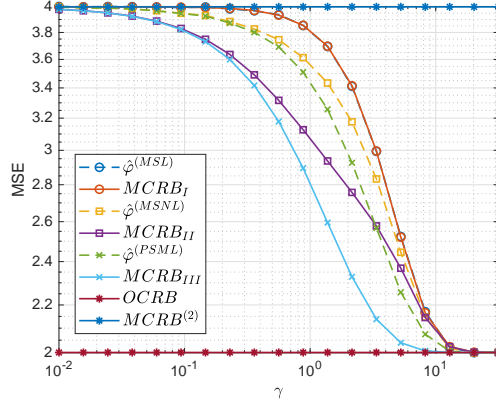


Figure 4.1: The MSE of the MSL, MSNL, PSML, and the oracle ML estimators, and PS-MCRBs and the oracle CRB versus the threshold, γ , where the true hypothesis is \mathcal{H}_1 .

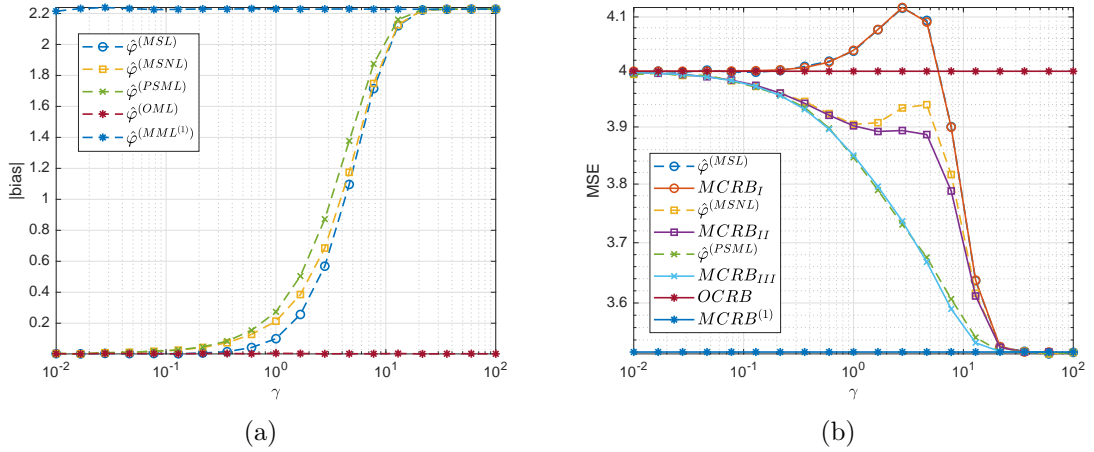


Figure 4.2: The bias (a) and MSE (b) of the MSL, MSNL, PSML, and the oracle ML estimators, and PS-MCRBs and the oracle CRB versus the threshold, γ , where the true hypothesis is \mathcal{H}_2 .

In Fig. 4.3, the performance of the MSL, MSNL, and PSML estimators is presented and compared with the corresponding PS-MCRBs and the oracle CRB versus M , the length of the parameter vector under hypothesis \mathcal{H}_1 , where the true hypothesis is \mathcal{H}_2 . The selection rule here is the BIC, i.e. obtained for $\gamma = (N - M) \log(N)$, where $N = 8$. The matrix $\mathbf{H} \in \mathbb{R}^{8 \times M}$ is taken such that its elements satisfy $\mathbf{H}_{i,j} = \frac{1}{1+|i-j|}$. The vector

$\boldsymbol{\theta}^{(2)}$ is set to the ones vector, i.e. $\theta_i^{(1)} = 1, i = 1, \dots, N$. It can be seen that the PSML estimator outperforms the MSNL estimator, and both outperform the commonly-used MSL estimator and that the proposed PS-MCRB for each interpretation is a valid bound that is more informative than the oracle CRB. We can see that in this scenario, the MSE of the post-selection estimators is non-monotonically w.r.t. M , this is since the value of M affects the performance, but in the case of the BIC is also affects the selection rule via the threshold γ .

In Fig. 4.4, the performance of the MSL, MSNL, and PSML estimators is presented and compared with the corresponding PS-MCRBs and the oracle CRB versus the signal to noise ratio (SNR) where $\text{SNR} \triangleq 10 \log \frac{\|\boldsymbol{\varphi}\|^2}{\sigma^2}$. In this case, the true hypothesis is \mathcal{H}_1 . The selection rule here is the AIC, i.e. obtained for $\gamma = 2(N - M)$, where $N = 3, M = 2$ and the matrix $\mathbf{H} \in \mathbb{R}^{3 \times 2}$ is taken such that its elements satisfy $\mathbf{H}_{i,j} = \frac{1}{1+|i-j|}$. The vector $\boldsymbol{\theta}^{(1)} = [1, 2]^T$. It can be seen that also in this scenario the PSML estimator outperforms the MSNL estimator, and both outperform the commonly-used MSL estimator. The proposed PS-MCRB for each interpretation is a valid bound that is more informative than the oracle CRB.

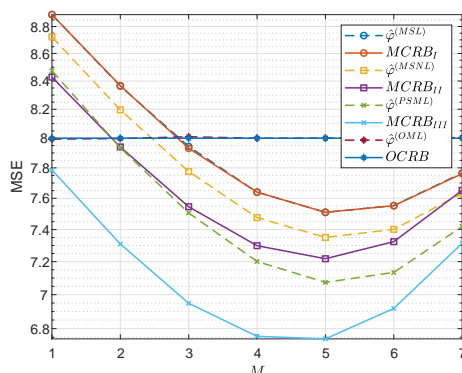


Figure 4.3: The MSE of the MSL, MSNL, PSML, and the oracle ML estimators, and PS-MCRBs and the oracle CRB versus M , where the true hypothesis is \mathcal{H}_2 for the BIC selection rule.

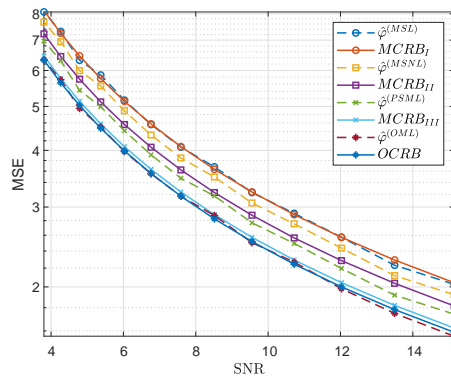


Figure 4.4: The MSE of the MSL, MSNL, PSML, and the oracle ML estimators, and PS-MCRBs and the oracle CRB versus the SNR, where the true hypothesis is \mathcal{H}_1 for the AIC selection rule.

Chapter 5

Bayesian Post-Model-Selection Estimation

In this chapter, we investigate the problem of post-model-selection estimation for the case of *random* parameters with a *deterministic* unknown support set, where the support set represents the unknown model. The analytical computation of the minimum mean-squared-error (MMSE) estimator and its corresponding performance are usually intractable [28, 157, 158]. This problem is even further exacerbated in the considered scenario when the exact observation model is unknown [14, 102, 159] and is a function of a deterministic, unknown, discrete-valued support set. Thus, derivation of lower bounds on the mean-squared-error is crucial for performance analysis and system design for post-model-selection estimation. To this end, we discuss the concept of coherent estimators that estimate the unselected parameters (by a given model-selection rule) to their prior mean. Then, we develop performance bounds for coherent estimators. Finally, we evaluate the suggested estimators and performance bounds via numerical simulations.

Different bounds have been developed in the literature for related estimation problems. The non-Bayesian Cramér-Rao bound has been developed for post-processing estimation of *deterministic* unknown parameters in the settings of estimation after *parameter-of-interest* selection with a known model [9, 65], estimation after detection [63, 80], and estimation after model-selection [13]. Hybrid Cramér-Rao-type bounds for joint random and deterministic unknown parameters [14, 102] assume *continuous*-valued deterministic parameters and, thus, are inappropriate for a *discrete*-valued support set, as in the considered scheme. A generalized Bayesian Cramér-Rao bound (BCRB) on the expected Optimal Subpattern Assignment (OSPA) [160], for tracking where both the number of targets and their positions should be estimated, has been proposed in [161]. However, the OSPA metric does not fit the considered scheme, in which the mapping between the

indices of the estimated vector and the unknown parameters is known. Finally, MSE performance bounds for the special case of sparse Bayesian learning and compressed sensing were derived and analyzed in the literature (see, e.g. [143, 162–164]). However, none of the existing bounds are appropriate for the post-model-selection scheme with an unknown random vector and deterministic support set.

The problem of Bayesian estimation after non-Bayesian model selection can be described as follows: we consider estimation of a random parameter vector, $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}} \subseteq \mathbb{R}^M$, with an *a-priori* probability density function (pdf), $f(\boldsymbol{\theta})$, that has a mean $\mathbb{E}[\boldsymbol{\theta}] = \boldsymbol{\mu}$. We consider the observation vector, $\mathbf{x} \in \Omega_{\mathbf{x}}$, with the conditional pdf $f(\mathbf{x}|\boldsymbol{\theta}_{\Lambda}; \Lambda)$, in which $\Lambda \in \mathcal{P}(\{1, \dots, M\})$ is an unknown deterministic set of indices, represents the unknown *model*. It is assumed that $f(\mathbf{x}|\boldsymbol{\theta}_{\Lambda}; \Lambda)$ belongs to a known set of pdfs, $\{f(\mathbf{x}|\boldsymbol{\theta}_{\Lambda_k}; \Lambda_k)\}_{k=1}^K$, where each support set of $\boldsymbol{\theta}$, Λ_k , represents a different candidate model. The *a-posteriori* pdf of the parameter vector, $\boldsymbol{\theta}$, given \mathbf{x} , under the k th hypothesis is denoted by $f(\boldsymbol{\theta}|\mathbf{x}; \Lambda_k)$. We assume that the elements of $\boldsymbol{\theta}$ are both independent and conditionally-independent, that is

$$f(\boldsymbol{\theta}|\mathbf{x}; \Lambda_k) = f(\boldsymbol{\theta}_{\Lambda_k}|\mathbf{x}; \Lambda_k)f(\boldsymbol{\theta}_{\Lambda_k^c}), \quad k = 1, \dots, K. \quad (5.1)$$

The goal here is to estimate the parameter vector $\boldsymbol{\theta}$ where the exact model, Λ , is unknown. However, model-selection and estimation performance are often two contradictory goals [72]. We consider here the practical procedure that is usually adopted in estimation after model selection. First, a model is selected by a predetermined selection-rule, which results in an estimated support set, $\hat{\Lambda} : \Omega_{\mathbf{x}} \rightarrow \mathcal{P}(\{1, \dots, M\})$. In the second stage, the parameters of the selected model, $\boldsymbol{\theta}_{\hat{\Lambda}}$, are estimated based on the same observation vector, \mathbf{x} , and the unselected parameters, $\boldsymbol{\theta}_{\hat{\Lambda}^c}$, are set to their prior mean. We consider this practice as “coherency” with the selection rule, similar to the non-Bayesian coherency definition in Definition 3.1.

Definition 5.1. *An estimator, $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \rightarrow \mathbb{R}^M$, is a coherent estimator of $\boldsymbol{\theta}$ w.r.t. the selection rule, $\hat{\Lambda}$, if*

$$\hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c} = \boldsymbol{\mu}_{\hat{\Lambda}^c}. \quad (5.2)$$

The MSE of an estimator, $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \rightarrow \mathbb{R}^M$, with a bounded second moment is defined as

$$\text{MSE}(\hat{\boldsymbol{\theta}}) \triangleq \mathbb{E} [(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}); \Lambda]. \quad (5.3)$$

The MSE from (5.3) can be decomposed as follows:

$$\text{MSE}(\hat{\boldsymbol{\theta}}) = \text{sMSE}(\hat{\boldsymbol{\theta}}) + \text{usMSE}(\hat{\boldsymbol{\theta}}), \quad (5.4)$$

where the selected MSE (sMSE) is

$$\text{sMSE}(\hat{\boldsymbol{\theta}}) \triangleq \text{E}[(\hat{\boldsymbol{\theta}}_{\hat{\Lambda}} - \boldsymbol{\theta}_{\hat{\Lambda}})^T (\hat{\boldsymbol{\theta}}_{\hat{\Lambda}} - \boldsymbol{\theta}_{\hat{\Lambda}}); \Lambda], \quad (5.5)$$

and the unselected MSE (usMSE) is

$$\text{usMSE}(\hat{\boldsymbol{\theta}}) \triangleq \text{E}[(\hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c} - \boldsymbol{\theta}_{\hat{\Lambda}^c})^T (\hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c} - \boldsymbol{\theta}_{\hat{\Lambda}^c}); \Lambda]. \quad (5.6)$$

By substituting (5.2) in (5.6), we obtain that the usMSE of any coherent estimator is given by

$$\text{usMSE}(\hat{\boldsymbol{\theta}}) = \text{E}[(\boldsymbol{\mu}_{\hat{\Lambda}^c} - \boldsymbol{\theta}_{\hat{\Lambda}^c})^T (\boldsymbol{\mu}_{\hat{\Lambda}^c} - \boldsymbol{\theta}_{\hat{\Lambda}^c}); \Lambda]. \quad (5.7)$$

The usMSE in (5.7) is only a function of the prior mean, $\boldsymbol{\mu}$, and the selected support set, $\hat{\Lambda}$. Therefore, we can conclude that the usMSE is identical for any coherent estimator. Thus, minimization of the MSE in (5.3) w.r.t. to a coherent estimator can be replaced by a minimization of the sMSE from (5.5).

5.1 Bayesian Post-Model-Selection Estimators

We present different estimators of $\boldsymbol{\theta}$ based on the observation vector, \mathbf{x} , for a given model-selection rule, $\hat{\Lambda}$.

5.1.1 Oracle MMSE

The oracle MMSE estimator [39, 40], which assumes knowledge of the true support set, Λ , is given by

$$\hat{\boldsymbol{\theta}}^{(\text{oMMSE})} = \text{E}[\boldsymbol{\theta} | \mathbf{x}; \Lambda]. \quad (5.8)$$

The oracle MMSE is not a coherent estimator in the sense of Definition 5.1 since it does not take into account the model-selection stage. Furthermore, it is not a practical estimator since the conditional expectation in (5.8) is a function of the unknown support set, Λ .

5.1.2 Oracle coherent MMSE (cMMSE)

The cMMSE estimator minimizes the MSE among the set of coherent estimators that satisfy (5.2). Thus, the cMMSE estimator is the solution of

$$\hat{\boldsymbol{\theta}}^{\text{cMMSE}} = \arg \min_{\hat{\boldsymbol{\theta}}} \text{MSE}(\hat{\boldsymbol{\theta}}), \quad \text{s.t. } \hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c} = \boldsymbol{\mu}_{\hat{\Lambda}^c}. \quad (5.9)$$

By using the constraint, $\hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c} = \boldsymbol{\mu}_{\hat{\Lambda}^c}$, and since according to (5.7) the usMSE is identical for any coherent estimator, the optimization in (5.9) is equivalent to

$$\begin{cases} \hat{\boldsymbol{\theta}}_{\hat{\Lambda}}^{(\text{cMMSE})} = \arg \min_{\hat{\boldsymbol{\theta}}_{\hat{\Lambda}}} \text{sMSE}(\hat{\boldsymbol{\theta}}) \\ \hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c}^{(\text{cMMSE})} = \boldsymbol{\mu}_{\hat{\Lambda}^c} \end{cases}. \quad (5.10)$$

By using the law of total expectation on (5.5), we obtain

$$\text{sMSE}(\hat{\boldsymbol{\theta}}) = \text{E} \left[\text{E}[(\hat{\boldsymbol{\theta}}_{\hat{\Lambda}} - \boldsymbol{\theta}_{\hat{\Lambda}})^T (\hat{\boldsymbol{\theta}}_{\hat{\Lambda}} - \boldsymbol{\theta}_{\hat{\Lambda}}) | \mathbf{x}; \Lambda]; \Lambda \right]. \quad (5.11)$$

By minimizing the inner expectation in (5.11) w.r.t $\hat{\boldsymbol{\theta}}_{\hat{\Lambda}}$, pointwise, for each $\mathbf{x} \in \Omega_{\mathbf{x}}$, similarly to the conventional MMSE derivations [2, pp. 55-56], and then substituting the result in (5.10) the cMMSE estimator is given by

$$\begin{cases} \hat{\boldsymbol{\theta}}_{\hat{\Lambda}}^{(\text{cMMSE})} = \text{E}[\boldsymbol{\theta}_{\hat{\Lambda}} | \mathbf{x}; \Lambda] \\ \hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c}^{(\text{cMMSE})} = \boldsymbol{\mu}_{\hat{\Lambda}^c} \end{cases}. \quad (5.12)$$

The cMMSE estimator is a coherent estimator. However, since it is a function of the unknown support set, Λ , it is not a practical estimator and can be used only as a benchmark.

5.1.3 Selected MMSE (sMMSE)

As a practical and coherent estimator, we consider the following sMMSE estimator:

$$\hat{\boldsymbol{\theta}}^{(\text{sMMSE})} = \sum_{k=1}^K \text{E}[\boldsymbol{\theta} | \mathbf{x}; \Lambda_k] \mathbb{1}_{\hat{\Lambda} = \Lambda_k}. \quad (5.13)$$

That is, the sMMSE estimator is the conditional mean w.r.t. the selected posterior pdf, $f(\boldsymbol{\theta} | \mathbf{x}; \Lambda_k)$ for $\hat{\Lambda} = \Lambda_k$, as defined in (5.1). By substituting (5.1) in (5.13) it can be verified that

$$\hat{\boldsymbol{\theta}}_{\hat{\Lambda}^c}^{(\text{sMMSE})} = \boldsymbol{\mu}_{\hat{\Lambda}^c}. \quad (5.14)$$

Hence, according to Definition 5.1 the sMMSE estimator is a coherent estimator. The sMMSE estimator has no optimality guarantees, since it is based on the selection, which may be wrong. Nevertheless, if the selection rule is consistent, i.e. if asymptotically it chooses the true model a.e. [150], then the sMMSE approaches a.e. the cMMSE estimator from (5.12).

5.1.4 Full MMSE (fMMSE)

As a naive but practical approach, one may assume the full support set of $\boldsymbol{\theta}$. This assumption leads to the fMMSE estimator, which is given by

$$\hat{\boldsymbol{\theta}}^{(\text{fMMSE})} = \text{E}[\boldsymbol{\theta}|\mathbf{x}; \{1, \dots, M\}]. \quad (5.15)$$

It can be verified that the fMMSE estimator is not a coherent estimator, since it ignores the selection.

5.2 Bayesian Post-model-selection bounds

Analytical computation of post-model-selection estimators and evaluation of their performance is usually intractable, especially in the considered setting, when the support set is unknown. Hence, derivation of lower bounds on the performance of post-model-selection estimators is crucial for performance analysis and system design. In this section, we derive three different Bayesian lower bounds on the sMSE of coherent estimators, as well as their modification to MSE lower bounds.

5.2.1 Coherent MMSE bound

In general, MSE lower bounds can be developed by evaluating the performance of estimators that use additional, side information that is not available to the MMSE estimator [41]. Similarly, the sMSE of the cMMSE estimator from (5.12) that uses the unknown support set can be used as a semi-oracle bound on the sMSE of estimators that do not use this side-information. That is, since the cMMSE estimator minimizes the sMSE among coherent estimators, its sMSE is a lower bound on the sMSE of any coherent estimator, $\hat{\boldsymbol{\theta}}$, i.e.

$$\begin{aligned} \text{sMSE}(\hat{\boldsymbol{\theta}}) &\geq \text{sMSE}(\hat{\boldsymbol{\theta}}^{(\text{cMMSE})}) \\ &= \text{Tr} \left(\text{E} \left[\mathbf{D}(\hat{\Lambda}) \left(\text{E}[\boldsymbol{\theta}\boldsymbol{\theta}^T|\mathbf{x}; \Lambda] - \text{E}[\boldsymbol{\theta}|\mathbf{x}; \Lambda]\text{E}[\boldsymbol{\theta}^T|\mathbf{x}; \Lambda] \right); \Lambda \right] \right), \end{aligned} \quad (5.16)$$

where $\mathbf{D}(\Lambda) \in \mathbb{R}^{M \times M}$ is a diagonal matrix with the elements:

$$[\mathbf{D}(\Lambda)]_{m,m} = \mathbb{1}_{m \in \Lambda}, \quad m = 1, \dots, M. \quad (5.17)$$

The last equality in (5.16) is obtained by substituting (5.12) in (5.11), since $\mathbf{D}(\hat{\Lambda})$ is only a function of \mathbf{x} . Similarly to the conventional MMSE estimator, often, a simple closed-form expression for the r.h.s. of (5.16) cannot be found. It is therefore useful to search for

more conservative lower bounds [165].

5.2.2 Selective Bayesian Cramér-Rao Bound (BCRB)

In this subsection, we present two Bayesian Cramér-Rao type lower bounds. The proofs for Theorem 5.1 and Theorem 5.2, as well as their achievability conditions, appear in Appendix D. We define the following regularity conditions:

C.1. The joint pdf, $f(\mathbf{x}, \boldsymbol{\theta}; \Lambda)$, is differentiable w.r.t. $\boldsymbol{\theta}$, $\forall \mathbf{x} \in \Omega_{\mathbf{x}}$, and the elements of $\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda) = \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{x}; \Lambda)$ are integrable functions w.r.t. $f(\mathbf{x}, \boldsymbol{\theta}; \Lambda)$.

C.2. The Bayesian Fisher information matrix (FIM) [157],

$$\mathbf{J}(\Lambda) \triangleq \mathbb{E}[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda); \Lambda], \quad (5.18)$$

which is parametrized by Λ , is a non-singular matrix.

C.3. The matrix

$$\mathbf{J}(\mathbf{x}, \Lambda) \triangleq \mathbb{E}[\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) \nabla_{\boldsymbol{\theta}}^T \log f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) | \mathbf{x}; \Lambda] \quad (5.19)$$

is a non-singular matrix, for a.e. $\mathbf{x} \in \Omega_{\mathbf{x}}$.

C.4. The posterior pdf satisfies $\boldsymbol{\theta} f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) \Big|_{\partial\Omega_{\boldsymbol{\theta}}} = \mathbf{0}$, a.e. $\mathbf{x} \in \Omega_{\mathbf{x}}$, where $\partial\Omega_{\boldsymbol{\theta}}$ refers to the boundary of $\Omega_{\boldsymbol{\theta}}$.

Theorem 5.1 (selective BCRB). *Under regularity conditions C.1, C.2 and C.4,*

$$sMSE(\hat{\boldsymbol{\theta}}) \geq sBCRB(\Lambda) \triangleq \text{Tr}(\mathbf{\Pi} \mathbf{J}^{-1} \mathbf{\Pi}), \quad (5.20)$$

for any coherent estimator, $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \rightarrow \mathbb{R}^M$, where $\mathbf{\Pi}$ is a diagonal matrix with the diagonal elements

$$[\mathbf{\Pi}]_{m,m} = [\mathbb{E}[\mathbf{D}(\hat{\Lambda})]]_{m,m} = \Pr(m \in \hat{\Lambda}). \quad (5.21)$$

The selective BCRB in (5.20) is achievable iff

$$\mathbf{D}(\hat{\Lambda})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{\Pi} \mathbf{J}^{-1}(\Lambda) \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{x}; \Lambda), \quad a.e.. \quad (5.22)$$

Proof. The proof for Theorem 5.1 appears in Appendix D.i. □

A tighter version of the selective BCRB is developed, in the following theorem, in a similar manner to the relation between the conventional BCRB and the tighter BCRB [157, 166].

Theorem 5.2 (selective Tighter BCRB (TBCRB)). *Under regularity conditions C.1, C.3 and C.4,*

$$sMSE(\hat{\boldsymbol{\theta}}) \geq sTBCRB(\Lambda) \triangleq \text{Tr} \left(\mathbb{E} \left[\mathbf{J}^{-1}(\mathbf{x}) \mathbf{D}(\hat{\Lambda}) \right] \right), \quad (5.23)$$

for any coherent estimator, $\hat{\boldsymbol{\theta}} : \Omega_{\mathbf{x}} \rightarrow \mathbb{R}^M$. The selective BCRB in (5.23) is achievable iff

$$\mathbf{D}(\hat{\Lambda})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{D}(\hat{\Lambda}) \mathbf{J}^{-1}(\mathbf{x}, \Lambda) \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta} | \mathbf{x}; \Lambda), \quad a.e.. \quad (5.24)$$

Proof. The proof for Theorem 5.2 appears in Appendix D.ii. □

Remark 5.1 (Order relation). *In Appendix D.iii we show that*

$$sTBCRB(\Lambda) \geq sBCRB(\Lambda). \quad (5.25)$$

However, it should be noted that the selective TBCRB from Theorem 5.2 may be intractable for problems in which the selective BCRB Theorem 5.1 is tractable. This is due to the fact that the Bayesian FIM in (5.18) is easier to evaluate than the one in (5.19), which is based on the a-posteriori pdf, $f(\boldsymbol{\theta} | \mathbf{x}; \Lambda)$.

Remark 5.2 (MSE lower bounds). *By using any lower bound on the sMSE, a lower bound on the MSE of any coherent estimator can be readily obtained as follows: Let B be a lower bound on the sMSE; by using the MSE decomposition from (5.4) and by the fact that the usMSE is identical for any coherent estimator, we obtain a lower bound on the MSE:*

$$\begin{aligned} MSE(\hat{\boldsymbol{\theta}}) &\geq B + usMSE(\hat{\boldsymbol{\theta}}) \\ &= B + \text{Tr} \left(\mathbb{E} \left[(\mathbf{I} - \mathbf{D}(\hat{\Lambda})) \mathbb{E}[(\boldsymbol{\mu} - \boldsymbol{\theta})(\boldsymbol{\mu} - \boldsymbol{\theta})^T | \mathbf{x}; \Lambda]; \Lambda \right] \right), \end{aligned} \quad (5.26)$$

where the last equality is obtained by substituting (5.7) and using the law of total expectation. Thus, the coherent MMSE bound, the selective BCRB, and the selective TBCRB, derived here, also imply MSE lower bounds that are obtained from substituting (5.16), (5.20) or (5.23) as B in (5.26).

Remark 5.3 (Oracle BCRB). *The oracle BCRB is a lower bound on the MSE, which is commonly used for post-model-selection performance analysis [167]. Under regularity conditions C.1, C.2 and C.4 the oracle BCRB is given by*

$$MSE(\hat{\boldsymbol{\theta}}) \geq \text{Tr} \left(\mathbf{J}^{-1}(\Lambda) \right), \quad (5.27)$$

for any estimator, $\hat{\boldsymbol{\theta}}$, where $\mathbf{J}(\Lambda)$ is defined in (5.18). The oracle BCRB is not a function of the selection rule, and thus, it is not a tight bound on the MSE of coherent estimators.

5.3 Simulations

In this section, we evaluate the performance of the estimators presented in Section 5.1 and the performance bounds from Section 5.2. We consider a linear Gaussian model:

$$\mathbf{x} = \mathbf{H}_\Lambda \boldsymbol{\theta}_\Lambda + \mathbf{w}, \quad (5.28)$$

where $\mathbf{H}_\Lambda \in \mathbb{R}^{N \times |\Lambda|}$ is a submatrix of a known matrix $\mathbf{H} \in \mathbb{R}^{N \times M}$, consisting of the columns indexed by Λ . The random vectors $\boldsymbol{\theta}$ and \mathbf{w} are mutually-independent zero-mean Gaussian vectors with non-singular covariance matrices, $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Sigma}_w$, respectively. This model can represent, for example, a sparse recovery problem of a random vector with a known measurement matrix. Solving model-selection procedures for the model in (5.28) is NP-hard problem. Thus, we assume the following element-wise energy-based selection rule:

$$m \in \hat{\Lambda} \text{ if } \left(\hat{\theta}_m^{(\text{fMMSE})} \right)^2 > \gamma_m, \quad \forall m = 1, \dots, M, \quad (5.29)$$

where γ_m are positive thresholds, and the fMMSE estimator from (5.15) for this setting is given by

$$\hat{\boldsymbol{\theta}}^{(\text{fMMSE})} = \mathbf{K} \boldsymbol{\Sigma}_w^{-1} \mathbf{x}, \quad (5.30)$$

in which $\mathbf{K} \triangleq (\mathbf{H}^T \boldsymbol{\Sigma}_w^{-1} \mathbf{H} + \boldsymbol{\Sigma}_\theta^{-1})^{-1} \mathbf{H}^T$. In the following we set $\gamma_m = F^{-1}(1-\alpha) [\mathbf{K} \boldsymbol{\Sigma}_w^{-1} \mathbf{K}^T]_{m,m}$ to obtain an element-wise Neyman-Pearson test [135] with false-alarm (FA) rate, α , where $F(\cdot)$ is the cumulative distribution function of a central χ^2 distribution with 1 degree of freedom. For this setting, the matrix from (5.19) coincides with the FIM from (5.18), given by

$$\mathbf{J}(\mathbf{x}, \Lambda) = \mathbf{J}(\Lambda) = \mathbf{D}(\Lambda) \mathbf{H}^T \boldsymbol{\Sigma}_w^{-1} \mathbf{H} \mathbf{D}(\Lambda) + \boldsymbol{\Sigma}_\theta^{-1}. \quad (5.31)$$

The oracle MMSE from (5.8) is given by

$$\hat{\boldsymbol{\theta}}^{(\text{oMMSE})} = \mathbf{J}^{-1}(\Lambda) \mathbf{D}(\Lambda) \mathbf{H}^T \boldsymbol{\Sigma}_w^{-1} \mathbf{x}, \quad (5.32)$$

the cMMSE estimator from (5.12) is given by

$$\hat{\boldsymbol{\theta}}^{(\text{cMMSE})} = \mathbf{D}(\hat{\Lambda}) \hat{\boldsymbol{\theta}}^{(\text{oMMSE})}, \quad (5.33)$$

and the sMMSE estimator from (5.13) is given by

$$\hat{\boldsymbol{\theta}}^{(\text{sMMSE})} = (\mathbf{D}(\hat{\Lambda}) \mathbf{H}^T \boldsymbol{\Sigma}_w^{-1} \mathbf{H} \mathbf{D}(\hat{\Lambda}) + \boldsymbol{\Sigma}_\theta^{-1})^{-1} \mathbf{D}(\hat{\Lambda}) \mathbf{H}^T \boldsymbol{\Sigma}_w^{-1} \mathbf{x}. \quad (5.34)$$

Since for the Gaussian case, the covariance matrix of $\boldsymbol{\theta}$ given \mathbf{x} is $\mathbf{J}^{-1}(\Lambda)$, the cMMSE bound from (5.16) is given by

$$\text{sMSE}(\hat{\boldsymbol{\theta}}^{(\text{cMMSE})}) = \text{Tr} \left(\mathbf{E} \left[\mathbf{J}^{-1}(\Lambda) \mathbf{D}(\hat{\Lambda}) \right] \right), \quad (5.35)$$

and therefore, coincides with the selective TBCRB from (5.23). Notice that

$$\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) = \mathbf{J}(\Lambda)(\hat{\boldsymbol{\theta}}^{(\text{oMMSE})} - \boldsymbol{\theta}), \quad (5.36)$$

so that the cMMSE estimator satisfies the achievability condition from (5.24), which is consistent with (5.35). The elements of the probabilities matrix from (5.21) are

$$[\boldsymbol{\Pi}]_{m,m} \triangleq 1 - F \left(\frac{\gamma_m}{[\mathbf{K}\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}(\mathbf{H}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{H}^T + \boldsymbol{\Sigma}_{\mathbf{w}})\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{K}^T]_{m,m}} \right). \quad (5.37)$$

Therefore, the selective BCRB from (5.20) is given by

$$\text{sTBCRB}(\hat{\boldsymbol{\theta}}) = \sum_{m=1}^M [\boldsymbol{\Pi}]_{m,m}^2 [\mathbf{J}^{-1}(\Lambda)]_{m,m}, \quad (5.38)$$

and the selective TBCRB from (5.23) is given by

$$\text{sTBCRB}(\hat{\boldsymbol{\theta}}) = \sum_{m=1}^M [\boldsymbol{\Pi}]_{m,m} [\mathbf{J}^{-1}(\Lambda)]_{m,m}. \quad (5.39)$$

In the following simulations, the matrix \mathbf{H} consists of a normalized 32×32 Hadamard matrix concatenated with a random 32×32 matrix, whose entries are drawn from a standard Gaussian distribution. The covariance matrices of the prior distribution and the noise are set to $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \mathbf{I}$ and $\boldsymbol{\Sigma}_{\mathbf{w}} = \sigma^2 \mathbf{I}$. In Fig. 5.1 the MSE of the estimators from (5.30) and (5.32)–(5.34) are compared with the oracle BCRB from (5.27), as well as the MSE version of the selective BCRB and the selective TBCRB obtained by substituting (5.38) and (5.39) in (5.26). For this case $\Lambda = \{1, \dots, 8\}$ and $\alpha = 0.01$. It can be seen that the selective bounds are much tighter than the oracle BCRB and that the selective TBCRB achieves the performance of the cMMSE estimator, which is the optimal coherent estimator. In Fig. 5.2 the sMSE of the cMMSE and sMMSE estimators are compared with the selective BCRB and the selective TBCRB from (5.38) and (5.39) versus the FA rate, α , for $\sigma^2 = 1$. The oracle BCRB is omitted from this figure since it is a bound on the MSE and not on the sMSE. It can be seen that the selective TBCRB is tighter than the selective BCRB, and is close to the performance of the practical sMMSE estimator.

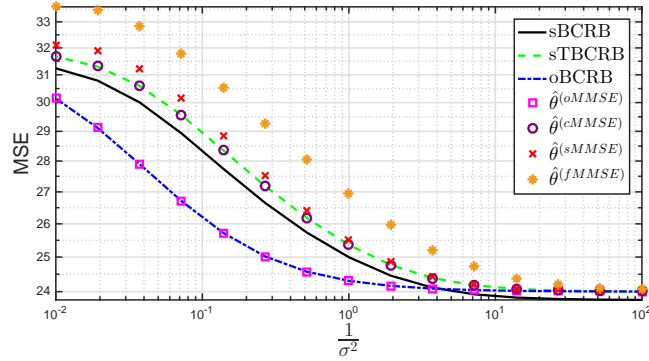


Figure 5.1: The MSE of the oracle MMSE, cMMSE, sMMSE, and fMMSE estimators compared to the oracle BCRB, selective BCRB, and selective TBCRB, versus $\frac{1}{\sigma^2}$.

Thus, it can be used to evaluate the performance of practical estimators and for system design.

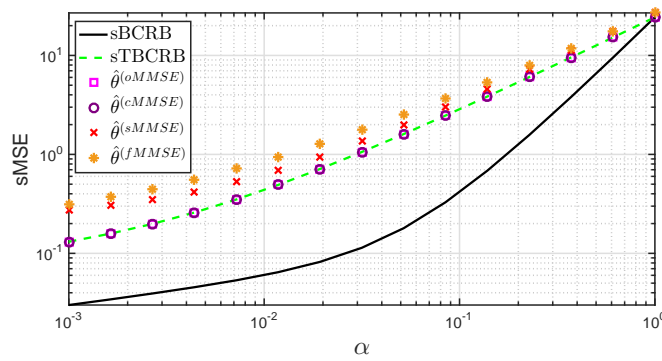


Figure 5.2: The sMSE of the cMMSE and sMMSE estimators compared with the selective BCRB and the selective TBCRB versus the FA rate, α .

Chapter 6

Summary and Future Research

This chapter summarizes this dissertation and discusses possible future research topics. In this work, we examine estimation problems that follows a preliminary selection procedure. In particular, we focus on two fundamental problems. Estimation after parameter selection and estimation after model selection. In the former, which the observation model is known and the parameters of interest to be estimated are selected based on the observed data. In the later, prior to the estimation stage, a model selection procedure is applied to select a model from a set of candidate models. Then, the parameters of the selected model are estimated. The significance of this research lies in the creation of a novel estimation framework tailored to account for the selection process. Specifically, we analyses the affect of the selection procedure on the following estimation. We introduce theoretical and practical estimators and streamlined algorithms for their practical implementation. Furthermore, we derive appropriate performance bounds that consider the selection procedure for performance analysis.

In Chapter 2, we derived practical algorithms for low-complexity estimation after parameter selection, where the selection of the parameter of interest is data-based and predetermined. We established a detailed model and theoretical results, including unbiasedness and the properties of the PSML estimator. Subsequently, we show that employing the MBP algorithm can converge to the PSML estimator. We incorporated the MBP-PSML estimator into two methods, namely the second-best PSML and SA-PSML, that avoid the need for calculating the probability of selection. Furthermore, we present the empirical Ψ -CRB and derive an efficient algorithm, which can be used as a low-complexity performance analysis tool. The proposed methods were implemented in several applications, and the results demonstrated substantial enhancements in terms of Ψ -bias and PSMSE compared to existing methods.

In Chapter 3, we discussed estimation methods for models with a preliminary data-

based selection stage, which selects the subset of parameters of interest. Our approach assumes coherent estimators that constrain the unselected parameters to zero, facilitating estimation in scenarios where the complete parameter vector is unidentifiable without a selection process. We introduced the SA-cPSML algorithm, designed to implement the coherent PSML estimator in practice. Simulations show that the proposed estimator, implemented by the SA-cPSML algorithm, outperforms existing methods.

In Chapter 4, we explored the paradigm of non-Bayesian estimation following model selection, framing this issue as estimation under model misspecification. We showed that the conventional approach of interpreting this scenario naively may yield an invalid analysis. At the same time, the natural pdf-corrected normalized interpretation, though valid, introduces an incoherent coupling among the parameters of various candidate models. Consequently, we advocated the adoption of a selective inference interpretation, which utilizes conditional likelihoods conditioned on the selection of a specific model. This selective approach is both valid and coherent, boasting desirable properties. We derived the corresponding MML estimators based on the three interpretations: the MSL, MSNL, and PSML estimators. We proposed the PS-MCRB, a novel lower bound on any post-model-selection unbiased estimator that incorporates the misspecification in the form of a model-selection procedure. We derived the PS-MCRBs for the three interpretations and analyzed their properties. Through simulations, we illustrated the relationships between the proposed estimators and bounds. Specifically, we demonstrated that in terms of MSE, the PSML and MSNL estimators, aligned with the normalized and selective inference interpretations, respectively, surpass the commonly used MSL estimator associated with the naive interpretation. We also established that the proposed PS-MCRBs under different interpretations offer more insightful information than the oracle CRB, with the selective inference interpretation yielding the most favourable bound and MSE.

In Chapter 5, we considered a Bayesian estimation framework following model selection, employing a predefined model-selection criterion. Within this context, we examined Bayesian coherent estimation after model selection, wherein estimators of the unselected parameters are assigned as their prior mean. We introduced diverse types of estimators, categorized as coherent/non-coherent and practical/theoretical. Additionally, we derived three Bayesian bounds on the sMSE of any coherent estimator: the coherent MMSE bound, the selective BCRB, and the selective TBCRB. Through simulations, we compared the presented estimators against these proposed bounds. The results demonstrated that these bounds are more stringent compared to the oracle BCRB, which neglects the model-selection stage, thereby serving as valuable tools for practical performance assessment.

Future research may include extensions of the non-Bayesian performance bound for estimation after parameter selection and estimation after model selection to large-error MSE lower bounds, such as the Barankin and Hammersley-Chapman-Robbins bounds. In addition, in the Bayesian framework different performance bounds may be considered, such as bounds from the Ziv-Zakai class.

Furthermore, future work could develop low-complexity methods for practical implementation of the proposed estimators for estimation after model selection where an analytical solution is intractable. We note that a pre-processing stage is a common practice in many signal processing and statistical inference applications. The presented framework can be generalized for general pre-processing procedures that are not necessarily a selection procedure. Generalization for the suggested estimation methods and performance analysis can be developed for cases where the estimation stage follows a predetermined data-based pre-processing procedure. This generalization can include pre-processing procedures that are implemented by neural networks or other learning, data-driven techniques.

An inherent challenge that frequently arises when employing post-selection estimation methods and establishing bounds is the presence of certain elements within the procedure that are often intractable. This difficulty becomes particularly pronounced when dealing with high-dimensional and complex problems. Recently, there has been a growing interest regarding approaches that integrate data-based solutions, i.e. learning techniques in model-based frameworks [168]. A similar strategy has been applied in the context of implementing performance bounds [169]. Hence, another direction for future research may explore, in a similar manner, custom-made data-based solutions that can be integrated into this framework in order to provide calculations that are intractable analytically.

In addition, the context of post-selection estimation and estimation under model misspecification can be generalized by considering different cost functions from the MSE, such as periodic or cyclic error [8, 95, 170].

Appendix A

A.i Existence of Ψ -unbiased estimator

In this appendix we show that for the two-stage model, which satisfies (2.60), if a mean-unbiased estimator of $\boldsymbol{\theta}$ based only on the second-stage observation vector, \mathbf{y} , exists, and under the assumption that $\Omega_{\mathbf{y}} \neq \emptyset$, then, for any selection rule, there exists an Ψ -unbiased estimator of $\boldsymbol{\theta}$.

Lemma A.1. *Let $\hat{\boldsymbol{\theta}}(\mathbf{y})$ be an estimator of $\boldsymbol{\theta}$ based only on the observation set \mathbf{y} . Assuming that $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is a mean-unbiased estimator, i.e.*

$$\mathbb{E}_{\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}] = \mathbf{0}, \quad (\text{A.1})$$

then $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is an Ψ -unbiased estimator for the two-stage model.

Proof. The mean unbiasedness in (A.1) implies that

$$\int_{\Omega_{\mathbf{y}}^{(m)}} (\hat{\theta}_m - \theta_m) f_m(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} = 0, \quad \forall m = 1, \dots, M. \quad (\text{A.2})$$

Therefore, for all $m = 1, \dots, M$ that $\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) \neq 0$, we obtain that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[\hat{\theta}_m - \theta_m | \Psi_{\mathbf{x}} = m] &= \int_{\mathcal{A}_m} \int_{\Omega_{\mathbf{y}}^{(m)}} (\hat{\theta}_m - \theta_m) f(\mathbf{x}, \mathbf{y} | \Psi_{\mathbf{x}} = m; \boldsymbol{\theta}) d\mathbf{y} d\mathbf{x} \\ &= \int_{\mathcal{A}_m} \frac{f(\mathbf{x}; \boldsymbol{\theta})}{\Pr(\Psi_{\mathbf{x}} = m; \boldsymbol{\theta})} \int_{\Omega_{\mathbf{y}}^{(m)}} (\hat{\theta}_m - \theta_m) f_m(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} d\mathbf{x} \\ &= 0, \end{aligned} \quad (\text{A.3})$$

where the first equality in (A.3) is obtained by substituting (2.62), the second equality is obtained by substituting (2.60), and the last equality is obtained by substituting (A.2). Therefore, the Ψ -unbiasedness condition from (2.6) holds and $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is an Ψ -unbiased estimator. \square

Appendix B

B.i Proof of Proposition 4.1

By using Definition 1.1 on the considered setting, an estimator $\hat{\boldsymbol{\theta}}$ is said to be an unbiased estimator of $\boldsymbol{\theta}$ in the Lehmann sense w.r.t. the MSSE from (4.5) if

$$\mathbb{E}_p[\mathcal{C}_{\text{MSSE}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\vartheta})] \preceq \mathbb{E}_p[\mathcal{C}_{\text{MSSE}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\eta})], \quad \forall \boldsymbol{\vartheta}, \boldsymbol{\eta} \in \Omega_{\boldsymbol{\theta}}, \quad (\text{B.1})$$

Notice that, by Definition 4.1, the pseudo-true vector, $\hat{\boldsymbol{\theta}}$, is determined by the true pdf, $p(\mathbf{x})$, and the assumed pdf, $f(\mathbf{x}; \cdot)$. By substituting the MSSE from (4.5) in (B.1) we can notice that

$$\begin{aligned} \mathbb{E}_p[\mathcal{C}_{\text{MSSE}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\eta})] &= \mathbb{E}_p[(\hat{\boldsymbol{\theta}} - \boldsymbol{\eta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\eta})^T] = \mathbb{E}_p[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta} + \boldsymbol{\vartheta} - \boldsymbol{\eta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta} + \boldsymbol{\vartheta} - \boldsymbol{\eta})^T] \\ &= \mathbb{E}_p[\mathcal{C}_{\text{MSSE}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\vartheta})] + \mathbb{E}_p[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})(\boldsymbol{\vartheta} - \boldsymbol{\eta})^T] + (\boldsymbol{\vartheta} - \boldsymbol{\eta})\mathbb{E}_p^T[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})] + (\boldsymbol{\vartheta} - \boldsymbol{\eta})(\boldsymbol{\vartheta} - \boldsymbol{\eta})^T. \end{aligned} \quad (\text{B.2})$$

Therefore, by substituting (B.2) in (B.1), we obtain that the condition in (B.1) can be written as

$$\mathbb{E}_p[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})(\boldsymbol{\eta} - \boldsymbol{\vartheta})^T] + (\boldsymbol{\eta} - \boldsymbol{\vartheta})\mathbb{E}_p^T[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})] \preceq (\boldsymbol{\vartheta} - \boldsymbol{\eta})(\boldsymbol{\vartheta} - \boldsymbol{\eta})^T. \quad (\text{B.3})$$

One can notice that (4.7) is a sufficient condition to satisfy the inequality in (B.3). In the following, we show that it is also a necessary condition. The inequality should hold for any $\boldsymbol{\eta} \in \Omega_{\boldsymbol{\theta}}$. In particular, for $\boldsymbol{\eta} = \boldsymbol{\vartheta} + \epsilon \mathbf{e}_m$, where \mathbf{e}_m is the standard m th unit vector for some $m \in \{1, \dots, M\}$ and $\epsilon > 0 \in \mathbb{R}$, sufficiently small that $\boldsymbol{\vartheta} + \epsilon \mathbf{e}_m \in \Omega_{\boldsymbol{\theta}}$. Substitution of such $\boldsymbol{\eta}$ in (B.3) results in

$$\mathbb{E}_p[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})]\epsilon \mathbf{e}_m^T + \epsilon \mathbf{e}_m \mathbb{E}_p^T[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta})] \preceq \epsilon^2 \mathbf{e}_m \mathbf{e}_m^T. \quad (\text{B.4})$$

We notice that a necessary condition for the matrix inequality in (B.4) is that

$$2\epsilon\mathbb{E}_p[(\hat{\theta}_m - \vartheta_m)] \leq \epsilon^2. \quad (\text{B.5})$$

Since ϵ can be infinitesimally small, and since the index m is arbitrary and (B.5) should be satisfied for any $m \in \{1, \dots, M\}$, then (4.7) is also a necessary condition to satisfy the inequality in (B.3), which concludes this proof.

B.ii Proof of Theorem 4.1

To prove Theorem 4.1, we propose the following lemma that considers a parametric true model described in the following. Let $(\Omega_{\mathbf{x}}, \mathcal{F}, P_{\gamma})$ denote a probability space, where $\Omega_{\mathbf{x}}$ is the observation space, \mathcal{F} is the σ -algebra, and P_{γ} is a probability measure on \mathcal{F} that is parameterized by a real deterministic parameter vector $\gamma \in \Omega_{\theta}$. We assume that the pdf w.r.t. P_{γ} exists and is denoted by $\tilde{p}(\cdot; \gamma)$. Let $\mathbf{x} \in \Omega_{\mathbf{x}}$ be a random observation vector, which is distributed according to P_{γ} , and let $\hat{\boldsymbol{\theta}}$ be an MS-unbiased estimator of $\boldsymbol{\vartheta}_{\gamma}$, which is the pseudo-true vector according to (4.2) w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$ as the true pdf. In addition, we assume the following regularity condition:

RC.4. For any $\gamma \in \Omega_{\theta}$ there is a neighborhood, \mathcal{N}_{γ} , such that for any $\gamma' \in \mathcal{N}_{\gamma}$

$$\left(\frac{1}{\tilde{p}(\mathbf{x}; \gamma)} \left| \frac{\partial \tilde{p}(\mathbf{x}; \gamma)}{\partial \gamma_i} \right| \Big|_{\gamma=\gamma'} \right) \leq m(\mathbf{x}), \quad (\text{B.6})$$

where $m(\mathbf{x})$ is a square-integrable function w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$.

We define:

$$\mathbf{J}_{\tilde{p}}(\gamma) \triangleq \mathbb{E}_{\tilde{p}} \left[\nabla_{\gamma} \log \tilde{p}(\mathbf{x}; \gamma) \nabla_{\gamma}^T \log \tilde{p}(\mathbf{x}; \gamma) \right], \quad (\text{B.7})$$

which is the conventional FIM w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$. In addition, we define

$$\mathbf{B}_{\tilde{p},f}(\boldsymbol{\theta}; \gamma) \triangleq \mathbb{E}_{\tilde{p}} [\nabla_{\gamma} \log \tilde{p}(\mathbf{x}; \gamma) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}; \boldsymbol{\theta})]. \quad (\text{B.8})$$

The following lemma is equivalent to [87, Th. 3.1].

Lemma B.1. *A general Parametric Bound*

Let $f(\mathbf{x}; \boldsymbol{\theta})$ be an assumed model that satisfies regularity conditions RC.1–RC.3 where the true model is $\tilde{p}(\mathbf{x}; \gamma)$ that satisfies regularity condition RC.4. The MSMSE from (4.6)

of any MS-unbiased estimator, $\hat{\boldsymbol{\theta}}$, is bounded by the following bound:

$$\mathbb{E}_{\tilde{p}} \left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T \right] \succeq \mathbf{A}_{\tilde{p}}^{-1}(\boldsymbol{\vartheta}_\gamma, \gamma) \mathbf{B}_{\tilde{p},f}^T(\boldsymbol{\vartheta}_\gamma; \gamma) \mathbf{J}_{\tilde{p}}^{-1}(\gamma) \mathbf{B}_{\tilde{p},f}(\boldsymbol{\vartheta}_\gamma; \gamma) \mathbf{A}_{\tilde{p}}^{-1}(\boldsymbol{\vartheta}_\gamma, \gamma), \quad (\text{B.9})$$

where $\mathbf{A}_{\tilde{p}}(\boldsymbol{\vartheta}_\gamma, \gamma)$ is defined in (4.9), only here, the expectation is w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$ and hence the dependency in γ . Notice that both sides of the inequality in (B.9) are functions of γ and the inequality is satisfied for any $\gamma \in \Omega_\theta$.

Proof of Lemma B.1. From the covariance form of the Cauchy-Schwarz inequality [157, pp. 33] we obtain

$$\begin{aligned} \mathbb{E}_{\tilde{p}} \left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T \right] \\ \succeq \mathbb{E}_{\tilde{p}} \left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma) \nabla_\gamma^T \log \tilde{p}(\mathbf{x}; \gamma) \right] \mathbf{J}_{\tilde{p}}^{-1}(\gamma) \mathbb{E}_{\tilde{p}} \left[\nabla_\gamma \log \tilde{p}(\mathbf{x}; \gamma) (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T \right]. \end{aligned} \quad (\text{B.10})$$

Since the estimator, $\hat{\boldsymbol{\theta}}$, is assumed to a MS-unbiased estimator of $\boldsymbol{\vartheta}_\gamma$, then

$$\mathbb{E}_{\tilde{p}}[\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma] = \int_{\Omega_{\mathbf{x}}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma) \tilde{p}(\mathbf{x}; \gamma) d\mathbf{x} = \mathbf{0}. \quad (\text{B.11})$$

By applying the derivative w.r.t. γ on both sides of (B.11), we obtain that

$$\nabla_\gamma \int_{\Omega_{\mathbf{x}}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T \tilde{p}(\mathbf{x}; \gamma) d\mathbf{x} = \mathbf{0}. \quad (\text{B.12})$$

Under regularity condition RC.4 we can replace the order of the derivative and the integration on the l.h.s. of (B.12) to obtain

$$\nabla_\gamma \int_{\Omega_{\mathbf{x}}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T \tilde{p}(\mathbf{x}; \gamma) d\mathbf{x} = \int_{\Omega_{\mathbf{x}}} \nabla_\gamma \{ (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T \tilde{p}(\mathbf{x}; \gamma) \} d\mathbf{x} = \mathbf{0}. \quad (\text{B.13})$$

By using the multiplication rule, we obtain that (B.13) can be rewritten as

$$\begin{aligned} \int_{\Omega_{\mathbf{x}}} \nabla_\gamma \{ (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T \tilde{p}(\mathbf{x}; \gamma) \} d\mathbf{x} &= \int_{\Omega_{\mathbf{x}}} \nabla_\gamma \tilde{p}(\mathbf{x}; \gamma) (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T d\mathbf{x} - \nabla_\gamma \boldsymbol{\vartheta}_\gamma^T \int_{\Omega_{\mathbf{x}}} \tilde{p}(\mathbf{x}; \gamma) d\mathbf{x} \\ &= \int_{\Omega_{\mathbf{x}}} \nabla_\gamma \log \tilde{p}(\mathbf{x}; \gamma) (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T \tilde{p}(\mathbf{x}; \gamma) d\mathbf{x} - \nabla_\gamma \boldsymbol{\vartheta}_\gamma^T. \end{aligned} \quad (\text{B.14})$$

We can notice that

$$\begin{aligned} \int_{\Omega_{\mathbf{x}}} \nabla_\gamma \tilde{p}(\mathbf{x}; \gamma) (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T d\mathbf{x} &= \int_{\Omega_{\mathbf{x}}} \nabla_\gamma \log \tilde{p}(\mathbf{x}; \gamma) (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T \tilde{p}(\mathbf{x}; \gamma) d\mathbf{x} \\ &= \mathbb{E}_{\tilde{p}}[\nabla_\gamma \log \tilde{p}(\mathbf{x}; \gamma) (\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_\gamma)^T]. \end{aligned} \quad (\text{B.15})$$

By substituting (B.15) in (B.14), then (B.12) implies that

$$\mathbb{E}_{\tilde{p}}[\nabla_{\gamma} \log \tilde{p}(\mathbf{x}; \gamma)(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_{\gamma})^T] = \nabla_{\gamma} \boldsymbol{\vartheta}_{\gamma}^T. \quad (\text{B.16})$$

Substitution of (B.16) in (B.10) results in

$$\mathbb{E}_{\tilde{p}} [(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_{\gamma})(\hat{\boldsymbol{\theta}} - \boldsymbol{\vartheta}_{\gamma})^T] \succeq \nabla_{\gamma}^T \boldsymbol{\vartheta}_{\gamma} \mathbf{J}_{\tilde{p}}^{-1}(\gamma) \nabla_{\gamma} \boldsymbol{\vartheta}_{\gamma}^T. \quad (\text{B.17})$$

We define the following notation:

$$\boldsymbol{\zeta}(\boldsymbol{\theta}, \gamma) \triangleq \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\tilde{p}}[\log f(\mathbf{x}; \boldsymbol{\theta})], \quad (\text{B.18})$$

where the dependency on γ stems from the expectation w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$. Since $\boldsymbol{\vartheta}_{\gamma}$ is the pseudo-true vector, according to (4.4) and under regularity condition RC.1

$$\boldsymbol{\zeta}(\boldsymbol{\theta}, \gamma) \Big|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}_{\gamma}} = \mathbf{0}. \quad (\text{B.19})$$

Recall that $\boldsymbol{\vartheta}_{\gamma}$ is a function of γ , hence, applying the gradient w.r.t. γ in (B.19) using the chain rule results in

$$\nabla_{\gamma} \{\boldsymbol{\zeta}^T(\boldsymbol{\vartheta}_{\gamma}, \gamma)\} = \nabla_{\gamma} \boldsymbol{\zeta}^T(\boldsymbol{\vartheta}_{\gamma}, \gamma) + \nabla_{\gamma} \boldsymbol{\vartheta}_{\gamma}^T \nabla_{\boldsymbol{\theta}} \boldsymbol{\zeta}(\boldsymbol{\vartheta}_{\gamma}, \gamma) = \mathbf{0}. \quad (\text{B.20})$$

Since the assumed pdf $f(\mathbf{x}; \boldsymbol{\theta})$ satisfies regularity condition RC.2 w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$, we can differentiate (B.18) under the integral sign. Therefore, we can notice that

$$\begin{aligned} \nabla_{\gamma} \boldsymbol{\zeta}^T(\boldsymbol{\theta}, \gamma) &= \int_{\Omega_{\mathbf{x}}} \nabla_{\gamma} \tilde{p}(\mathbf{x}; \gamma) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \\ &= \mathbb{E}_{\tilde{p}}[\nabla_{\gamma} \log \tilde{p}(\mathbf{x}; \gamma) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}; \boldsymbol{\theta})] = \mathbf{B}_{\tilde{p},f}(\boldsymbol{\theta}; \gamma). \end{aligned} \quad (\text{B.21})$$

In a similar manner, we can notice that

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{\zeta}^T(\boldsymbol{\theta}, \gamma) = \mathbf{A}_{\tilde{p}}(\boldsymbol{\theta}; \gamma). \quad (\text{B.22})$$

Thus, by substituting (B.21) and (B.22) in (B.20) we obtain that

$$\nabla_{\gamma} \{\boldsymbol{\zeta}^T(\boldsymbol{\vartheta}_{\gamma}, \gamma)\} = \mathbf{B}_{\tilde{p},f}(\boldsymbol{\vartheta}_{\gamma}; \gamma) + \nabla_{\gamma} \boldsymbol{\vartheta}_{\gamma}^T \mathbf{A}_{\tilde{p}}(\boldsymbol{\vartheta}_{\gamma}; \gamma) = \mathbf{0}. \quad (\text{B.23})$$

Thus, under the assumption that $\mathbf{A}_{\tilde{p}}(\boldsymbol{\vartheta}_\gamma; \gamma)$ is not a singular matrix,

$$\nabla_\gamma \boldsymbol{\vartheta}_\gamma^T = -\mathbf{B}_{\tilde{p},f}(\boldsymbol{\vartheta}_\gamma; \gamma) \mathbf{A}_{\tilde{p}}^{-1}(\boldsymbol{\vartheta}_\gamma; \gamma). \quad (\text{B.24})$$

Substitution of (B.24) in (B.16) and then in (B.10) results in (B.9), which concludes the proof of Lemma B.1. \square

In the following, we define a specific form for $\tilde{p}(\mathbf{x}; \gamma)$ that binds the true pdf, $p(\mathbf{x})$, and the assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, from Section 4.1. First, we will show that if the assumed pdf satisfies regularity conditions RC.1–RC.3 w.r.t. $p(\mathbf{x})$ it is also satisfied w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$ as Lemma B.1 requires i.e. the regularity conditions of Theorem 4.1 are sufficient for Lemma B.1. Then we use Lemma B.1 to prove Theorem 4.1.

Let us define the following pdf:

$$\tilde{p}(\mathbf{x}; \gamma) = \frac{1}{c(\gamma)} \left(1 + \exp \left(1 - \frac{f(\mathbf{x}; \gamma)}{f(\mathbf{x}; \boldsymbol{\vartheta})} \right) \right) p(\mathbf{x}), \quad (\text{B.25})$$

where

$$c(\gamma) \triangleq \int_{\Omega_{\mathbf{x}}} \left(1 + \exp \left(1 - \frac{f(\mathbf{x}; \gamma)}{f(\mathbf{x}; \boldsymbol{\vartheta})} \right) \right) p(\mathbf{x}) d\mathbf{x} \quad (\text{B.26})$$

is a normalization factor, and here, $\boldsymbol{\vartheta}$ is the pseudo-true vector w.r.t. $p(\mathbf{x})$. We can notice that for $\gamma = \boldsymbol{\vartheta}$ we obtain

$$\tilde{p}(\mathbf{x}; \boldsymbol{\vartheta}) = p(\mathbf{x}). \quad (\text{B.27})$$

Lemma B.2. *If the assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, satisfies regularity conditions RC.1–RC.3 w.r.t. $p(\mathbf{x})$ as required in Theorem 4.1, it also satisfies all the assumptions for Lemma B.1. In particular, the assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, is strictly positive for any measurable set in $\Omega_{\mathbf{x}}$ w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$. In addition, regularity conditions RC.1–RC.3 and RC.4 are satisfied w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$.*

Proof. First, let us assume that there is a set of observation vectors $\mathcal{D} \subseteq \Omega_{\mathbf{x}}$, that is measurable w.r.t. $\tilde{p}(\mathbf{x}; \gamma)$ but has a zero measure w.r.t. $p(\mathbf{x})$. We notice that

$$1 \leq \left(1 + \exp \left(1 - \frac{f(\mathbf{x}; \gamma)}{f(\mathbf{x}; \boldsymbol{\vartheta})} \right) \right) \leq 1 + e, \quad (\text{B.28})$$

and by integrating (B.28) w.r.t. $p(\mathbf{x})$ we obtain

$$1 \leq c(\gamma) \leq 1 + e. \quad (\text{B.29})$$

Therefore,

$$\int_{\mathcal{D}} \tilde{p}(\mathbf{x}; \boldsymbol{\gamma}) d\mathbf{x} \geq (1 + e) \int_{\mathcal{D}} p(\mathbf{x}) d\mathbf{x} = 0, \quad (\text{B.30})$$

which contradicts the assumption that \mathcal{D} is a measurable set. This implies that any measurable set w.r.t. $\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})$ is also measurable w.r.t. $p(\mathbf{x})$. Therefore, the assumption that the assumed pdf, $f(\mathbf{x}; \boldsymbol{\theta})$, is strictly positive w.r.t. $p(\mathbf{x})$ also applies w.r.t. $\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})$.

Let $m(\mathbf{x})$ is square-integrable function w.r.t. $p(\mathbf{x})$. Then by using (B.28) and (B.29) we can verify that

$$m^2(\mathbf{x})\tilde{p}(\mathbf{x}; \boldsymbol{\gamma}) \leq (1 + e)m^2(\mathbf{x})p(\mathbf{x}). \quad (\text{B.31})$$

Therefore, if $m(\mathbf{x})$ is square-integrable function w.r.t. $p(\mathbf{x})$ it is also square-integrable function w.r.t. $\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})$. This concludes that regularity conditions RC.2 and RC.3 w.r.t. $p(\mathbf{x})$ implies the same regularity condition w.r.t. $\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})$.

We assume that $f(\mathbf{x}; \boldsymbol{\theta})$ is twice continuously differentiable w.r.t. $\boldsymbol{\theta}$, therefore, regularity condition RC.2 w.r.t. $p(\mathbf{x})$ implies that there is a neighborhood of the pseudo-true parameter vector, $\boldsymbol{\vartheta}$, such that $\mathbf{A}(\boldsymbol{\theta})$, which is the Hessian matrix of $E_p[\log f(\mathbf{x}; \boldsymbol{\theta})]$, is a negative definite matrix. Since we show that regularity condition RC.2 applies w.r.t. $\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})$ then $\mathbf{A}_{\tilde{p}}(\boldsymbol{\theta}; \boldsymbol{\gamma})$ which is the Hessian matrix of $E_{\tilde{p}}[\log f(\mathbf{x}; \boldsymbol{\theta})]$ is continuous. Since $\tilde{p}(\mathbf{x}; \boldsymbol{\vartheta}) = p(\mathbf{x})$ then $\mathbf{A}_{\tilde{p}}(\boldsymbol{\theta}; \boldsymbol{\vartheta}) = \mathbf{A}(\boldsymbol{\theta})$. Hence, there is a neighborhood of $\boldsymbol{\vartheta}$ such that $\mathbf{A}_{\tilde{p}}(\boldsymbol{\theta}; \boldsymbol{\gamma})$ is also a negative definite matrix and $E_{\tilde{p}}[\log f(\mathbf{x}; \boldsymbol{\theta})]$ is concave w.r.t. $\boldsymbol{\theta}$. Therefore, there is a neighborhood of $\boldsymbol{\vartheta}$ such that regularity condition RC.1 also applies w.r.t. $\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})$. Finally, we will show that the additional regularity condition RC.4 is guaranteed. We notice that

$$\frac{1}{\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})} \nabla_{\boldsymbol{\gamma}} \tilde{p}(\mathbf{x}; \boldsymbol{\gamma}') = \frac{\tilde{p}(\mathbf{x}; \boldsymbol{\gamma}')}{\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})} \nabla_{\boldsymbol{\gamma}} \log \tilde{p}(\mathbf{x}; \boldsymbol{\gamma}'). \quad (\text{B.32})$$

By substituting (B.25) and using (B.28) and (B.29) we obtain:

$$\frac{\tilde{p}(\mathbf{x}; \boldsymbol{\gamma}')}{\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})} = \frac{c(\boldsymbol{\gamma}')}{c(\boldsymbol{\gamma})} \frac{1 + \exp\left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma}')}{f(\mathbf{x}; \boldsymbol{\vartheta})}\right)}{1 + \exp\left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma})}{f(\mathbf{x}; \boldsymbol{\vartheta})}\right)} \leq (1 + e)^2. \quad (\text{B.33})$$

Therefore, to guarantee regularity condition RC.4, it is sufficient to show that the derivative of $\log \tilde{p}(\mathbf{x}; \boldsymbol{\gamma})$ is bounded.

$$\nabla_{\boldsymbol{\gamma}} \log \tilde{p}(\mathbf{x}; \boldsymbol{\gamma}) = \nabla_{\boldsymbol{\gamma}} \log \left(1 + \exp \left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma})}{f(\mathbf{x}; \boldsymbol{\vartheta})} \right) \right) - \nabla_{\boldsymbol{\gamma}} \log c(\boldsymbol{\gamma}). \quad (\text{B.34})$$

From (B.29), under the assumption that $f(\mathbf{x}; \boldsymbol{\theta})$ satisfies regularity condition RC.3, one

can verify that the i th element of the vector is

$$\exp\left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma})}{f(\mathbf{x}; \boldsymbol{\vartheta})}\right) \frac{1}{f(\mathbf{x}; \boldsymbol{\vartheta})} [\nabla_{\boldsymbol{\gamma}} f(\mathbf{x}; \boldsymbol{\gamma})]_i \leq em(\mathbf{x}). \quad (\text{B.35})$$

Therefore, according to Lebesgue's dominated convergence theorem, the gradient of $\log c(\boldsymbol{\gamma})$ is given by

$$\begin{aligned} \nabla_{\boldsymbol{\gamma}} \log c(\boldsymbol{\gamma}) &= \frac{1}{c(\boldsymbol{\gamma})} \nabla_{\boldsymbol{\gamma}} c(\boldsymbol{\gamma}) = \frac{1}{c(\boldsymbol{\gamma})} \int_{\Omega_{\mathbf{x}}} \nabla_{\boldsymbol{\gamma}} \left(1 + \exp\left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma})}{f(\mathbf{x}; \boldsymbol{\vartheta})}\right)\right) p(\mathbf{x}) d\mathbf{x} \\ &= -\frac{1}{c(\boldsymbol{\gamma})} \int_{\Omega_{\mathbf{x}}} \exp\left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma})}{f(\mathbf{x}; \boldsymbol{\vartheta})}\right) \frac{1}{f(\mathbf{x}; \boldsymbol{\vartheta})} \nabla_{\boldsymbol{\gamma}} f(\mathbf{x}; \boldsymbol{\gamma}) p(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (\text{B.36})$$

Substitution of (B.36) in (B.34) results in

$$\begin{aligned} \nabla_{\boldsymbol{\gamma}} \log \tilde{p}(\mathbf{x}; \boldsymbol{\gamma}) &= -\frac{1}{1 + \exp\left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma})}{f(\mathbf{x}; \boldsymbol{\vartheta})}\right)} \exp\left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma})}{f(\mathbf{x}; \boldsymbol{\vartheta})}\right) \frac{1}{f(\mathbf{x}; \boldsymbol{\vartheta})} \nabla_{\boldsymbol{\gamma}} f(\mathbf{x}; \boldsymbol{\gamma}) \\ &+ \frac{1}{c(\boldsymbol{\gamma})} \int_{\Omega_{\mathbf{x}}} \exp\left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma})}{f(\mathbf{x}; \boldsymbol{\vartheta})}\right) \frac{1}{f(\mathbf{x}; \boldsymbol{\vartheta})} \nabla_{\boldsymbol{\gamma}} f(\mathbf{x}; \boldsymbol{\gamma}) p(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (\text{B.37})$$

By using regularity condition RC.3 and (B.29), there is a neighborhood of $\boldsymbol{\vartheta}$ such that

$$\left| \frac{\partial \log \tilde{p}(\mathbf{x}; \boldsymbol{\gamma})}{\partial \gamma_i} \right| \leq em(\mathbf{x}) + \int_{\Omega_{\mathbf{x}}} em(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (\text{B.38})$$

Hence, the assumptions of Theorem 4.1 are sufficient for Lemma B.1. \square

In the following, we show the relation between the gradient of $\tilde{p}(\mathbf{x}; \boldsymbol{\gamma})$ to the gradient of $f(\mathbf{x}; \boldsymbol{\theta})$: We notice that evaluating (B.36) at $\boldsymbol{\gamma} = \boldsymbol{\vartheta}$, results in

$$\nabla_{\boldsymbol{\gamma}} \log c(\boldsymbol{\gamma}) \Big|_{\boldsymbol{\vartheta}} = -\frac{1}{2} \int_{\Omega_{\mathbf{x}}} \nabla_{\boldsymbol{\gamma}} \log f(\mathbf{x}; \boldsymbol{\vartheta}) p(\mathbf{x}) d\mathbf{x} = -\frac{1}{2} \nabla_{\boldsymbol{\theta}} \mathbb{E}_p [\log f(\mathbf{x}; \boldsymbol{\theta})] \Big|_{\boldsymbol{\vartheta}} = \mathbf{0}, \quad (\text{B.39})$$

since $\boldsymbol{\vartheta}$ is the maximizer of $\mathbb{E}_p [\log f(\mathbf{x}; \boldsymbol{\theta})]$. Therefore, evaluating (B.34) at $\boldsymbol{\gamma} = \boldsymbol{\vartheta}$ results in

$$\nabla_{\boldsymbol{\gamma}} \log \tilde{p}(\mathbf{x}; \boldsymbol{\gamma}) \Big|_{\boldsymbol{\vartheta}} = \nabla_{\boldsymbol{\gamma}} \log \left(1 + \exp\left(1 - \frac{f(\mathbf{x}; \boldsymbol{\gamma})}{f(\mathbf{x}; \boldsymbol{\vartheta})}\right)\right) \Big|_{\boldsymbol{\vartheta}} = -\frac{1}{2} \nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\vartheta}). \quad (\text{B.40})$$

Substitution of (B.40) in (B.7) and (B.21), in addition to the fact that $\tilde{p}(\mathbf{x}; \boldsymbol{\vartheta}) = p(\mathbf{x})$, results in

$$\mathbf{J}_{\tilde{p}}(\boldsymbol{\vartheta}) = \frac{1}{4} \mathbb{E}_p [\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\vartheta}) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}; \boldsymbol{\vartheta})] = \frac{1}{4} \mathbf{B}(\boldsymbol{\vartheta}) \quad (\text{B.41})$$

and

$$\mathbf{B}_{\tilde{p},f}(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}) = -\frac{1}{2}\mathbb{E}_p[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\vartheta}) \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}; \boldsymbol{\vartheta})] = -\frac{1}{2}\mathbf{B}(\boldsymbol{\vartheta}), \quad (\text{B.42})$$

respectively. The inequality in (B.9) holds for any $\boldsymbol{\gamma}$, in particular for $\boldsymbol{\gamma} = \boldsymbol{\vartheta}$. Therefore, substitution of (B.41) and (B.42) in (B.9) results in the bound in (4.8), which concludes this proof.

Appendix C

C.i Detailed developments for the example in (4.79)

In Section 5.3, $\mathbf{x} \in \mathbb{R}^N$ is assumed to be an observation vector such that (see (4.79)) $\mathbf{x} = \boldsymbol{\varphi} + \mathbf{w}$, where $\boldsymbol{\varphi} \in \mathbb{R}^N$ is an unknown deterministic parameter vector to be estimated, and \mathbf{w} is a white Gaussian noise with zero mean and a known covariance matrix, $\sigma^2 \mathbf{I}$. The two hypotheses regarding $\boldsymbol{\varphi}$ are given in (4.80). Thus, it can be seen that in the considered setting, the candidate pdfs are both Gaussian, with means $\mathbf{H}\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$, respectively, i.e.

$$f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) = \begin{cases} (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}^{(1)}\|^2}, & k = 1 \\ (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \|\mathbf{x} - \boldsymbol{\theta}^{(2)}\|^2}, & k = 2. \end{cases} \quad (\text{C.1})$$

In addition, under the considered GLRT selection rule from (4.83), the true and assumed probabilities of selection are given by (4.86) and (4.88), respectively. In particular, for the considered setting, $\pi_1(\boldsymbol{\theta}^{(1)})$ is not a function of the parameter $\boldsymbol{\theta}^{(1)}$. Hence, in the following, we use the notation, $\pi_1(\boldsymbol{\theta}^{(1)}) = \pi_1$.

C.i.1 Estimators

In this subsection, we derive the estimators from Section 5.1.

C.i.1.1 Naive Interpretation- MSL Estimator

The MSL estimator from Subsection 4.4.1 is given by the maximization in (4.1), which is obtained via the solution to the following equation

$$\nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) = \mathbf{0}, \quad \forall \mathbf{x} \in \mathcal{A}_k. \quad (\text{C.2})$$

The gradient of (C.1) w.r.t. $\boldsymbol{\theta}^{(k)}$ is given by

$$\nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) = \begin{cases} \frac{1}{\sigma^2} \mathbf{H}^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}^{(1)}), & k = 1 \\ \frac{1}{\sigma^2} (\mathbf{x} - \boldsymbol{\theta}^{(2)}), & k = 2. \end{cases} \quad (\text{C.3})$$

C.i.1.2 Normalized Interpretation- MSNL estimator

By substituting (4.88) in (4.20), we obtain that for the considered setting, the normalization factor is given by

$$\alpha(\boldsymbol{\theta}) = \pi_1 + \pi_2(\boldsymbol{\theta}^{(2)}) = F_r(\gamma; 0) + 1 - F_r(\gamma; \lambda^{(2)}). \quad (\text{C.4})$$

The MSNL estimator is given by the maximization in (4.31), i.e. as the solution of the following score equation

$$\nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) - \nabla_{\boldsymbol{\theta}^{(k)}} \log \alpha_k(\boldsymbol{\theta}^{(k)}) = \mathbf{0}, \quad \forall \mathbf{x} \in \mathcal{A}_k. \quad (\text{C.5})$$

where

$$\alpha_k(\boldsymbol{\theta}^{(k)}) \triangleq \pi_k(\boldsymbol{\theta}^{(k)}) + \sum_{l \neq k} \pi_l. \quad (\text{C.6})$$

Since $\alpha(\boldsymbol{\theta})$ in (C.4) is not a function of $\boldsymbol{\theta}^{(1)}$, $\alpha_1(\boldsymbol{\theta}^{(1)})$ from (C.6) is not a function of $\boldsymbol{\theta}^{(1)}$, and thus, under $k = 1$, (C.5) is reduced to the same score equation as in (C.2). Therefore, for $k = 1$ (i.e. $\mathbf{x} \in \mathcal{A}_1$) we obtain (4.30). On the other hand, for $\mathbf{x} \in \mathcal{A}_2$ the MSNL estimator of $\boldsymbol{\theta}^{(2)}$ is obtained by the maximization in (C.5) w.r.t. $\boldsymbol{\theta}^{(2)}$. Since π_1 is not a function of $\boldsymbol{\theta}^{(1)}$ then $\alpha_2(\boldsymbol{\theta}^{(2)})$ is identical to $\alpha(\boldsymbol{\theta})$ in (C.4). Therefore, (C.5) is given by

$$\nabla_{\boldsymbol{\theta}^{(2)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(2)}) - \nabla_{\boldsymbol{\theta}^{(2)}} \log \alpha(\boldsymbol{\theta}) = \mathbf{0}. \quad (\text{C.7})$$

In order to derive the last term in (C.7), we note that the cdf of non-central χ^2 with r degrees of freedom is given by

$$F_r(\gamma, \lambda) = \sum_{m=0}^{\infty} e^{-\frac{\lambda}{2}} \frac{\left(\frac{\lambda}{2}\right)^m}{m!} F_{r+2m}(\gamma, 0). \quad (\text{C.8})$$

The derivative of (C.8) w.r.t. the non-centrality parameter, λ , is given by [171]

$$\begin{aligned}
\frac{\partial F_r(\gamma, \lambda)}{\partial \lambda} &= -\frac{1}{2}F_r(\gamma, \lambda) + \frac{1}{2} \sum_{m=1}^{\infty} e^{-\frac{\lambda}{2}} \frac{\left(\frac{\lambda}{2}\right)^{m-1}}{(m-1)!} F_{r+2m}(\gamma, 0) \\
&= -\frac{1}{2}F_r(\gamma, \lambda) + \frac{1}{2} \sum_{l=0}^{\infty} e^{-\frac{\lambda}{2}} \frac{\left(\frac{\lambda}{2}\right)^l}{(l)!} F_{r+2+2l}(\gamma, 0) \\
&= -\frac{1}{2}F_r(\gamma, \lambda) + \frac{1}{2}F_{r+2}(\gamma, \lambda). \tag{C.9}
\end{aligned}$$

By applying the chain rule, and using (4.87) and $\pi_2(\boldsymbol{\theta}^{(2)}) = 1 - F_r(\gamma; \lambda^{(2)})$ from (4.88), we obtain

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}^{(2)}} \pi_2(\boldsymbol{\theta}^{(2)}) &= \frac{\partial \pi_2(\boldsymbol{\theta}^{(2)})}{\partial \lambda^{(2)}} \nabla_{\boldsymbol{\theta}^{(2)}} \lambda^{(2)} \\
&= \frac{1}{\sigma^2} (F_r(\gamma, \lambda^{(2)}) - F_{r+2}(\gamma, \lambda^{(2)})) \mathbf{P}_{\mathbf{H}}^{\perp} \boldsymbol{\theta}^{(2)}. \tag{C.10}
\end{aligned}$$

Hence, (C.10) implies that the gradient of the right term on the l.h.s. of (C.7) w.r.t. $\boldsymbol{\theta}^{(2)}$ is given by

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}^{(2)}} \log \alpha(\boldsymbol{\theta}) &= \frac{\nabla_{\boldsymbol{\theta}^{(2)}} \pi_2(\boldsymbol{\theta}^{(2)})}{\alpha(\boldsymbol{\theta})} \\
&= \frac{1}{\sigma^2} \frac{F_r(\gamma, \lambda^{(2)}) - F_{r+2}(\gamma, \lambda^{(2)})}{F_r(\gamma; 0) + 1 - F_r(\gamma; \lambda^{(2)})} \mathbf{P}_{\mathbf{H}}^{\perp} \boldsymbol{\theta}^{(2)}. \tag{C.11}
\end{aligned}$$

By substituting (C.3) and (C.11) in (C.7), we obtain that the score equation of the MSL estimator for $\mathbf{x} \in \mathcal{A}_2$ is reduced to (4.90).

C.i.1.3 Selective Inference Interpretation- PSML estimator

The PSML estimator is given by the maximization in (4.35), i.e. as the solution to the following score equation

$$\nabla_{\boldsymbol{\theta}^{(k)}} \log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) - \nabla_{\boldsymbol{\theta}^{(k)}} \log \pi_k(\boldsymbol{\theta}^{(k)}) = \mathbf{0}, \quad \forall \mathbf{x} \in \mathcal{A}_k. \tag{C.12}$$

Since π_1 from (4.88) is not a function of $\boldsymbol{\theta}^{(1)}$, under $k = 1$, (C.12) is reduced to the same score equation as in (C.2). Therefore, for $k = 1$ (i.e. $\mathbf{x} \in \mathcal{A}_1$) we obtain (4.91).

For $\mathbf{x} \in \mathcal{A}_2$, the PSML estimator of $\boldsymbol{\theta}^{(2)}$ is obtained by solving (C.12). By using

(C.10), we can verify that

$$\begin{aligned}\nabla_{\boldsymbol{\theta}^{(2)}} \log \pi_2(\boldsymbol{\theta}^{(2)}) &= \frac{\nabla_{\boldsymbol{\theta}^{(2)}} \pi_2(\boldsymbol{\theta}^{(2)})}{\pi_2(\boldsymbol{\theta}^{(2)})} \\ &= \frac{1}{\sigma^2} \frac{F_r(\gamma, \lambda^{(2)}) - F_{r+2}(\gamma, \lambda^{(2)})}{1 - F_r(\gamma; \lambda^{(2)})} \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\theta}^{(2)}.\end{aligned}\quad (\text{C.13})$$

By substituting (C.3) and (C.13) in (C.12), the score equation of the PSML estimator for $\mathbf{x} \in \mathcal{A}_2$ is reduced to (4.92).

C.i.2 Post-Selection-Pseudo-true Parameters

In this subsection, we develop the PS-pseudo-true parameter vectors from Section 4.6 for each interpretation.

C.i.2.1 Naive Interpretation

The PS-pseudo-true parameter vector under the naive interpretation is given by the maximization in (4.43), i.e. as the solution of

$$\nabla_{\boldsymbol{\theta}^{(k)}} \mathbb{E}_p \left[\log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) | \Psi = k \right] = \mathbf{0}, \quad \forall k \in \{1, 2\} \quad (\text{C.14})$$

Since the true pdf, $p(\mathbf{x}; \boldsymbol{\varphi})$ is not a function of the parameters $\boldsymbol{\theta}^{(k)}$, by using (C.3) we obtain that

$$\nabla_{\boldsymbol{\theta}^{(k)}} \mathbb{E}_p \left[\log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) | \Psi = k \right] = \begin{cases} \frac{1}{\sigma^2} \mathbf{H}^T (\boldsymbol{\mu}^{(1)} - \mathbf{H} \boldsymbol{\theta}^{(1)}), & k = 1 \\ \frac{1}{\sigma^2} (\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)}), & k = 2. \end{cases} \quad (\text{C.15})$$

where $\boldsymbol{\mu}^{(k)}$, $k = 1, 2$ is the conditional expectation of \mathbf{x} given the selection, as defined in (4.94). In Appendix C.ii of this report, we derive a closed-form expression for $\boldsymbol{\mu}^{(k)}$, $k = 1, 2$.

Setting the gradient from (C.15) to zero results in (4.93), i.e. in the following term:

$$\boldsymbol{\vartheta}_I^{(k)} = \begin{cases} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \boldsymbol{\mu}^{(1)}, & k = 1 \\ \boldsymbol{\mu}^{(2)}, & k = 2. \end{cases} \quad (\text{C.16})$$

C.i.2.2 Normalized Interpretation

The PS-pseudo-true parameter vector under the normalized interpretation is given by the maximization in (4.48). Therefore, it is given by the solution of the following equation

$$\nabla_{\boldsymbol{\theta}^{(k)}} \mathbb{E}_p \left[\log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) | \Psi = k \right] - \nabla_{\boldsymbol{\theta}^{(k)}} \log \alpha_k(\boldsymbol{\theta}^{(k)}) = \mathbf{0}. \quad (\text{C.17})$$

In the considered setting $\alpha(\boldsymbol{\theta})$ in (C.4) is not a function of $\boldsymbol{\theta}^{(1)}$, then α_k is also not a function of $\boldsymbol{\theta}^{(1)}$ (for both $k = 1, 2$). Thus, for $k = 1$ the solution of (C.17) is given by

$$\boldsymbol{\vartheta}_{II}^{(1)} = \boldsymbol{\vartheta}_I^{(1)} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \boldsymbol{\mu}^{(1)}. \quad (\text{C.18})$$

By using (C.11) and (C.15), $\boldsymbol{\vartheta}_{II}^{(2)}$ is given as the solution of the following equation

$$\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)} - \frac{F_r(\gamma; \lambda^{(2)}) - F_{r+2}(\gamma; \lambda^{(2)})}{F_r(\gamma; 0) + 1 - F_r(\gamma; \lambda^{(2)})} \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\theta}^{(2)} = \mathbf{0}. \quad (\text{C.19})$$

C.i.2.3 Selective Inference Interpretation

The PS-pseudo-true parameter vector under the selective inference interpretation is given by the maximization in (4.49). Thus, it is given by the solution of the following equation

$$\nabla_{\boldsymbol{\theta}^{(k)}} \mathbb{E}_p \left[\log f_k(\mathbf{x}; \boldsymbol{\theta}^{(k)}) | \Psi = k \right] - \nabla_{\boldsymbol{\theta}^{(k)}} \log \pi_k(\boldsymbol{\theta}^{(k)}) = \mathbf{0}. \quad (\text{C.20})$$

Similarly to the normalized interpretation, since π_1 is not a function of $\boldsymbol{\theta}^{(1)}$ we obtain that (see also (4.97))

$$\boldsymbol{\vartheta}_{III}^{(1)} = \boldsymbol{\vartheta}_I^{(1)} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \boldsymbol{\mu}^{(1)}. \quad (\text{C.21})$$

By substituting (C.13) and (C.15) in (C.20), $\boldsymbol{\vartheta}_{III}^{(2)}$ is given by the solution of the following equation.

$$\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)} - \frac{F_r(\gamma; \lambda^{(2)}) - F_{r+2}(\gamma; \lambda^{(2)})}{1 - F_r(\gamma; \lambda^{(2)})} \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\theta}^{(2)} = \mathbf{0}. \quad (\text{C.22})$$

In Appendix C.ii.1 in this report we show that

$$\boldsymbol{\mu}^{(k)} = \boldsymbol{\varphi} + (-1)^k \frac{1}{p_k} (F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)) \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\varphi}. \quad (\text{C.23})$$

Hence, by substituting $\boldsymbol{\mu}^{(2)}$ in (C.22), we can verify that the solution of (C.22) is obtained for $\boldsymbol{\vartheta}_{III}^{(2)} = \boldsymbol{\varphi}$ as appears in (4.98).

C.i.3 Post-selection PS-MCRB

In this subsection, we derive the different PS-MCRBs from Section 5.2, for each interpretation. In particular we will derive the matrices $\mathbf{A}_i^{(k)}(\boldsymbol{\vartheta}_i^{(k)})$, $\mathbf{B}_i^{(k)}(\boldsymbol{\vartheta}_i^{(k)})$, $i \in \{I, II, III\}$.

C.i.3.1 Naive Interpretation

By using (C.1), the Hessian matrix of the log-likelihood under each of the hypotheses is given by the matrix in (4.99), which is not a function of $\boldsymbol{\theta}^{(k)}$ for any k . Substitution (4.99) in (4.66) results in $\mathbf{A}_I^{(k)}(\boldsymbol{\vartheta}_I^{(k)})$ in (4.100).

The k th outer-product form information matrix is given by (4.67). By substituting (C.3) in (4.67) we obtain that

$$\mathbf{B}_I^{(1)}(\boldsymbol{\theta}^{(1)}) = \frac{1}{\sigma^4} \mathbf{H}^T \mathbb{E}_p[(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}^{(1)})(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}^{(1)})^T | \Psi = 1] \mathbf{H}, \quad (\text{C.24})$$

and respectively,

$$\mathbf{B}_I^{(2)}(\boldsymbol{\theta}^{(1)}) = \frac{1}{\sigma^4} \mathbb{E}_p[(\mathbf{x} - \boldsymbol{\theta}^{(2)})(\mathbf{x} - \boldsymbol{\theta}^{(2)})^T | \Psi = 2]. \quad (\text{C.25})$$

We can rewrite the expectation on the l.h.s. of (C.24) as follows

$$\mathbb{E}[(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}^{(1)})(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}^{(1)})^T | \Psi = 1] = \boldsymbol{\Sigma}_{\mathbf{x}}^{(1)} + (\boldsymbol{\mu}^{(1)} - \mathbf{H}\boldsymbol{\theta}^{(1)})(\boldsymbol{\mu}^{(1)} - \mathbf{H}\boldsymbol{\theta}^{(1)})^T, \quad (\text{C.26})$$

where $\boldsymbol{\Sigma}_{\mathbf{x}}^{(k)}$, defined in (4.104), is the conditional covariance matrix of \mathbf{x} given $\Psi = k$ w.r.t. the true pdf. A closed-form expression for $\boldsymbol{\Sigma}_{\mathbf{x}}^{(k)}$ appears in (C.61) in the following. Substitution of (C.26) in (C.24) results in

$$\mathbf{B}_I^{(1)}(\boldsymbol{\theta}^{(1)}) = \frac{1}{\sigma^4} \mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{x}}^{(1)} \mathbf{H}^T + \frac{1}{\sigma^4} \mathbf{H}^T (\boldsymbol{\mu}^{(1)} - \mathbf{H}\boldsymbol{\theta}^{(1)})(\boldsymbol{\mu}^{(1)} - \mathbf{H}\boldsymbol{\theta}^{(1)})^T \mathbf{H}^T. \quad (\text{C.27})$$

The PS-pseudo-true parameter vector, according to the naive interpretation, $\boldsymbol{\vartheta}_I^{(1)}$, satisfies (C.14), which implies that

$$\boldsymbol{\mu}^{(1)} - \mathbf{H}\boldsymbol{\vartheta}_I^{(1)} = \boldsymbol{\mu}^{(1)} - \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \boldsymbol{\mu}^{(1)} = \mathbf{P}_{\mathbf{H}}^{\perp} \boldsymbol{\mu}^{(1)}. \quad (\text{C.28})$$

Since $\mathbf{P}_{\mathbf{H}}^{\perp}$ is the orthogonal projection to \mathbf{H} , substitution of (C.28) in (C.27) results in

$$\mathbf{B}_I^{(1)}(\boldsymbol{\vartheta}_I^{(1)}) = \frac{1}{\sigma^4} \mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{x}}^{(1)} \mathbf{H}. \quad (\text{C.29})$$

In a similar manner to (C.26), we can write

$$\mathbb{E}[(\mathbf{x} - \boldsymbol{\theta}^{(2)})(\mathbf{x} - \boldsymbol{\theta}^{(2)})^T | \Psi = 2] = \boldsymbol{\Sigma}_{\mathbf{x}}^{(2)} + (\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)})(\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)})^T. \quad (\text{C.30})$$

Substitution of (C.30) in (C.25) results in

$$\mathbf{B}_I^{(2)}(\boldsymbol{\theta}^{(2)}) = \frac{1}{\sigma^4} \boldsymbol{\Sigma}_{\mathbf{x}}^{(2)} + \frac{1}{\sigma^4} (\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)})(\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)})^T. \quad (\text{C.31})$$

By substituting (C.16) in (C.31), the naive interpretation outer-product form information matrix is given by

$$\mathbf{B}_I^{(2)}(\boldsymbol{\vartheta}_I^{(2)}) = \frac{1}{\sigma^4} \boldsymbol{\Sigma}_{\mathbf{x}}^{(2)}. \quad (\text{C.32})$$

C.i.3.2 Normalized Interpretation

In the considered setting, $\alpha(\boldsymbol{\theta})$ is not a function of $\boldsymbol{\theta}^{(1)}$. Thus, the right term on the r.h.s of (4.69) vanishes, and the Hessian form of the information matrix for the normalized interpretation is given by

$$\mathbf{A}_{II}^{(1)}(\boldsymbol{\vartheta}_{II}) = -\frac{1}{\sigma^2} \mathbf{H}^T \mathbf{H}. \quad (\text{C.33})$$

In order to derive $\mathbf{A}_{II}^{(2)}(\boldsymbol{\theta}_{II})$, we first compute the Hessian of $\log \alpha(\boldsymbol{\theta})$, which satisfies

$$\nabla_{\boldsymbol{\theta}^{(2)}}^2 \log \alpha(\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}^{(2)}}^2 \pi_2(\boldsymbol{\theta}^{(2)})}{\alpha(\boldsymbol{\theta})} - \nabla_{\boldsymbol{\theta}^{(2)}} \log \alpha(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}^{(2)}}^T \log \alpha(\boldsymbol{\theta}), \quad (\text{C.34})$$

where the equality is obtained by using the chain rule and the definition of $\alpha(\boldsymbol{\theta})$. The derivative of (C.9) w.r.t. the non centrality parameter, λ , is

$$\frac{\partial^2 F_r(\gamma, \lambda)}{\partial \lambda^2} = -\frac{1}{2} \frac{\partial}{\partial \lambda} (F_r(\gamma, \lambda) - F_{r+2}(\gamma, \lambda)) = \frac{1}{4} (F_r(\gamma, \lambda) - 2F_{r+2}(\gamma, \lambda) + F_{r+4}(\gamma, \lambda)). \quad (\text{C.35})$$

Recall that $\pi_2(\boldsymbol{\theta}^{(2)}) = 1 - F_r(\gamma; \lambda^{(2)})$ from (4.88). Thus, by using the chain rule and the derivative in (C.35) we obtain

$$\begin{aligned} \nabla_{\boldsymbol{\theta}^{(2)}}^2 \pi_2(\boldsymbol{\theta}^{(2)}) &= \frac{1}{\sigma^2} (F_r(\gamma, \lambda^{(2)}) - F_{r+2}(\gamma, \lambda^{(2)})) \mathbf{P}_{\mathbf{H}}^\perp \\ &\quad - \frac{1}{\sigma^4} (F_r(\gamma, \lambda^{(2)}) - 2F_{r+2}(\gamma, \lambda^{(2)}) + F_{r+4}(\gamma, \lambda^{(2)})) \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\theta}^{(2)} (\boldsymbol{\theta}^{(2)})^T \mathbf{P}_{\mathbf{H}}^\perp. \end{aligned} \quad (\text{C.36})$$

Substitution of (C.11) and (C.36) in (C.34), and then substituting the result in (4.69) results in the following closed-form expression for (4.101):

$$\begin{aligned} \mathbf{A}_{II}^{(2)}(\boldsymbol{\vartheta}_{II}^{(2)}) &= -\frac{1}{\sigma^2} \left(\mathbf{I} + \frac{F_r(\gamma; \lambda^{(2)}) - F_{r+2}(\gamma; \lambda^{(2)})}{F_r(\gamma; 0) + 1 - F_r(\gamma; \lambda^{(2)})} \mathbf{P}_{\mathbf{H}}^\perp \right) \\ &\quad + \frac{1}{\sigma^4} \left(\frac{F_r(\gamma; \lambda^{(2)}) - 2F_{r+2}(\gamma; \lambda^{(2)}) + F_{r+4}(\gamma; \lambda^{(2)})}{F_r(\gamma; 0) + 1 - F_r(\gamma; \lambda^{(2)})} \right. \\ &\quad \left. + \left(\frac{F_r(\gamma; \lambda^{(2)}) - F_{r+2}(\gamma; \lambda^{(2)})}{F_r(\gamma; 0) + 1 - F_r(\gamma; \lambda^{(2)})} \right)^2 \right) \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\vartheta}_{II}^{(k)} (\boldsymbol{\vartheta}_{II}^{(k)})^T \mathbf{P}_{\mathbf{H}}^\perp. \end{aligned} \quad (\text{C.37})$$

The outer-product form information matrix under the normalized interpretation is given by (4.72). Since $\alpha(\boldsymbol{\theta})$ is not a function of $\boldsymbol{\theta}^{(1)}$ and since $\boldsymbol{\vartheta}_{II}^{(1)} = \boldsymbol{\vartheta}_I^{(1)}$, then

$$\mathbf{B}_{II}^{(1)}(\boldsymbol{\vartheta}_{II}^{(1)}) = \mathbf{B}_I^{(1)}(\boldsymbol{\vartheta}_I^{(1)}) = \frac{1}{\sigma^4} \mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{x}}^{(1)} \mathbf{H}^T. \quad (\text{C.38})$$

Substitution of (C.31) results in

$$\mathbf{B}_I^{(2)}(\boldsymbol{\theta}^{(2)}) = \frac{1}{\sigma^4} \boldsymbol{\Sigma}_{\mathbf{x}}^{(2)} + \frac{1}{\sigma^4} (\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)}) (\boldsymbol{\mu}^{(2)} - \boldsymbol{\theta}^{(2)})^T - \nabla_{\boldsymbol{\theta}^{(k)}} \log \alpha(\boldsymbol{\vartheta}_{II}) \nabla_{\boldsymbol{\theta}^{(k)}}^T \log \alpha(\boldsymbol{\vartheta}_{II}). \quad (\text{C.39})$$

Recall that by the definition, the PS-pseudo-true vector, $\boldsymbol{\vartheta}_{II}^{(2)}$ satisfies (C.5) i.e. $\nabla_{\boldsymbol{\theta}^{(2)}} \log \alpha(\boldsymbol{\vartheta}_{II}^{(2)}) = \frac{1}{\sigma^2} (\boldsymbol{\mu}^{(2)} - \boldsymbol{\vartheta}_{II}^{(2)})$ and thus

$$\mathbf{B}_{II}^{(2)}(\boldsymbol{\vartheta}_{II}^{(k)}) = \frac{1}{\sigma^4} \boldsymbol{\Sigma}_{\mathbf{x}}^{(2)}. \quad (\text{C.40})$$

C.i.3.3 Selective Inference Interpretation

The Hessian form of the Information matrix under the normalized interpretation is given by (4.74). Similarly to the normalized interpretation, since in our settings, π_1 is not a function of $\boldsymbol{\theta}^{(1)}$, thus, we obtain (4.102). By using the chain rule, substitution of (C.36) in (4.74) results in the following closed form expression for (4.102):

$$\begin{aligned} \mathbf{A}_{III}^{(2)}(\boldsymbol{\vartheta}_{III}^{(2)}) &= -\frac{1}{\sigma^2} \left(\mathbf{I} + \frac{F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)}{1 - F_r(\gamma; \lambda)} \mathbf{P}_{\mathbf{H}}^\perp \right) \\ &\quad + \frac{1}{\sigma^4} \left(\frac{F_r(\gamma; \lambda) - 2F_{r+2}(\gamma; \lambda) + F_{r+4}(\gamma; \lambda)}{1 - F_r(\gamma; \lambda)} + \left(\frac{F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)}{1 - F_r(\gamma; \lambda)} \right)^2 \right) \mathbf{P}_{\mathbf{H}}^\perp \boldsymbol{\varphi} \boldsymbol{\varphi}^T \mathbf{P}_{\mathbf{H}}^\perp. \end{aligned} \quad (\text{C.41})$$

The outer-product form information matrix under the selective inference interpretation is given by (4.77). Similarly to the normalized interpretation, since π_1 is not a function of

$\boldsymbol{\theta}^{(1)}$. Since $\boldsymbol{\vartheta}_{III}^{(1)} = \boldsymbol{\vartheta}_I^{(1)}$ we obtain that

$$\mathbf{B}_{III}^{(1)}(\boldsymbol{\vartheta}_{III}^{(1)}) = \mathbf{B}_I^{(1)}(\boldsymbol{\vartheta}_{III}^{(1)}) = \frac{1}{\sigma^4} \mathbf{H}^T \boldsymbol{\Sigma}_{\mathbf{x}}^{(1)} \mathbf{H}. \quad (\text{C.42})$$

Substitution of (C.31) results in

$$\mathbf{B}_{III}^{(2)}(\boldsymbol{\vartheta}_{III}^{(k)}) = \frac{1}{\sigma^4} \boldsymbol{\Sigma}_{\mathbf{x}}^{(2)} + \frac{1}{\sigma^4} (\boldsymbol{\mu}^{(2)} - \boldsymbol{\vartheta}_{III}^{(k)}) (\boldsymbol{\mu}^{(2)} - \boldsymbol{\vartheta}_{III}^{(k)})^T - \nabla_{\boldsymbol{\theta}^{(k)}} \log \pi_2(\boldsymbol{\vartheta}_{III}^{(2)}) \nabla_{\boldsymbol{\theta}^{(2)}}^T \log \pi_2(\boldsymbol{\vartheta}_{III}^{(2)}). \quad (\text{C.43})$$

Since $\boldsymbol{\vartheta}_{III}^{(2)}$ satisfies (C.20), (C.43) results in

$$\mathbf{B}_{III}^{(2)}(\boldsymbol{\vartheta}_{III}^{(k)}) = \frac{1}{\sigma^4} \boldsymbol{\Sigma}_{\mathbf{x}}^{(2)}. \quad (\text{C.44})$$

To conclude, we obtain that the outer-product form of the k th post-model-selection information matrices are identical for all the interpretations and are given by (4.103).

C.ii Conditional Expectations

In this section, we derive analytic, closed forms expressions for the conditional expectations, $\boldsymbol{\mu}^{(k)}$, and covariance matrices, $\boldsymbol{\Sigma}_{\mathbf{x}}^{(k)}$, from (4.94) and (4.104), respectively. To this end, in the following proposition, we derive the moment-generating function (MGF), for this case.

Proposition C.1. *Let $\mathbf{x} \in \mathbb{R}^N$, an observation vector such that*

$$\mathbf{x} = \boldsymbol{\varphi} + \mathbf{w}, \quad (\text{C.45})$$

where $\boldsymbol{\varphi} \in \mathbb{R}^N$ is an unknown deterministic parameter vector, and \mathbf{w} is a white Gaussian noise with zero mean and a known covariance matrix, $\sigma^2 \mathbf{I}$, i.e. $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\varphi}, \sigma^2 \mathbf{I})$. The MGF of \mathbf{x} given $\mathbf{x} \in \mathcal{A}_k$, $m_k(\mathbf{t})$, is given by

$$m_k(\mathbf{t}) \triangleq \mathbb{E}_p[e^{\mathbf{t}^T \mathbf{x}} | \mathbf{x} \in \mathcal{A}_k] = \frac{1}{p_k} e^{\mathbf{t}^T \boldsymbol{\varphi} + \frac{\sigma^2}{2} \mathbf{t}^T \mathbf{t}} p_{k,\mathbf{t}}, \quad (\text{C.46})$$

where

$$p_{k,\mathbf{t}} \triangleq \int_{\mathcal{A}_k} p(\mathbf{x}; \boldsymbol{\varphi} + \sigma^2 \mathbf{t}) d\mathbf{x}, \quad k \in \{1, 2\}, \quad (\text{C.47})$$

is the probability that $\mathbf{x} \in \mathcal{A}_k$ if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\varphi} + \sigma^2 \mathbf{t}, \sigma^2 \mathbf{I})$. Notice that in particular $p_{k,\mathbf{0}} = p_k$.

Proof. The MGF, $m_k(\mathbf{t})$, is given by

$$m_k(\mathbf{t}) = \int_{\Omega_{\mathbf{x}}} e^{\mathbf{t}^T \mathbf{x}} p(\mathbf{x} | \mathbf{x} \in \mathcal{A}_k; \boldsymbol{\varphi}) d\mathbf{x}. \quad (\text{C.48})$$

By using Bayes rule, (C.48) can be rewritten as

$$m_k(\mathbf{t}) = \frac{1}{p_k} \int_{\mathcal{A}_k} e^{\mathbf{t}^T \mathbf{x}} p(\mathbf{x}; \boldsymbol{\varphi}) d\mathbf{x}. \quad (\text{C.49})$$

Substitution of $p(\mathbf{x}; \boldsymbol{\varphi})$, a Gaussian pdf, results in

$$\begin{aligned} & \frac{1}{p_k} \int_{\mathcal{A}_k} e^{\mathbf{t}^T \mathbf{x}} p(\mathbf{x}; \boldsymbol{\varphi}) d\mathbf{x} \\ &= \frac{1}{p_k} \int_{\mathcal{A}_k} (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \|\mathbf{x} - \boldsymbol{\varphi}\|^2 + \mathbf{t}^T \mathbf{x}} d\mathbf{x} \\ &= \frac{1}{p_k} \int_{\mathcal{A}_k} (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \|\mathbf{x} - (\boldsymbol{\varphi} + \sigma^2 \mathbf{t})\|^2 + \mathbf{t}^T \boldsymbol{\varphi} + \frac{\sigma^2}{2} \mathbf{t}^T \mathbf{t}} d\mathbf{x} \\ &= \frac{1}{p_k} e^{\mathbf{t}^T \boldsymbol{\varphi} + \frac{\sigma^2}{2} \mathbf{t}^T \mathbf{t}} \int_{\mathcal{A}_k} (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \|\mathbf{x} - (\boldsymbol{\varphi} + \sigma^2 \mathbf{t})\|^2} d\mathbf{x}. \end{aligned} \quad (\text{C.50})$$

Notice that

$$(2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \|\mathbf{x} - (\boldsymbol{\varphi} + \sigma^2 \mathbf{t})\|^2} = p(\mathbf{x}; (\boldsymbol{\varphi} + \sigma^2 \mathbf{t})) \quad (\text{C.51})$$

which results in (C.46). \square

The moments of \mathbf{x} given the k th selection are obtained by the derivatives of the k th MGF, $m_k(\mathbf{t})$.

C.ii.1 Conditional Expectations

In the settings in Section 5.3, we consider the GLRT selection rule from (4.83), $\psi(\mathbf{x}) \sim \chi_r^2(\lambda)$. Therefore, $p_{k,\mathbf{t}}$ from (C.47) are given by

$$p_{k,\mathbf{t}} = \begin{cases} F_r(\gamma; \lambda_{\mathbf{t}}), & k = 1 \\ 1 - F_r(\gamma; \lambda_{\mathbf{t}}), & k = 2, \end{cases} \quad (\text{C.52})$$

where

$$\lambda_{\mathbf{t}} \triangleq \frac{(\boldsymbol{\varphi} + \sigma^2 \mathbf{t})^T \mathbf{P}_{\mathbf{H}}^{\perp} (\boldsymbol{\varphi} + \sigma^2 \mathbf{t})}{\sigma^2}. \quad (\text{C.53})$$

By substitution of and by using the chain rule, we obtain that

$$\nabla_{\mathbf{t}} m_k(\mathbf{t}) = m_k(\mathbf{t}) (\boldsymbol{\varphi} + \sigma^2 \mathbf{t}) + \frac{1}{p_k} e^{\mathbf{t}^T \boldsymbol{\varphi} + \frac{\sigma^2}{2} \mathbf{t}^T \mathbf{t}} \nabla_{\mathbf{t}} p_{k,\mathbf{t}}. \quad (\text{C.54})$$

By substituting $p_{k,\mathbf{t}}$ from (C.52) and by using the derivative from (C.9) we obtain that

$$\nabla_{\mathbf{t}} p_{k,\mathbf{t}} = \frac{\partial p_{k,\mathbf{t}}}{\partial \lambda_{\mathbf{t}}} \nabla_{\mathbf{t}} \lambda_{\mathbf{t}} = (-1)^k (F_r(\gamma; \lambda_{\mathbf{t}}) - F_{r+2}(\gamma; \lambda_{\mathbf{t}})) \mathbf{P}_{\mathbf{H}}^{\perp} (\boldsymbol{\varphi} + \sigma^2 \mathbf{t}). \quad (\text{C.55})$$

$$\nabla_{\mathbf{t}} m_k(\mathbf{t}) = m_k(\mathbf{t}) (\boldsymbol{\varphi} + \sigma^2 \mathbf{t}) + (-1)^k \frac{1}{p_k} e^{\mathbf{t}^T \boldsymbol{\varphi} + \frac{\sigma^2}{2} \mathbf{t}^T \mathbf{t}} (F_r(\gamma; \lambda_{\mathbf{t}}) - F_{r+2}(\gamma; \lambda_{\mathbf{t}})) \mathbf{P}_{\mathbf{H}}^{\perp} (\boldsymbol{\varphi} + \sigma^2 \mathbf{t}). \quad (\text{C.56})$$

Evaluation of (C.56) at $\mathbf{t} = \mathbf{0}$ results in

$$\boldsymbol{\mu}^{(k)} = \nabla_{\mathbf{t}} m_k(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{0}} = \boldsymbol{\varphi} + (-1)^k \frac{1}{p_k} (F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)) \mathbf{P}_{\mathbf{H}}^{\perp} \boldsymbol{\varphi}. \quad (\text{C.57})$$

C.ii.2 Conditional Covariance Matrices

By taking the gradient w.r.t. \mathbf{t} on (C.56) we obtain that

$$\nabla_{\mathbf{t}}^2 m_k(\mathbf{t}) = m_k(\mathbf{t}) \sigma^2 \mathbf{I} + \nabla_{\mathbf{t}} m_k(\mathbf{t}) (\boldsymbol{\varphi} + \sigma^2 \mathbf{t})^T + \frac{1}{p_k} e^{\mathbf{t}^T \boldsymbol{\varphi} + \frac{\sigma^2}{2} \mathbf{t}^T \mathbf{t}} \left((\boldsymbol{\varphi} + \sigma^2 \mathbf{t}) \nabla_{\mathbf{t}}^T p_{k,\mathbf{t}} + \nabla_{\mathbf{t}}^2 p_{k,\mathbf{t}} \right) \quad (\text{C.58})$$

where $\nabla_{\mathbf{t}} p_{k,\mathbf{t}}$ is given in (C.55) and

$$\begin{aligned} \nabla_{\mathbf{t}}^2 p_{k,\mathbf{t}} &= (-1)^{k-1} (F_r(\gamma; \lambda_{\mathbf{t}}) - 2F_{r+2}(\gamma; \lambda_{\mathbf{t}}) + F_{r+4}(\gamma; \lambda_{\mathbf{t}})) \mathbf{P}_{\mathbf{H}}^{\perp} (\boldsymbol{\varphi} + \sigma^2 \mathbf{t}) (\boldsymbol{\varphi} + \sigma^2 \mathbf{t})^T \mathbf{P}_{\mathbf{H}}^{\perp} \\ &\quad + (-1)^k (F_r(\gamma; \lambda_{\mathbf{t}}) - F_{r+2}(\gamma; \lambda_{\mathbf{t}})) \mathbf{P}_{\mathbf{H}}^{\perp} \sigma^2. \end{aligned} \quad (\text{C.59})$$

Evaluation of (C.58) at $\mathbf{t} = \mathbf{0}$ results in

$$\begin{aligned} \nabla_{\mathbf{t}}^2 m_k(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{0}} &= \sigma^2 \mathbf{I} + \boldsymbol{\mu}^{(1)} \boldsymbol{\varphi}^T + \frac{1}{p_k} (-1)^k (F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)) \boldsymbol{\varphi} \boldsymbol{\varphi}^T \mathbf{P}_{\mathbf{H}}^{\perp} \\ &\quad + (-1)^{k-1} \frac{1}{p_k} (F_r(\gamma; \lambda) - 2F_{r+2}(\gamma; \lambda) + F_{r+4}(\gamma; \lambda)) \mathbf{P}_{\mathbf{H}}^{\perp} \boldsymbol{\varphi} \boldsymbol{\varphi}^T \mathbf{P}_{\mathbf{H}}^{\perp} \\ &\quad + (-1)^k \frac{1}{p_k} (F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)) \mathbf{P}_{\mathbf{H}}^{\perp} \sigma^2. \end{aligned} \quad (\text{C.60})$$

Therefore, by substituting (C.57) and (C.60) we obtain that

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{x}}^{(k)} &= \nabla_{\mathbf{t}}^2 m_k(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{0}} - \boldsymbol{\mu}^{(k)} (\boldsymbol{\mu}^{(k)})^T \\ &= \sigma^2 \mathbf{I} + (-1)^{k-1} \frac{1}{p_k} (F_r(\gamma; \lambda) - 2F_{r+2}(\gamma; \lambda) + F_{r+4}(\gamma; \lambda)) \mathbf{P}_{\mathbf{H}}^{\perp} \boldsymbol{\varphi} \boldsymbol{\varphi}^T \mathbf{P}_{\mathbf{H}}^{\perp} \\ &\quad + (-1)^k \frac{1}{p_k} (F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)) \mathbf{P}_{\mathbf{H}}^{\perp} \sigma^2 - \left(\frac{1}{p_k} (F_r(\gamma; \lambda) - F_{r+2}(\gamma; \lambda)) \right)^2 \mathbf{P}_{\mathbf{H}}^{\perp} \boldsymbol{\varphi} \boldsymbol{\varphi}^T \mathbf{P}_{\mathbf{H}}^{\perp}, \end{aligned} \quad (\text{C.61})$$

Appendix D

D.i Proof of Theorem 5.1

Let $\boldsymbol{\xi} : \Omega_{\mathbf{x}} \times \Omega_{\boldsymbol{\theta}} \rightarrow \mathbb{R}^M$ and $\mathbf{A} \in \mathbb{R}^{M \times M}$, then under regularity condition C.1

$$\mathbb{E}[(\boldsymbol{\xi}(\mathbf{x}, \boldsymbol{\theta}) - \mathbf{A} \nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda))^T (\boldsymbol{\xi}(\mathbf{x}, \boldsymbol{\theta}) - \mathbf{A} \nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda)); \Lambda] \geq 0, \quad (\text{D.1})$$

where the equality holds *iff*

$$\boldsymbol{\xi}(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{A} \nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda) \quad \text{a.e.} \quad (\text{D.2})$$

The inequality in (D.1) implies that

$$\begin{aligned} & \mathbb{E}[\boldsymbol{\xi}(\mathbf{x}, \boldsymbol{\theta})^T \boldsymbol{\xi}(\mathbf{x}, \boldsymbol{\theta}); \Lambda] \geq \\ & \mathbb{E}[2\boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}) \mathbf{A} \nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda) - \nabla_{\boldsymbol{\theta}}^T \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda) \mathbf{A}^T \mathbf{A} \nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda); \Lambda] \\ & = \text{Tr} \left(2\mathbf{A} \mathbb{E}[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda) \boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}); \Lambda] - \mathbf{A} \mathbf{J}(\Lambda) \mathbf{A}^T \right), \end{aligned} \quad (\text{D.3})$$

where the last equality is obtained by substituting (5.18) and using the trace operator properties. By using the product rule, it can be verified that

$$\begin{aligned} & \mathbb{E}[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda) \boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}); \Lambda] = \\ & \int_{\Omega_{\mathbf{x}}} \int_{\Omega_{\boldsymbol{\theta}}} \nabla_{\boldsymbol{\theta}} \left(\boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) \right) d\boldsymbol{\theta} f(\mathbf{x}; \Lambda) d\mathbf{x} - \int_{\Omega_{\mathbf{x}}} \int_{\Omega_{\boldsymbol{\theta}}} f(\mathbf{x}, \boldsymbol{\theta}; \Lambda) \nabla_{\boldsymbol{\theta}} \boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{x}. \end{aligned} \quad (\text{D.4})$$

By setting $\boldsymbol{\xi}(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{D}(\hat{\Lambda})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, and using the Leibniz integral rule and regularity condition C.4, one obtains

$$\int_{\Omega_{\boldsymbol{\theta}}} \nabla_{\boldsymbol{\theta}} \left(\boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) \right) d\boldsymbol{\theta} = \mathbf{0}, \quad \text{a.e. } \mathbf{x} \in \Omega_{\mathbf{x}}. \quad (\text{D.5})$$

Since $\mathbf{D}(\hat{\Lambda})$ and $\hat{\boldsymbol{\theta}}$ are not a function of $\boldsymbol{\theta}$, we obtain

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{D}(\hat{\Lambda}) = -\mathbf{D}(\hat{\Lambda}). \quad (\text{D.6})$$

By substituting (D.5) and (D.6) in (D.4), we obtain

$$\mathbb{E}[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{x}, \boldsymbol{\theta}; \Lambda) \boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}); \Lambda] = \boldsymbol{\Pi}. \quad (\text{D.7})$$

Substitution of (D.7) in (D.3) and using (5.5), results in

$$\text{sMSE}(\hat{\boldsymbol{\theta}}) \geq \text{Tr} \left(2\mathbf{A}\boldsymbol{\Pi} - \mathbf{A}\mathbf{J}(\Lambda)\mathbf{A}^T \right). \quad (\text{D.8})$$

Since the inequality in (D.8) holds for any $\mathbf{A} \in \mathbb{R}^{M \times M}$, and since the r.h.s of (D.8) is a concave function w.r.t \mathbf{A} , by setting the derivative to zero it can be verified that the choice

$$\mathbf{A} = \boldsymbol{\Pi}\mathbf{J}^{-1}(\Lambda) \quad (\text{D.9})$$

maximizes (D.8). By substituting (D.9) in (D.8) we obtain that the lower bound in (5.20) is satisfied. By substituting (D.9) in (D.2) we obtain that the equality holds *iff* a.e. (5.22) holds.

D.ii Proof of Theorem 5.2

Consider the inequality from (D.1) when in this case we allow the matrix \mathbf{A} to be a function of \mathbf{x} . In this case, by using the law of total expectation, (D.1) implies that

$$\begin{aligned} \mathbb{E}[\boldsymbol{\xi}(\mathbf{x}, \boldsymbol{\theta})^T \boldsymbol{\xi}(\mathbf{x}, \boldsymbol{\theta}); \Lambda] &\geq \\ \mathbb{E}[\text{Tr}(2\mathbf{A}(\mathbf{x})\mathbb{E}[\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) \boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta})|\mathbf{x}; \Lambda] - \mathbf{A}(\mathbf{x})\mathbf{J}(\mathbf{x}, \Lambda)\mathbf{A}^T(\mathbf{x})); \Lambda], \end{aligned} \quad (\text{D.10})$$

where the last equality is obtained by substituting (5.19). By using the product rule, it can be verified that

$$\mathbb{E}[\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) \boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta})|\mathbf{x}; \Lambda] = \int_{\Omega_{\boldsymbol{\theta}}} \nabla_{\boldsymbol{\theta}} \left(\boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) \right) d\boldsymbol{\theta} - \int_{\Omega_{\boldsymbol{\theta}}} f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) \nabla_{\boldsymbol{\theta}} \boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (\text{D.11})$$

By setting $\boldsymbol{\xi}(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{D}(\hat{\Lambda})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, and by substituting (D.5) and (D.6) in (D.11), we obtain

$$\mathbb{E}[\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{x}; \Lambda) \boldsymbol{\xi}^T(\mathbf{x}, \boldsymbol{\theta})|\mathbf{x}; \Lambda] = \mathbf{D}(\hat{\Lambda}). \quad (\text{D.12})$$

Substitution of (D.12) in (D.10) and using (5.5) results in

$$\text{sMSE}(\hat{\boldsymbol{\theta}}) \geq \text{E} \left[\text{Tr} \left(2\mathbf{A}(\mathbf{x})\mathbf{D}(\hat{\Lambda}) - \mathbf{A}(\mathbf{x})\mathbf{J}(\mathbf{x}, \Lambda)\mathbf{A}^T(\mathbf{x}) \right); \Lambda \right]. \quad (\text{D.13})$$

Since the argument of the expectation in (D.13) is a concave function w.r.t $\mathbf{A}(\mathbf{x})$, by setting the derivative to zero it can be verified that the choice

$$\mathbf{A}(\mathbf{x}) = \mathbf{D}(\hat{\Lambda})\mathbf{J}^{-1}(\mathbf{x}, \Lambda) \quad (\text{D.14})$$

maximizes (D.13) for a.e. $\mathbf{x} \in \Omega_{\mathbf{x}}$. By substituting (D.14) in (D.13) we obtain that (5.20) is satisfied and by substituting (D.14) in (D.2) we obtain that the equality holds in (5.20) *iff* a.e. (5.24) holds.

D.iii Proof of Remark 5.1 (order relation)

It can be verified that substitution of \mathbf{A} from (D.9) in (D.13) results in the selective BCRB bound from (5.20). Since (D.14) maximizes the bound in (D.13), the selective TBCRB is tighter than the selective BCRB. This result is due to the fact that in Appendix D.ii we allow the matrix \mathbf{A} to be a function of \mathbf{x} , while in Appendix D.i the matrix \mathbf{A} is restricted to be a deterministic matrix.

Bibliography

- [1] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993. 2, 3, 4, 16, 20, 27, 35, 49, 51, 79, 85
- [2] H. L. Van Trees, *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004. 2, 3, 4, 92
- [3] H. V. Poor, *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013. 2
- [4] E. Lehmann, “A general concept of unbiasedness,” *The Annals of Mathematical Statistics*, vol. 22, no. 4, pp. 587–592, 1951. 2
- [5] K. Todros and J. Tabrikian, “Uniformly best biased estimators in non-Bayesian parameter estimation,” *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7635–7647, 2011. 2
- [6] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006. 2, 13
- [7] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006. 2, 15, 24, 30
- [8] T. Routtenberg and J. Tabrikian, “Non-Bayesian periodic Cramér-Rao bound,” *IEEE Trans. Signal Process.*, vol. 61, no. 4, pp. 1019–1032, 2013. 2, 3, 15, 101
- [9] T. Routtenberg and L. Tong, “Estimation after parameter selection: Performance analysis and estimation methods,” *IEEE Trans. Signal Process.*, vol. 64, no. 20, pp. 5268–5281, Oct 2016. 2, 3, 6, 14, 17, 27, 31, 46, 48, 50, 64, 66, 69, 89
- [10] —, “The Cramér-Rao bound for estimation-after-selection,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 414–418. 2, 3, 14, 27, 31
- [11] E. Nitzan, T. Routtenberg, and J. Tabrikian, “Optimal biased estimation using Lehmann-unbiasedness,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4496–4500. 2, 3

BIBLIOGRAPHY

- [12] —, “Cramér-Rao bound for constrained parameter estimation using Lehmann-unbiasedness,” *IEEE Trans. Signal Process.*, 2018. 2, 3, 15, 66
- [13] E. Meir and T. Rotttenberg, “Cramér-Rao bound for estimation after model selection and its application to sparse vector estimation,” *IEEE Trans. Signal Process.*, vol. 69, pp. 2284–2301, 2021. 2, 3, 6, 15, 46, 47, 60, 62, 64, 69, 79, 89
- [14] S. Bar and J. Tabrikian, “Bayesian estimation in the presence of deterministic nuisance parameters- part I: Performance bounds,” *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6632–6646, 2015. 2, 89
- [15] —, “Bayesian estimation in the presence of deterministic nuisance parameters- part II: Estimation methods,” *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6647–6658, 2015. 2
- [16] Y. Noam and H. Messer, “Notes on the tightness of the hybrid Cramér-Rao lower bound,” *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2074–2084, 2009. 2
- [17] C. Radhakrishna Rao, “Information and accuracy attainable in the estimation of statistical parameters,” *Bulletin of the Calcutta Mathematical Society*, vol. 37, no. 3, pp. 81–91, 1945. 3
- [18] H. Cramér, “A contribution to the theory of statistical estimation,” *Scandinavian Actuarial Journal*, vol. 1946, no. 1, pp. 85–94, 1946. 3
- [19] A. Bhattacharyya, “On some analogues of the amount of information and their use in statistical estimation,” *Sankhyā: The Indian Journal of Statistics*, pp. 1–14, 1946. 3, 4
- [20] E. W. Barankin, “Locally best unbiased estimates,” *The Annals of Mathematical Statistics*, vol. 20, no. 4, pp. 477–501, 1949. 3
- [21] J. M. Hammersley, “On estimating restricted parameters,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 12, no. 2, pp. 192–240, 1950. 3
- [22] D. G. Chapman and H. Robbins, “Minimum variance estimation without regularity assumptions,” *The Annals of Mathematical Statistics*, pp. 581–586, 1951. 3
- [23] R. McAulay and L. Seidman, “A useful form of the Barankin lower bound and its application to ppm threshold analysis,” *IEEE Trans. Inf. Theory*, vol. 15, no. 2, pp. 273–279, 1969. 3
- [24] R. McAulay and E. Hofstetter, “Barankin bounds on parameter estimation,” *IEEE Trans. Inf. Theory*, vol. 17, no. 6, pp. 669–676, 1971. 3

BIBLIOGRAPHY

- [25] J. S. Abel, “A bound on mean-square-estimate error,” *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 1675–1680, 1993. 3
- [26] A. Quinlan, E. Chaumette, and P. Larzabal, “A direct method to generate approximations of the Barankin bound,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 3. IEEE, 2006, pp. III–III. 3
- [27] K. Todros and J. Tabrikian, “General classes of performance lower bounds for parameter estimation- part I: Non-Bayesian bounds for unbiased estimators,” *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5045–5063, 2010. 3
- [28] E. Weinstein and A. J. Weiss, “A general class of lower bounds in parameter estimation,” *IEEE Trans. Inf. Theory*, vol. 34, no. 2, pp. 338–342, Mar. 1988. 4, 89
- [29] B. Bobrovsky and M. Zakai, “A lower bound on the estimation error for certain diffusion processes,” *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 45–52, 1976. 4
- [30] A. Renaux, P. Forster, P. Larzabal, and C. Richmond, “The Bayesian abel bound on the mean square error,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 3. IEEE, 2006, pp. III–III. 4
- [31] K. Todros and J. Tabrikian, “General classes of performance lower bounds for parameter estimation- part II: Bayesian bounds,” *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5064–5082, 2010. 4
- [32] I. Reuven and H. Messer, “A Barankin-type lower bound on the estimation error of a hybrid parameter vector,” *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 1084–1093, 1997. 4
- [33] J. Ziv and M. Zakai, “Some lower bounds on signal parameter estimation,” *IEEE Trans. Inf. Theory*, vol. 15, no. 3, pp. 386–391, 1969. 4
- [34] S. Bellini and G. Tartara, “Bounds on error in signal parameter estimation,” *IEEE Transactions on Communications*, vol. 22, no. 3, pp. 340–342, 1974. 4
- [35] D. Chazan, M. Zakai, and J. Ziv, “Improved lower bounds on signal parameter estimation,” *IEEE Trans. Inf. Theory*, vol. 21, no. 1, pp. 90–93, 1975. 4
- [36] E. Weinstein, “Relations between Belini-Tartara, Chazan-Zakai-Ziv, and Wax-Ziv lower bounds,” *IEEE Trans. Inf. Theory*, vol. 34, no. 2, pp. 342–343, 1988. 4
- [37] K. L. Bell, Y. Steinberg, Y. Ephraim, and H. L. Van Trees, “Extended Ziv-Zakai lower bound for vector parameter estimation,” *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 624–637, 1997. 4

BIBLIOGRAPHY

- [38] K. L. Bell, Y. Ephraim, and H. L. Van Trees, “Explicit Ziv-Zakai lower bound for bearing estimation,” *IEEE Trans. Signal Process.*, vol. 44, no. 11, pp. 2810–2824, 1996. 4
- [39] J. S. Turek, I. Yavneh, and M. Elad, “On MMSE and MAP denoising under sparse representation modeling over a unitary dictionary,” *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3526–3535, Aug. 2011. 5, 91
- [40] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, “Shrinkage algorithms for MMSE covariance estimation,” *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5016–5029, Oct. 2010. 5, 91
- [41] M. Fauß, H. V. Poor, and A. Dytso, “MMSE bounds under Kullback–Leibler divergence constraints on the joint input-output distribution,” in *2020 54th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2020, pp. 1477–1478. 5, 93
- [42] S. Haykin, “Cognitive radio: brain-empowered wireless communications,” *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, 2005. 5, 41
- [43] E. Biglieri, A. J. Goldsmith, L. J. Greenstein, H. V. Poor, and N. B. Mandayam, *Principles of cognitive radio*. Cambridge University Press, 2013. 5, 41, 42
- [44] E. Vul, C. Harris, P. Winkielman, and H. Pashler, “Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition,” *Perspectives on psychological science*, vol. 4, no. 3, pp. 274–290, 2009. 5, 26
- [45] J. D. Rosenblatt and Y. Benjamini, “Selective correlations; not voodoo,” *Neuroimage*, vol. 103, pp. 401–410, 2014. 5
- [46] Y. Benjamini and A. Meir, “Selective correlations-the conditional estimators,” *arXiv preprint arXiv:1412.3242*, 2014. 5
- [47] P. F. Thall, R. Simon, and S. S. Ellenberg, “A two-stage design for choosing among several experimental treatments and a control in clinical trials,” *Biometrics*, pp. 537–547, 1989. 5, 6, 13, 14
- [48] L. Shen, “An improved method of evaluating drug effect in a multiple dose clinical trial,” *Statistics in medicine*, vol. 20, no. 13, pp. 1913–1929, 2001. 5, 6
- [49] N. Stallard, S. Todd, and J. Whitehead, “Estimation following selection of the largest of two normal means,” *Journal of Statistical Planning and Inference*, vol. 138, no. 6, pp. 1629–1638, 2008. 5, 6, 13, 14, 34
- [50] J. Bowden and E. Glimm, “Unbiased estimation of selected treatment means in two-stage trials,” *Biometrical Journal*, vol. 50, no. 4, pp. 515–527, 2008. 5, 6, 13, 14, 15, 31, 34

BIBLIOGRAPHY

- [51] P. Bauer, F. Koenig, W. Brannath, and M. Posch, “Selection and bias: two hostile brothers,” *Statistics in Medicine*, vol. 29, no. 1, pp. 1–13, 2010. 5, 13, 14
- [52] M. W. Sill and A. R. Sampson, “Drop-the-losers design: Binomial case,” *Computational statistics & data analysis*, vol. 53, no. 3, pp. 586–595, 2009. 5, 6, 13, 14, 34, 38
- [53] N. Mukhopadhyay and T. K. Solanky, *Multistage Selection and Ranking Procedures: Second Order Asymptotics*. CRC Press, 1994, vol. 142. 5
- [54] J. Whitehead, “On the bias of maximum likelihood estimation following a sequential test,” *Biometrika*, vol. 73, no. 3, pp. 573–581, 1986. 5
- [55] B. Efron, “Tweedie’s formula and selection bias,” *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011. 5, 46
- [56] A. Somekh-Baruch, A. Leshem, and V. Saligrama, “On the non-existence of unbiased estimators in constrained estimation problems,” *IEEE Trans. Inf. Theory*, 2017. 5
- [57] S. Reid, J. Taylor, and R. Tibshirani, “Post-selection point and interval estimation of signal sizes in Gaussian samples,” *Canadian Journal of Statistics*, vol. 45, no. 2, pp. 128–148, 2017. 5, 15, 31
- [58] M. Carreras and W. Brannath, “Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection,” *Statistics in Medicine*, vol. 32, no. 10, pp. 1677–1690, 2013. 5, 15, 31
- [59] A. Cohen and H. B. Sackrowitz, “Two stage conditionally unbiased estimators of the selected mean,” *Statistics & Probability Letters*, vol. 8, no. 3, pp. 273–278, 1989. 5, 6, 13, 30, 31, 34, 36
- [60] D. S. Robertson, A. T. Prevost, and J. Bowden, “Accounting for selection and correlation in the analysis of two-stage genome-wide association studies,” *Biostatistics*, vol. 17, no. 4, pp. 634–649, 2016. 5, 15, 31, 34, 36
- [61] D. S. Robertson and E. Glimm, “Conditionally unbiased estimation in the normal setting with unknown variances,” *Communications in Statistics-Theory and Methods*, pp. 1–12, 2018. 5, 15, 31
- [62] T. Routtenberg, “Two-stage estimation after parameter selection,” in *2016 IEEE Statistical Signal Processing Workshop (SSP)*, June 2016, pp. 1–5. 6, 13, 31
- [63] E. Chaumette, P. Larzabal, and P. Forster, “On the influence of a detection step on lower bounds for deterministic parameter estimation,” *IEEE Trans. Signal Process.*, vol. 53, no. 11, pp. 4080–4090, 2005. 6, 27, 60, 89

BIBLIOGRAPHY

- [64] T. Weiss, T. Routtenberg, and H. Messer, “Total performance evaluation of intensity estimation after detection,” *Signal Processing*, p. 108042, 2021. 6
- [65] N. Harel and T. Routtenberg, “Low-complexity methods for estimation after parameter selection,” *IEEE Trans. Signal Process.*, vol. 68, pp. 1152–1167, Jan. 2020. 6, 8, 10, 48, 50, 52, 69, 89
- [66] —, “Post-parameter-selection maximum-likelihood estimation,” in *2021 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2021, pp. 446–450. 6, 8, 10, 64
- [67] E. Bashan, R. Raich, and A. O. Hero, “Optimal two-stage search for sparse targets using convex criteria,” *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5389–5402, 2008. 6, 46
- [68] J. A. Bazerque and G. B. Giannakis, “Distributed spectrum sensing for cognitive radio networks by exploiting sparsity,” *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1847–1862, 2010. 6
- [69] E. J. Msechu and G. B. Giannakis, “Sensor-centric data reduction for estimation with wsns via censoring and quantization,” *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 400–414, 2012. 6
- [70] X. Shen, H.-C. Huang, and J. Ye, “Inference after model selection,” *J. Am. Stat. Assoc.*, vol. 99, no. 467, pp. 751–762, 2004. 6, 60
- [71] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao, “Valid post-selection inference,” *The Annals of Statistics*, pp. 802–837, 2013. 6, 60, 61, 65
- [72] B. Efron, “Estimation and accuracy after model selection,” *Journal of the American Statistical Association*, vol. 109, no. 507, pp. 991–1007, 2014. 6, 60, 90
- [73] H. Leeb and B. M. Pötscher, “Model selection and inference: Facts and fiction,” *Econometric Theory*, vol. 21, no. 1, pp. 21–59, 2005. 6, 60
- [74] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor, “Exact post-selection inference, with application to the lasso,” *The Annals of Statistics*, vol. 44, no. 3, pp. 907–927, 2016. 6, 60, 64
- [75] W. Fithian, D. Sun, and J. Taylor, “Optimal inference after model selection,” *arXiv preprint arXiv:1410.2597*, 2014. 6, 60, 64
- [76] S. Zöllner and J. K. Pritchard, “Overcoming the winner’s curse: estimating penetrance parameters from case-control data,” *The American Journal of Human Genetics*, vol. 80, no. 4, pp. 605–615, 2007. 6, 64

BIBLIOGRAPHY

- [77] R. Heller, A. Meir, and N. Chatterjee, “Post-selection estimation and testing following aggregate association tests,” *Journal of the Royal Statistical Society Series B*, vol. 81, no. 3, pp. 547–573, 2019. 6, 64, 66
- [78] A. Belloni and V. Chernozhukov, “Least squares after model selection in high-dimensional sparse models,” *Bernoulli*, pp. 521–547, 2013. 6, 60, 61, 65
- [79] S. Cohen, T. Rountenberg, and L. Tong, “Non-Bayesian parametric missing-mass estimation,” *IEEE Trans. Signal Process.*, vol. 70, pp. 3709–3725, 2022. 6
- [80] E. Chaumette and P. Larzabal, “Cramér-Rao bound conditioned by the energy detector,” *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 477–480, 2007. 6, 60, 89
- [81] N. Harel and T. Rountenberg, “Bayesian post-model-selection estimation,” *IEEE Signal Process. Lett.*, vol. 28, pp. 175–179, Jan. 2021. 6, 9, 47, 60, 62, 64
- [82] P. J. Huber, “The behavior of maximum likelihood estimates under nonstandard conditions,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather Modification; University of California Press: Berkeley, CA, USA*, 1967, p. 221. 7, 55, 57
- [83] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” *Selected papers of H. Akaike*, pp. 199–213, 1998. 7, 55
- [84] H. White, “Maximum likelihood estimation of misspecified models,” *Econometrica*, vol. 50, no. 1, pp. 1–25, 1982. 7, 47, 50, 55, 56, 57, 64
- [85] R. H. Berk, “Limiting behavior of posterior distributions when the model is incorrect,” *The Annals of Mathematical Statistics*, vol. 37, no. 1, pp. 51–58, 1966. 7, 55
- [86] O. Bunke and X. Milhaud, “Asymptotic behavior of Bayes estimates under possibly incorrect models,” *The Annals of Statistics*, vol. 26, no. 2, pp. 617–644, 1998. 7, 55
- [87] Q. H. Vuong, “Cramér-Rao bounds for misspecified models,” California Institute of Technology, Division of the Humanities and Social Sciences, Working Papers 652, Oct. 1986. 7, 55, 56, 57, 58, 64, 72, 75, 79, 104
- [88] C. D. Richmond and L. L. Horowitz, “Parameter bounds on estimation accuracy under model misspecification,” *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2263–2278, 2015. 7, 55, 56, 64
- [89] S. Fortunati, F. Gini, and M. S. Greco, “The misspecified Cramér-Rao bound and its application to scatter matrix estimation in complex elliptically symmetric distributions,” *IEEE Trans. Signal Process.*, vol. 64, no. 9, pp. 2387–2399, 2016. 7, 55, 56, 64

BIBLIOGRAPHY

- [90] S. Fortunati, F. Gini, M. S. Greco, and C. D. Richmond, “Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 142–157, Nov. 2017. 7, 47, 50, 55, 56, 57, 72
- [91] S. Fortunati, F. Gini, M. S. Greco, A. M. Zoubir, and M. Rangaswamy, “Semiparametric CRB and Slepian-Bangs formulas for complex elliptically symmetric distributions,” *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5352–5364, 2019. 7, 55
- [92] M. Levy-Israel, I. Bilik, and J. Tabrikian, “MCRB on DOA estimation for automotive mimo radar in the presence of multipath,” *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–12, 2023. 7, 55
- [93] S. Fortunati, F. Gini, and M. S. Greco, “The constrained misspecified Cramér–Rao bound,” *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 718–721, 2016. 7, 55
- [94] L. T. Thanh, K. Abed-Meraim, and N. L. Trung, “Misspecified Cramér-Rao bounds for blind channel estimation under channel order misspecification,” *IEEE Trans. Signal Process.*, vol. 69, pp. 5372–5385, 2021. 7, 55
- [95] M. Khatib, N. Harel, Y. Ben-Horin, Y. Radzyner, and T. Routtenberg, “Cyclic misspecified cramer-rao bound for periodic parameter estimation,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 9911–9915. 7, 10, 55, 101
- [96] O. Krauz and J. Tabrikian, “Composite hypothesis tests for detection of modeling misspecification,” *IEEE Trans. Signal Process.*, vol. 70, pp. 351–365, 2021. 7, 55
- [97] N. E. Rosenthal and J. Tabrikian, “Model selection via misspecified Cramér-Rao bound minimization,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 5762–5766. 7, 55
- [98] A. Weiss, A. Lancho, Y. Bu, and G. W. Wornell, “A bilateral bound on the mean-square error for estimation in model mismatch,” *arXiv preprint arXiv:2305.08207*, 2023. 7, 55
- [99] N. Harel and T. Routtenberg, “Non-Bayesian post-model-selection estimation as estimation under model misspecification,” *IEEE Trans. Signal Process.*, vol. 72, pp. 3641–3657, 2024. 9
- [100] M. Wolf and C. Nadeu, “Channel selection measures for multi-microphone speech recognition,” *Speech Communication*, vol. 57, pp. 170–180, 2014. 14
- [101] F. Gini, “Estimation strategies in the presence of nuisance parameters,” *Signal processing*, vol. 55, no. 2, pp. 241–245, 1996. 14

BIBLIOGRAPHY

- [102] Y. Noam and H. Messer, “Notes on the tightness of the hybrid Cramér–Rao lower bound,” *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2074–2084, Feb. 2009. 14, 89
- [103] S. Bar and J. Tabrikian, “The risk-unbiased Cramér–Rao bound for non-Bayesian multivariate parameter estimation,” *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4920–4934, 2018. 14, 15
- [104] H. Sackrowitz and E. Samuel-Cahn, “Estimation of the mean of a selected negative exponential population,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 46, no. 2, pp. 242–249, 1984. 15, 31
- [105] M. Posch, F. Koenig, M. Branson, W. Brannath, C. Dunger-Baldauf, and P. Bauer, “Testing and estimation in flexible group sequential designs with adaptive treatment selection,” *Statistics in medicine*, vol. 24, no. 24, pp. 3697–3714, 2005. 15, 31
- [106] I. Goodd and R. A. Gaskins, “Nonparametric roughness penalties for probability densities,” *Biometrika*, vol. 58, no. 2, pp. 255–277, 1971. 17, 66
- [107] Y. C. Eldar, “Minimum variance in biased estimation: Bounds and asymptotically optimal estimators,” *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1915–1930, 2004. 17
- [108] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, pp. 1–22, 1977. 17
- [109] A. Meir and M. Drton, “Tractable post-selection maximum likelihood inference for the Lasso,” *arXiv preprint arXiv:1705.09417*, 2017. 17
- [110] R. Heller, A. Meir, and N. Chatterjee, “Post-selection estimation and testing following aggregated association tests,” *arXiv preprint arXiv:1711.00497*, 2017. 17
- [111] E. B. Andersen, “Asymptotic properties of conditional maximum-likelihood estimators,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 283–301, 1970. 17
- [112] P. K. Sen, “Asymptotic properties of maximum likelihood estimators based on conditional specification,” *The annals of Statistics*, pp. 1019–1033, 1979. 17
- [113] P. X.-K. Song, Y. Fan, and J. D. Kalbfleisch, “Maximization by parts in likelihood inference,” *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1145–1158, 2005. 18, 19, 21
- [114] B. Porat, *Digital processing of random signals: theory and methods*. Courier Dover Publications, 2008. 19

BIBLIOGRAPHY

- [115] R. Zamir, “A proof of the fisher information inequality via a data processing argument,” *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1246–1250, 1998. 21, 32
- [116] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 1990. 21, 27, 32
- [117] R. E. Bechhofer, “A single-sample multiple decision procedure for ranking means of normal populations with known variances,” *The Annals of Mathematical Statistics*, pp. 16–39, 1954. 23
- [118] S. S. Gupta, “On some multiple decision (selection and ranking) rules,” *Technometrics*, vol. 7, no. 2, pp. 225–245, 1965. 23
- [119] J. Kiefer, J. Wolfowitz *et al.*, “Stochastic estimation of the maximum of a regression function,” *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, 1952. 24
- [120] J. R. Blum, “Multidimensional stochastic approximation methods,” *The Annals of Mathematical Statistics*, pp. 737–744, 1954. 24
- [121] H. Robbins and S. Monro, “A stochastic approximation method,” in *Herbert Robbins Selected Papers*. Springer, 1985, pp. 102–109. 24
- [122] C. Robert and G. Casella, *Monte Carlo statistical methods*. Springer Science & Business Media, 2013. 25
- [123] K. Sricharan, R. Raich, and A. O. Hero, “Estimation of nonlinear functionals of densities with confidence,” *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4135–4159, 2012. 26
- [124] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967. 26, 43
- [125] K. Fukunaga, *Introduction to Statistical Pattern Recognition (2Nd Ed.)*. San Diego, CA, USA: Academic Press Professional, Inc., 1990. 26, 43
- [126] V. Berisha and A. O. Hero, “Empirical non-parametric estimation of the fisher information,” *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 988–992, 2015. 28
- [127] J. C. Spall, “Monte carlo computation of the Fisher information matrix in nonstandard settings,” *Journal of Computational and Graphical Statistics*, vol. 14, no. 4, pp. 889–909, 2005. 28
- [128] W. James and C. Stein, “Estimation with quadratic loss,” in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Univ of California Press, 1961, pp. 361–379. 36

BIBLIOGRAPHY

- [129] M. E. Bock, “Minimax estimators of the mean of a multivariate normal distribution,” *The Annals of Statistics*, pp. 209–218, 1975. 36
- [130] S. Bubeck, N. Cesa-Bianchi *et al.*, “Regret analysis of stochastic and nonstochastic multi-armed bandit problems,” *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012. 37
- [131] K. Liu and Q. Zhao, “Distributed learning in multi-armed bandit with multiple players,” *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, 2010. 37
- [132] H. Jiang, L. Lai, R. Fan, and H. V. Poor, “Optimal selection of channel sensing order in cognitive radio,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 297–307, 2009. 41
- [133] H. Urkowitz, “Energy detection of unknown deterministic signals,” *Proc. IEEE*, vol. 55, no. 4, pp. 523–531, 1967. 42
- [134] Z. Quan, S. Cui, A. H. Sayed, and H. V. Poor, “Optimal multiband joint detection for spectrum sensing in cognitive radio networks,” *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1128–1140, 2009. 42
- [135] S. M. Kay, “Fundamentals of statistical signal processing, vol. II: Detection theory,” *Signal Processing. Upper Saddle River, NJ: Prentice Hall*, 1998. 42, 81, 96
- [136] A. Jovicic and P. Viswanath, “Cognitive radio: An information-theoretic perspective,” *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 3945–3958, 2009. 43
- [137] K. M. Thilina, K. W. Choi, N. Saquib, and E. Hossain, “Machine learning techniques for cooperative spectrum sensing in cognitive radio networks,” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2209–2221, 2013. 43
- [138] R. L. Salamwade and D. M. Sakate, “Robust variable selection via penalized MT-estimator in generalized linear models,” *Communications in Statistics-Theory and Methods*, pp. 1–13, 2021. 46
- [139] E. Candes and T. Tao, “The Dantzig selector: Statistical estimation when p is much larger than n ,” *The annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007. 46, 60, 61, 65
- [140] A. M. Bruckstein, D. L. Donoho, and M. Elad, “From sparse solutions of systems of equations to sparse modeling of signals and images,” *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009. 46
- [141] T. T. Cai and L. Wang, “Orthogonal matching pursuit for sparse signal recovery with noise,” *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4680–4688, Jul. 2011. 46, 47, 53

BIBLIOGRAPHY

- [142] M. E. Tipping, “Sparse Bayesian learning and the relevance vector machine,” *Journal of Machine Learning Research*, vol. 1, no. 211, p. 244, Jun. 2001. 46
- [143] A. Koochakzadeh and P. Pal, “On saturation of the Cramér Rao bound for sparse Bayesian learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 3081–3085. 46, 90
- [144] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao, “Valid post-selection inference,” *The Annals of Statistics*, vol. 41, no. 2, pp. 802–837, Apr. 2013. 46
- [145] P. J. Huber, “The behavior of maximum likelihood estimates under nonstandard conditions,” in *Proc. of the fifth Berkeley symp. on math. stat. and prob.*, vol. 1, no. 1. University of California Press, 1967, pp. 221–233. 47
- [146] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York, NY, USA: Springer, 1998. 48, 69
- [147] K. Fukumizu, “Likelihood ratio of unidentifiable models and multilayer neural networks,” *Annals of Statistics*, vol. 31, no. 3, pp. 833–851, 2003. 52
- [148] T. M. Cover and J. A. Thomas, “Elements of information theory second edition solutions to problems,” *Internet Access*, pp. 19–20, 2006. 56
- [149] P. J. Schreier and L. L. Scharf, *Statistical signal processing of complex-valued data: the theory of improper and noncircular signals*. Cambridge university press, 2010. 59
- [150] H. Bozdogan, “Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions,” *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987. 60, 61, 65, 92
- [151] R. Nishii, “Maximum likelihood principle and model selection when the true model is unspecified,” in *Multivariate Statistics and Probability*. Elsevier, 1989, pp. 392–403. 60, 61, 65
- [152] K. P. Burnham and D. R. Anderson, “Practical use of the information-theoretic approach,” in *Model selection and inference*. Springer, 1998, pp. 75–117. 60, 61, 65
- [153] J. Ding, V. Tarokh, and Y. Yang, “Model selection techniques: An overview,” *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 16–34, 2018. 60, 61, 65, 80
- [154] J. C. Ye, Y. Bresler, and P. Moulin, “Cramér-Rao bounds for parametric shape estimation in inverse problems,” *IEEE Trans. Image Process.*, vol. 12, no. 1, pp. 71–84, 2003. 79
- [155] Z. Ben-Haim and Y. C. Eldar, “The Cramér-Rao bound for estimating a sparse parameter vector,” *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3384–3389, Jun. 2010. 79

BIBLIOGRAPHY

- [156] P. Stoica and Y. Selen, “Model-order selection: a review of information criterion rules,” *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004. 80
- [157] H. L. Van Trees and K. L. Bell, “Bayesian bounds for parameter estimation and nonlinear filtering/tracking,” *AMC*, vol. 10, p. 12, 2007. 89, 94, 105
- [158] A. Renaux, P. Forster, P. Larzabal, C. D. Richmond, and A. Nehorai, “A fresh look at the Bayesian bounds of the Weiss-Weinstein family,” *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5334–5352, June 2008. 89
- [159] A. Yeredor, “The joint MAP-ML criterion and its relation to ML and to extended least-squares,” *IEEE Trans. Signal Process.*, vol. 48, no. 12, pp. 3484–3492, Dec. 2000. 89
- [160] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, “A consistent metric for performance evaluation of multi-object filters,” *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, July 2008. 89
- [161] M. Rezaeian and B.-N. Vo, “Error bounds for joint detection and estimation of a single object with random finite set observation,” *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1493–1506, Mar. 2009. 89
- [162] R. Prasad and C. R. Murthy, “Cramér-Rao-type bounds for sparse Bayesian learning,” *IEEE Trans. Signal Process.*, vol. 61, no. 3, pp. 622–632, Oct. 2012. 90
- [163] R. Boyer, R. Couillet, B. Fleury, and P. Larzabal, “Large-system estimation performance in noisy compressed sensing with random support of known cardinality- a Bayesian analysis,” *IEEE Trans. Signal Process.*, vol. 64, no. 21, pp. 5525–5535, 2016. 90
- [164] R. Boyer and P. Larzabal, “Sparsity-based estimation bounds with corrupted measurements,” *Signal Processing*, vol. 143, pp. 86–93, 2018. 90
- [165] E. Weinstein and A. Weiss, “Lower bounds on the mean square estimation error,” *Proceedings of the IEEE*, vol. 73, no. 9, pp. 1433–1434, Sep. 1985. 94
- [166] L. Bacharach, C. Fritsche, U. Orguner, and E. Chaumette, “A tighter Bayesian Cramér-Rao bound,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5277–5281. 94
- [167] Z. Ben-Haim and Y. C. Eldar, “The Cramér-Rao bound for estimating a sparse parameter vector,” *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3384–3389, Mar. 2010. 95
- [168] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, “Model-based deep learning,” *Proceedings of the IEEE*, 2023. 101

BIBLIOGRAPHY

- [169] H. V. Habi, H. Messer, and Y. Bresler, “Learning to bound: A generative Cramér-Rao bound,” *IEEE Trans. Signal Process.*, 2023. 101
- [170] T. Routtenberg and J. Tabrikian, “Cyclic Barankin-type bounds for non-Bayesian periodic parameter estimation,” *IEEE Trans. Signal Process.*, vol. 62, no. 13, pp. 3321–3336, 2014. 101
- [171] J. D. Cohen, “Noncentral chi-square: Some observations on recurrence,” *The American Statistician*, vol. 42, no. 2, pp. 120–122, 1988. 113

תקציר

שערוך לאחר בחירה מתאר מקרה בו שלב בחירה מקדים, מבוסס־נתונים קובע את בעיית השערוך. בעבודה זאת, אנו חוקרים שני תרחישים יסודיים של בעיה זו. שערוך לאחר בחירת מודל ושערוך לאחר בחירת פרמטרים. בשערוך לאחר בחירת מודל, ייתכן כי מודל המדידות לא ידוע. לכן, טרם ביצוע שערוך מתבצעת בחירה של מודל מתוך סט מודלים מועמדים. לאחר מכן, מתבצע שערוך של הפרמטרים במודל הנבחר. בשערוך לאחר בחירת פרמטרים המודל ידוע, אך, תהליך בחירה מסויים קובע על סמך הנתונים מהם הפרמטרים הרצויים לשערוך בעוד האחרים נחשבים לפרמטרים טורדניים. בשני המקרים לתהליך הבחירה יש השפעה על השערוך שאחריו לדוגמה השפעה שבאה לידי ביטוי בתוספת של הטיה. לאורך מחקר זה, חקרנו היבטים שונים של שתי הבעיות, בפרט, שיטות שערוך הגדרות מתאימות להטיה וחסמי ביצועים.

תרומתה של עבודת דוקטורט זו מפורטת להלן. ראשית, עבור שערוך אחר בחירת פרמטרים אנו מגדירים את השגיאה הריבועית הממוצעת לאחר בחירה, (PSMSE), כמדד ביצועים מתאים שלוקח בחשבון את הליך הבחירה. אנו מציגים קריטריון הטיה מתאים, חוסר ההטיה במובן להמן ביחס לשגיאה הריבועית לאחר בחירה. אנו מציגים את המשעריך הסבירות המירבית לאחר בחירה, PSML, כמשעריך מתאים לבעיה זו. מכיוון שלמשעריך זה אין ביטוי אנליטי סגור בדרך כלל, פיתחנו שיטת קירוב חדשה בסיבוכיות נמוכה לחישוב המשעריך. נוסף על כך, הצגנו חסם מתאים מסוג קרמר־ראו לבעיה ופיתחנו שיטה יעילה לחישוב אמפירי שלו.

הרחבנו את המודל למקרה בו ייתכן שהמודל לא ניתן לזיהוי, מקרה בו לא ניתן לשעריך את כל הפרמטרים. במקרה זה, מתבצע שלב בחירה שמזהה את על בסיס המדידות את הפרמטרים המשמעותיים ביותר כשלב מקדים לשערוך. אנו מציגים את משעריך הסבירות המירבית לאחר בחירה הקוהרנטי, coherent PSML, כמעריך מתאים לבעיה זו ומפתחים אלגוריתם פרקטי למימוש משעריך זה. אנחנו מראים כי המשעריך המוצע משיג ביצועים טובים יותר משיטות מקובלות אחרות בבעיה זו.

שנית, אנו מתייחסים לבעיית שערוך לאחר בחירת מודל. שערוך לאחר בחירת מודל קשורה קשר הדוק לרעיון של שערוך תחת מודל שגוי. בעוד שהספרות בנושא שערוך תחת מודל שגוי נותנת מסגרת להתייחסות למודל השגוי היא אינה מתייחסת לתהליך הבחירה שמוביל לשגיאת המודל. בשערוך לאחר בחירת מודל אנחנו מתייחסים למספר מודלים בתור כמועמדים להיות המודל המשוער. לכן, הפרשנות עבור המודל המשוער בבעיה זו אינה ישירה. אנו מציגים שלוש פרשנויות שונות כדי להגדיר את הבעיה של שערוך לא סייסיאני לאחר בחירת מודל כבעיית שערוך תחת מודל שגוי. כל אחת מהן מגדירה משעריך סבירות מירבית תחת מודל שגוי מתאים וכן חסם מסוג קרמר־ראו תחת שגיאת מודל. לסיום, אנו שוקלים בעיית שערוך בייסיאני לאחר בחירת מודל לשערוך פרמטרים אקראיים תחת מודל דטרמניסטי לא ידוע. אנחנו מציגים שמערכים שונים שחסמים שונים לבעיה זו. בפרט, את חסם קרמר־ראו הסלקטיבי וחסם קרמר־ראו הסלקטיבי ההדוק כחסמים על השגיאה הריבועית הממוצעת.

מילות מפתח:

שערוך בייסיאני, חסם קרמר־ראו, חוסר הטיה במובן להמן, חסמים תחתונים, שגיאה ריבועית ממוצעת, שגיאת מודל, בחירת מודל, שערוך לא בייסיאני, שערוך פרמטרי, בחירת פרמטרים, הסקה סלקטיבית

תורת השערוך לאחר בחירה

מחקר לשם מילוי חלקי של הדרישות לקבלת תואר "דוקטור לפילוסופיה"

מאת

הראל

נדב

הוגש לסינאט אוניברסיטת בן גוריון בנגב

אישור המנחה

אישור דיקן בית הספר ללימודי מחקר מתקדמים ע"ש קרייטמן

20.5.24

יב' אייר התשע"ד

באר שבע